Phonetics Information Base and Lexicon

Steven Paul Moran

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2012

Reading Committee:

Emily M. Bender, Chair

Richard Wright, Chair

Scott Farrar

Sharon Hargus

Program Authorized to Offer Degree:
Department of Linguistics

University of Washington

Abstract

Phonetics Information Base and Lexicon

Steven Paul Moran

Co-Chairs of the Supervisory Committee:
Associate Professor Emily M. Bender
Department of Linguistics

Associate Professor Richard Wright
Department of Linguistics

In this dissertation, I investigate the linguistic and technological challenges involved in creating a cross-linguistic data set to undertake phonological typology. I then address the question of whether more sophisticated, knowledge-based approaches to data modeling, coupled with a broad cross-linguistic data set, can extend previous typological observations and provide new ways of querying segment inventories. The model that I implement facilitates testing typological observations by aligning data models to questions that typologists wish to ask. The technological infrastructure that I create is conducive to data sharing, extensibility and reproducibility of results. I use the data set and data models in this work to validate and extend previous typological observations.

In doing so, I revisit the typological facts proposed in the linguistics literature about the size, shape and composition of segment inventories in the world's languages and find that they remain similar even with a much larger sample of languages. I also show that as the number of segment inventories increases, the number of distinct segments also continues to increase. And when vowel systems grow beyond the basic cardinal vowels, they do so first by length and nasalization, and then diphthongization.

Moving beyond segments, I show that distinctive feature sets in general lack the typological representation needed to straightforwardly map sets of features to the segment types found in a broad set of language descriptions. Therefore, I extend a distinctive feature

set, devise a method to computationally encode features by combining feature vectors and assigning them to segment types, and create a system in which users can query by feature, by sets of features that define natural classes, or by omitting features in queries to utilize the underspecification of segments. I use this system and reinvestigate proposed descriptive universals about phonological systems and find that some, but not all universals hold up to the more rigorous testing made possible with this larger data set and a graph data model.

Lastly, I reevaluate one of the many purported correlations between a non-linguistic factor and language: the claim that there exists a relationship between population size and phoneme inventory size. I show that this finding is actually an artifact of a small data set, which constrains the use of more nuanced statistical approaches that can control for the genealogical relatedness of languages. Thus, in this work I illustrate how researchers can leverage the data set and data models that I have implemented to investigate different aspects of languages' phonological systems, including the possible impact of non-linguistic factors on phonology.

TABLE OF CONTENTS

LIST OF FIGURES

Figure Number Page

LIST OF TABLES

Table Number Page

ACKNOWLEDGMENTS

Procrastination took several forms while I worked on this dissertation. One that I enjoyed was reading the acknowledgments in the theses from which I extracted data for my own research. There was a clear theme among those acknowledgments that is particularly relevant here: although my name appears on the title page, this dissertation is the outcome of an enormous group effort. In this work, the data come from thousands of linguists, many of whom lived far away from their homes and family to work with groups of speakers of languages very different than their own. Many of these linguists risked their well-being and some of them lost their lives undertaking fieldwork. Their sacrifice is not forgotten.

I would like to express my sincerest gratitude to my advisors Emily Bender and Richard Wright for their encouragement, guidance and patience, and for sharing with me a wealth of knowledge and enthusiasm while allowing me to create my own path. I would like to thank the members of my dissertation committee: Scott Farrar for inspiration in knowledge representation and ontologies, and Sharon Hargus for copious and detailed comments (even after I dumped the entire dissertation on her desk just a few weeks before the defense). William Lewis was around for the original vision of PHOIBLE; in fact, he coined the acronym. Joyce Parvi and Mike Furr keep things running smoothly in the Linguistics Department at the University of Washington and they made my time there enjoyable, humorous and hassle-free. Daniel McCloy deserves special thanks for always answering my questions, leading the segment mappings, and for devising the statistical approach in Chapter 7. Thanks to Richard John Harvey for key exchanges in database principles and design, and to Taras Zakharko and Jelena Prokić for discussions on statistical approaches and coding in R. Jeff Good set me on the path to interoperability and he was integral to coordinating my first visit to the Max Planck Institute in Leipzig. I owe a debt of gratitude to Bernard Comrie for supporting visits to MPI and I'm thankful to Martin Haspelmath for insightful discus-

sions while I was there. One could say that Michael Cysouw has been a shadow dissertation committee member; he was at MPI when many of the ideas for this work were first being formulated, and he is still around now and providing invaluable input as I complete this work.

Many thanks go to colleagues who contributed directly to PHOIBLE's development: Morgana Davids, Scott Drellishak, David Ellison, Kelley Kilanski, Michael McAuliffe, Brandon Plasters, Kevin Pittman, Cameron Rule, Daniel Smith, Daniel Veja, and Christopher Green and Tristan Purvis. Several people created open source tools that saved me a great deal of time with formatting and data entry: Jim Fox, Swen Grawunder Harald Hammarström, Hans-Jörg Bibiko and Weston Ruter.

I'm grateful to close friends and family for their moral support and spirits: Patrick Blaine, Joshua Brown, Sara Pizzoni and Bill Jenkins (at "the office"), Andrea Henderson, Prashant Nagaraja, Chris Stamm, Gudrun Stamm, Alena Witzlack-Makarevich and Jack Wolf.

Laurie Poulson was not only my mentor at UW, but she was a great friend and a source of inspiration throughout the years. Laurie I wish you were here to witness the completion of this dissertation, but even much more so, I wish you would have been given the time on this Earth to finish yours.

Finally, to Shauna Eggers. I would not have made it this far or written *this* dissertation without your unwavering love and support. Thank you so much.

This research was made possible by several sources of financial support. While I was a student at UW, I received research assistantships from Paul Aoki and Bob Majors at the Language Learning Center and from Tom Lewis at UW Information Technology. The Linguistics Department also supported my studies with several teaching assistantships and I am grateful to the faculty for giving me those opportunities. A grant awarded by the University of Washington's Royalty Research Fund supported work on PHOIBLE in 2009.

DEDICATION

to Shauna, my mate,
and Laurie, my mentor

Chapter 1

INTRODUCTION

This thesis is broadly concerned with identifying and overcoming the linguistic and technological challenges involved in:

1. creating a cross-linguistic data set to undertake phonological typology

2. modeling this data set in ways that facilitate testing typological observations by aligning the data models to questions that typologists wish to ask

3. instantiating technological infrastructure that is conducive to data sharing, extensibility and reproducibility of results

4. using the data set and data models in this work to validate and extend previous typological observations

The central thesis of this dissertation is that more sophisticated, knowledge-based approaches to data modeling, coupled with a larger cross-linguistic data set, will extend previous typological observations by allowing researchers to query segment inventories at the level of distinctive features. Thus we can ask if previous observations in phonological typology are validated on a larger scale and we can investigate what are the new observations that can be made.

Phonological typology typically involves comparing languages by the number or types of sounds, or *segments* when encoded by graphic symbols, that they contain. My work draws on linguistic research in segmental phonology and distinctive feature theory, and on computational research in data modeling and knowledge representation. In this work my colleagues and I have created a cross-linguistic data set and I have modeled this data set in ways that allow researchers to investigate the variation of phonological systems across languages at the level of segments and at the level of distinctive features.

The motivation behind this work was to collect a much larger and broader cross-linguistic sample of phonological inventories than what was previously available and to model the data in an interoperable way so that users could federate disparate linguistic and non-linguistic information and pose novel questions on the combined data set. I call the resource that I have developed the Phonetics Information Base and Lexicon (PHOIBLE).[1] PHOIBLE incorporates the segment inventories from the Stanford Phonology Archive (SPA; Crothers et al. 1979), the UCLA Phonological Segment Inventory Database (UPSID; Maddieson 1984, Maddieson and Precoda 1990) and the *Systèmes alphabátiques des langues africaines* (AA; Hartell 1993, Chanard 2006). The genealogical and geographical coverage of these combined inventories is expanded by the work that my colleagues and I have undertaken in extracting phonological inventory data from hundreds of grammars and phonological descriptions.[2] This combined data sample contains 1336 segment inventories, which represent 1089 distinct languages, or roughly 16% of the world's estimated 6909 languages, as listed in the Ethnologue (Lewis, 2009).[3] Inventories range in detail from phonemic descriptions to fuller phonological descriptions including phonemes, allophones, their conditioning environments and additional information like phonological rules and a description of marginal sounds. The PHOIBLE data set is illustrated in Figure 1.1.

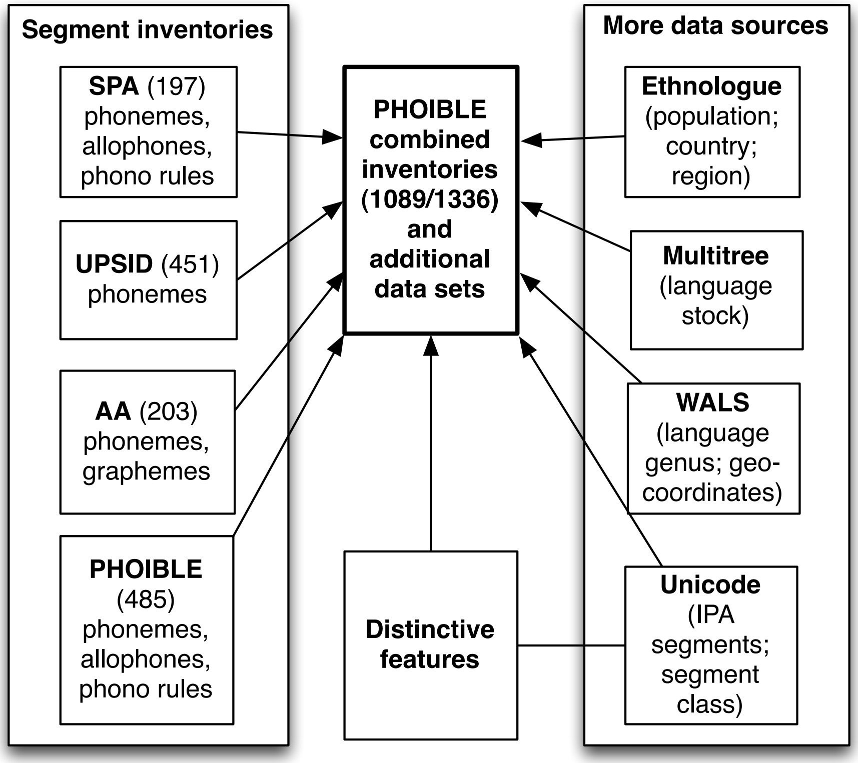
A major challenge in this work has been addressing the question of how to bring together these segment inventory databases, which are heterogeneous in format, encoding and content, into an accessible data model that is extensible and which can integrate additional linguistic and non-linguistic information. Before the integration processes and the resulting data models could be instantiated, however, there were many methodological considerations at the linguistic and technological levels that had to be identified and addressed, which I do in Chapter 2. I begin by defining the conventions and linguistic and technological terminology used throughout this work in Section 2.1. In Section 2.2 I provide a brief background

[1] `http://phoible.org/`

[2] See Appendix B.

[3] See Chapter 4 for details regarding the data set.

Figure 1.1: PHOIBLE overview


Segment inventories
SPA (197) phonemes, allophones, phono rules
UPSID (451) phonemes
AA (203) phonemes, graphemes
PHOIBLE (485) phonemes, allophones, phono rules
PHOIBLE combined inventories (1089/1336) and additional data sets
Distinctive features
More data sources
Ethnologue (population; country; region)
Multitree (language stock)
WALS (language genus; geo-coordinates)
Unicode (IPA segments; segment class)

on the fundamental linguistic theories pertinent to this work: segmental phonology and distinctive feature theory. Then in Section 2.3 I describe the theoretical and technological challenges in developing a cross-linguistic segment inventory data set, which involve undertaking typology with databases (Section 2.3.1), statistical sampling (2.3.2), data and analysis (2.3.3), linguistic segments (2.3.4), standardization (2.3.5) and metadata and data provenance (2.3.6).

From the beginning my goal has been to create a tool for typology that is extensible and that can also interoperate with additional linguistic and non-linguistic data sets. Although the inventories in PHOIBLE represent a convenience sample, i.e. a set of languages chosen from sources that are readily available, each segment inventory is associated with data regarding its genealogical affiliation, including its language family stock from the Ethnologue

(Lewis, 2009) via Multitree[4] and its language genus from the World Atlas of Language Structures (WALS; Haspelmath et al. 2008). Geographical information for each language also comes from the Ethnologue (country and geographic region) and WALS (geo-coordinates). Genealogical and geographic information is pertinent to statistical sampling in linguistic typology so that factors of shared descent and areal diffusion can be accounted for and can be used to inform statistical observations. Non-linguistic information, such as demographic data, is also included so that various cross-cultural and cross-disciplinary studies can be undertaken.[5]

In this work, as explained in Chapters 3 and 4, syntactic and semantic interoperability are achieved by extracting the segment inventory data from various disparate formats, bringing the data together into one data set that adheres to a well-defined standard of segments and their distinctive features, and then modeling the data set into formal data models that are aligned to questions that typologists wish to ask. Section 3.1 begins with a brief overview of several data models and examples. I then describe in detail in Section 3.2 the three PHOIBLE data models (flat file tables, a relational database and an RDF graph) and I provide many examples of how a user might query each. In Section 3.3 I discuss aspects of knowledge representation and how formal logic constraints can be integrated with the PHOIBLE RDF graph to create a ‘knowledge base’, i.e. a collection of assertions about phonological inventories and data related to those languages in a formal knowledge representation language. The graph model coupled with a knowledge representation formalism allows researchers to manipulate aspects of the PHOIBLE data set, such as specifying that the distinction between long and short vowels should be collapsed or that diphthongs should be ignored in a query, without changing the underlying data and thus allowing the researcher to apply his or her own analytical preferences to the data. Additionally, I have defined an ontology to encode concepts and their relationships in the data, so that a vocabulary of phonetic features has been given hierarchical structure to represent feature geometries, which can then be used to query the PHOIBLE data set or selected portions of

[4] http://multitree.linguistlist.org

[5] I give an example in Chapter 7.

it. Users can extend this ontology or define their own ontologies to interact with the data in PHOIBLE in different ways.

In Chapter 4 I provide an overview of PHOIBLE. Section 4.2 discusses my motivation for building PHOIBLE and in Section 4.3 I discuss how I processed and merged the different segment inventory databases into one cross-linguistic data set, highlighting the challenges particular to each data source. In Section 4.4 I evaluate the genealogical coverage of the combined segment inventories.

As I will show in this work, there is no one-data-model-fits-all approach for investigating questions in phonological typology. Data are ideally modeled in ways that are flexible such that different typological observations can be tested in appropriate ways and the same questions can be approached from multiple perspectives.[6] In Chapter 5 I revisit the typological facts put forth in the literature about segments and segment inventories and evaluate these claims against the expanded PHOIBLE data set. In Section 5.2 I provide some background and in Section 5.3 I use the denormalized table format of the PHOIBLE data set and load the data tables into statistical software to examine and illustrate properties of segment inventories and the distribution of segments cross-linguistically. Interestingly, as new inventories are added to the PHOIBLE data set, new distinct segment types continue to appear showing an increase in segment types that is quadratic. In Section 5.4 I show that many of the observations made by Maddieson (1984) about segment inventories, such as average inventory size, etc., are still valid even in a much broader and larger cross-linguistic data set. I also implement a statistical sampling technique to account for effects of genealogical skew because the PHOIBLE data set is not inherently genealogically balanced.[7] Another topic of typological interest, particularly in the area of investigating language complexity in phonological systems, is the balance between consonants and vowels across inventories. This topic is investigated in Section 5.5. In Section 5.6 I revisit Crothers's (1978) observation that vowel systems in most languages contain /i, a, u/. With the table data model,

[6] PHOIBLE is a tool for typological comparisons and description, not a tool for modeling acquisition or probing cognitive function.

[7] See Section 4.4 for a discussion of PHOIBLE's genealogical coverage and Appendix A for a list of its genealogical coverage by language family.

I use the multi-dimensional scaling statistical technique to visualize implications in vowel systems and how they tend to expand after /i, a, u/.

Another goal of my work is to provide novel access to phonological inventories and their associated data at a level deeper than the segment, that is, at the level of distinctive features. Chapter 6 is concerned with distinctive features and how to model them and use them to investigate phonological inventories at the sub-segment level. In Section 6.2 I provide a discussion of distinctive features and in Section 6.3 I show that current distinctive feature sets have poor typological coverage. Therefore in Section 6.4 I devise and implement a computational approach to assign distinctive feature vectors to segment types undefined in traditional distinctive feature sets. Finally, in Section 6.5 I use the distinctive features in a graph model, combined with the segment inventories in PHOIBLE, to investigate descriptive universals put forth about phonological systems in the world's languages and show that not all languages have coronals, as was previously proposed (Hyman, 2008) and rebutted (Blevins, 2009).

In Chapter 7 I present a case study using the PHOIBLE database to investigate one of many claims regarding societal effects on language structure. I use the segment inventory and demographic data and apply a hierarchical linear model to show that there is no correlation between population size and phoneme inventory size (Haudricourt, 1961; Trudgill, 1997, 2002; Hay and Bauer, 2007), once one accounts for the non-independence of data points due to genealogical factors inherent in cross-linguistic data sets.

Lastly, in Chapter 8 I provide my concluding remarks and then discuss my contributions to the field in Section 8.2. In Section 8.3 I discuss the 'LExicon' part of PHOIBLE and the challenges involved in linking lexicons to segment inventories. In Section 8.4 I lay out avenues for future research.

Chapter 2

BACKGROUND

I begin this chapter by defining the conventions and the linguistic and technological terminology used throughout this work. In Section 2.2 I provide an overview of segmental phonology and distinctive feature theory, which are the frameworks that I develop technological infrastructure for undertaking studies in phonological typology. In Section 2.3 I discuss the challenges involved in developing this infrastructure and the general issues in large cross-linguistic typological studies. My goal in this chapter is to situate the pertinent theories and technologies within the context of the development of PHOIBLE, which I describe in detail in Chapters 3 & 4. In later chapters I use PHOIBLE to investigate issues of phonological typology at the segment and feature levels.

2.1 Conventions and terminology

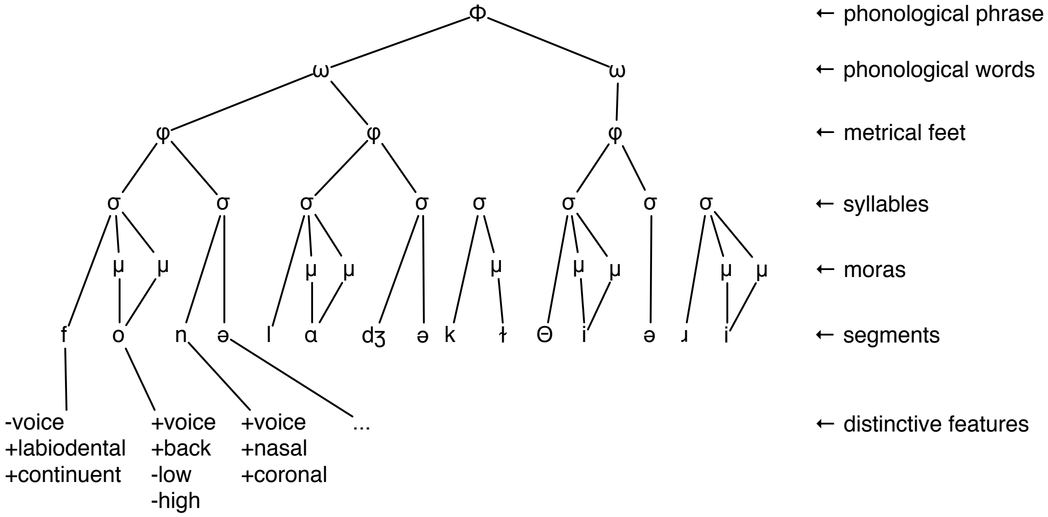
2.1.1 Conventions

All phonemic and phonetic representations are given in the International Phonetic Alphabet (IPA) (International Phonetic Association, 2005), unless noted otherwise. Standard conventions are used for distinguishing between graphemic < >, phonemic / / and phonetic representations [ ]. For character data information, I follow the Unicode Standard's notational conventions (The Unicode Consortium, 2007). Character names are represented in small capital letters (e.g. LATIN SMALL LETTER SCHWA) and code points are expressed as *U*+*n* where *n* is a four to six digit hexadecimal number, e.g. U+0256, which is rendered as the glyph <ə>. When I refer to a relational database table or column name, I use the `Courier monospace font`.

2.1.2 Linguistic terminology for phonology

Phonological theory can be divided into segmental and prosodic phonology. Prosodic phonology is concerned with suprasegmental phenomena, i.e. features and structures at a higher level than the segment, such as tone, stress, moras, syllables, metrical feet, phonological words and intonation. An illustration is provided in Figure 2.1.

Figure 2.1: Prosodic and segmental structure (adapted from Howe 2003, 2)



The present work deals mainly with the segment and features below the segment. A **segment** is an abstraction of a articulatory or auditory unit of speech production or perception. Segments are discrete (separate and individual) and are serially ordered, so as to model the speech stream as a temporal sequence of distinct states. A segment is called a **phone** if it is an unanalyzed sound in a language, i.e. it is an identifiable unit in the speech stream, but it has not been analyzed as contrastive or not. A contrastive set of segments in a language determines the language's phonemes. A **phoneme** is a minimally distinctive sound in a particular language variety.[1] An **allophone** is a phonetic variant of a phoneme that occurs

[1] Phonemes are theoretical constructs, determined by a linguist who has studied the sounds of a particular language, and chosen a set of contrastive segments based on phonological principles. Thus the set of phonemes in a language may be contested by different linguists.

in free variation or in complementary distribution with other phonetically similar segments.

Each spoken language uses a set of consonants and vowels to form words (all languages have consonants and vowels; many also have tone). This set is called a **segment inventory** and it is typically stated in terms of a language variety-specific set of phonemes, as analyzed by a linguist.[2] A segment inventory describes the speech sounds used by speakers of a particular language and encodes the phonetic dimensions employed by the phonological system to form meaningful contrasts. The notion of a segment inventory has been defined as an abstraction over the set of distinctive segments used by a particular language variety's phonological system, as defined by the set of distinctive features employed by the language (Clements, 2009, 19).

A segment is comprised of a set of distinctive features, as defined by a particular distinctive feature theory. In distinctive feature theory, segments are modeled as bundles of distinctive features. **Distinctive features** are the basic phonetic units of a segment and are typically modeled by their articulatory and/or acoustic properties as binary feature values. The IPA provides symbols as a shorthand for representing articulatory features, e.g. the segment <p> (phonemically /p/ or allophonically [p]) is a voiceless bilabial plosive. In the Hayes 2009 feature set, this sound is modeled with the distinctive features [−voice, +labial, −delayed release, etc.], which serve to contrast <p> with all other sounds.

In this work I will make a few further distinctions between different kinds of segments. I define a type-token distinction among segments in the world's languages. On the one hand, a segment can be used to encode a particular sound in a particular language, e.g. the German <i> sound. I refer to this kind of segment as a **segment token**; it is language-specific because the auditory properties of a segment like <i> as spoken by native speakers of German or English varies measurably.[3] On the other hand, a segment may be used to encode an abstract class of segments that may pattern in similar ways across languages, e.g. German, English and many other languages have an <i> sound. For this abstract sense, I

[2] A segment inventory may also include contrastive autosegments (e.g. tone, stress, other prosodic features) and a description of the set of allophones as determined by the linguist. Segment inventories in the world's languages range widely in size and shape. See Chapter 5 for details.

[3] In fact, we can say that segment tokens are language-variety specific. For example, the <r> sound in many dialects of German is pronounced noticeably different, thus adding to an individual's accent.

refer to the set of similar segments across languages as a **segment type**. To confuse matters, linguistic segments and diacritics can combine into what has also been labeled *segment types* in the literature (Sagey, 1986; Clements and Hume, 1995). I will refer to these three different types of segments (simple, complex and contour) as **segment classes**.[4]

2.1.3 Linguistic terminology for writing systems

Transcription is a scientific procedure, and also the result of that procedure, for representing the sounds of human speech. It incorporates a set of unambiguous symbols to represent distinctive speech sounds with conventions that specify how these symbols should be combined. IPA is a commonly used transcription system that provides a medium for transcribing languages at both phonetic and phonemic levels (narrow and broad transcriptions). In this thesis, a **transcription system** is a system of symbols and rules for graphically transcribing the sounds of a language variety. A **practical orthography** is a phonemic writing system designed for practical use by speakers. The mapping relation between phonemes and graphemes in practical orthographies is purposely shallow, i.e. there is a systematic and faithful one-to-one mapping from a phoneme to a grapheme.[5] The IPA is often used by field linguists in the development of practical orthographies for languages without writing systems. Practical orthographies are a kind of orthography. An **orthography** specifies the symbols, punctuation, and the rules in which a language is correctly written in a standardized way. All orthographies are language-specific.

Orthographies and transcription systems are both kinds of **writing systems**. A writing system is a symbolic system that uses visible or tactile signs to represent language in a systematic way. The term writing system has two mutually exclusive meanings. First, it may refer to the way a particular language is written, i.e. the writing system of a particular language. For example, the Serbian writing system use two scripts: Latin and Cyrillic. Second, writing system may refer to an abstract type of writing system, i.e. how scripts

[4] Complex and contour segment classes pose challenges in assigning distinctive features to segments. Segment classes and the assignment of features to segment types are described in Section 6.4.

[5] Practical orthographies are intended to jump-start written materials development by correlating a writing system with its sound units (cf. Meinhof and Jones 1928).

have been classified according to the way that they encode sounds or words in languages. For example, the Latin and Cyrillic scripts are both alphabets. Over the years linguists have typologized writing systems in a variety of ways, with the tripartite classification of logography, syllabary, and alphabet remaining the most popular, even though there are at least half a dozen different types of writing systems (Daniels, 1990, 1996).

A logographic writing system uses symbols that visually represent words or morphemes. A prototypically cited example is the Chinese writing system, although it is more appropriately classified as a logosyllabary. A syllabary uses symbols to denote syllables; for example, Japanese Kana are syllabic scripts. An alphabet relates symbols to sounds for consonants and vowels. A purely consonantary system is called an abjad (the Arabic script is the most wide-spread example) and an abugida is a type of writing system that uses symbols to encode units of a consonant accompanied by a specific vowel, e.g. Indic scripts (Daniels, 1990). Featural systems are less common and encode phonological features within the shapes of the symbols represented in the script. Korean Hangul is the most cited example. A writing system may also contain features of more than one system type.[6]

The term **script** refers to a collection of distinct symbols as employed by one or more writing systems.[7] For example, both Serbian and Russian are written with subsets of the Cyrillic script. A type of writing system can also be written with different scripts, e.g. the alphabet can be written in Latin and Cyrillic scripts (Coulmas, 1999). And a language, like Serbian or Japanese, can be written in different scripts.

In the terminology of writing systems, a **character** is both a general term for any self-contained element and a conventional term for a unit in the Chinese writing system (Daniels, 1996). In technological terminology, a **character** refers to the electronic encoding of a component in a writing system that has semantic value.[8] Different definitions for the term *character* are confusing. For example, although a Chinese character may be encoded as a single basic unanalyzable unit electronically, it may be the case that at a more fine-grained

[6] See discussions and examples in Sampson 1985b; Daniels 1990, 1996; Coulmas 2003.

[7] Note the term *script* also refers to a short computer program written in a programming language, e.g. her script parses out the headwords from an online dictionary.

[8] See Section 2.1.4.

level of analysis the internal structure of the character is comprised of smaller semantic and phonetic units that should be considered graphemes (Sproat, 2000).

A **grapheme** is the basic, minimally distinctive symbol of a particular writing system. Like the phoneme is an abstract representation of a distinct sound in a language, the term grapheme was modeled after phoneme and represents a contrastive graphical unit in a writing system.[9] Conditioned or free variants of a grapheme are called **allographs**; for example, the distinctive forms of Hebrew letters used at the end of a word are conditioned, and the different forms of letters like <a> or <ɑ> and <g> or <ɡ> are in free variation (Daniels and Bright, 1996).

A script may employ multiple graphemes to represent a single phoneme. For example, the graphemes <c> and <h> when conjoined in English orthography represent one phoneme in English, the digraph <ch> pronounced /ʧ/ or /k/. The opposite is also found in writing systems, where a single grapheme represents two or more phonemes, e.g. <x> in English orthography represents a combination of the phonemes /k/ and /s/. A **glyph** refers to a symbol with a particular shape.[10] It may correspond to a single grapheme or multiple graphemes. A **diacritic** is a mark, or series of marks, that may be above, below, or through glyphs. Diacritics are sometimes used to distinguish homophonous words and are more often used to indicate a modified pronunciation (Daniels and Bright, 1996, xli).

2.1.4 Technological terminology

On personal computers, "exotic" writing systems and phonetic transcription systems were long constrained to the American Standard Code for Information Interchange (ASCII) character encoding scheme, which meant that users could either use and adopt the (extended) Latin alphabet or they could utilize the small number of code points in ASCII to assign new symbols to its code points as rendered and defined in a different font.[11] To alleviate

[9] See Kohrt 1986 for a historical overview of the term grapheme.

[10] The Unicode Standard makes a distinction between glyphs and characters. A *glyph* is a concrete representation of a character when rendered with a font. A *character* is an abstract representation of a grapheme and is represented by a code point. See Section 2.1.4.

[11] See Section 2.3.5.

this problem, the Unicode Consortium set itself the ambitious goal of developing a single universal character encoding to provide a unique number, i.e. a code point, for every character in the world's writing systems.[12] In this work I adhere to the Unicode Standard for encoding linguistic data and I use some of its jargon.[13]

The term **character** refers to the basic unit for encoding a Unicode character. The Unicode Consortium (2007) defines a character as either:

1. The smallest component of written language that has semantic value; refers to the abstract meaning and/or shape, rather than a specific shape (see also glyph),[14] though in code tables some form of visual representation is essential for the reader's understanding.

2. Synonym for abstract character.[15]

3. The basic unit of encoding for the Unicode character encoding.

4. The English name for the ideographic written elements of Chinese origin.

Unfortunately, the term character can be quite confusing due to its alternative definitions and because in general the word character means different things to different people. A Unicode character is an abstraction of a set graphemes that are encoded as a single unit of information for representing textual data. Unicode defines the term grapheme as:

1. A minimally distinctive unit of writing in the context of a particular writing system.

[12] A **character encoding** represents a range of non-negative integers called a **code space**. A **code point** is a unique non-negative integer within a certain range, or in other words, a code space. An abstract character, for example a LATIN SMALL LETTER P, is then mapped to a particular code point such as U+0070. That encoded character is rendered on a computer screen (or printed) as a glyph depending on the font and the context in which the character appears.

[13] The glossary of Unicode terms resides at: `http://unicode.org/glossary/`.

[14] Unicode defines *glyph* as: "(1) An abstract form that represents one or more glyph images. (2) A synonym for glyph image. In displaying Unicode character data, one or more glyphs may be selected to depict a particular character. These glyphs are selected by a rendering engine during composition and layout processing."

[15] Unicode defines *abstract character* as: "A unit of information used for the organization, control, or representation of textual data."

2. What a user thinks of as a character.

Whereas a grapheme is a minimally distinctive unit in a particular language-specific writing system, Unicode does not encode different characters (think graphemes) for different languages. For example, on the one hand English, French and German have the same code point for <j>, even though each is pronounced differently and belongs to a different writing system.[16] They all, however, belong to the same script. On the other hand, the characters rendered as <p> and <р> are assigned different code points because they belong to different scripts, even though they are homoglyphs; the former is a LATIN SMALL LETTER P at code point U+0070 and the latter a CYRILLIC SMALL LETTER ER at U+0440.

Confusion ensues because the Unicode Consortium's decisions regarding characters and code points can sometimes be seen as going against this principle of grapheme abstraction. Unicode says it encodes characters and not glyphs. For example, <g>, <*g*>, <**g**>, <***g***>, <g>, <g>, <g>, <g>, and so on, are different glyphs of the same character.[17] However, in the IPA Extensions block,[18] there are several characters that could be considered glyphs, or variants, of the same grapheme in the Latin block, e.g. <ɑ> vs <a> and <g> vs <g>.[19] Nevertheless, other characters like <p>, <ŋ>, <β> do not appear in the IPA Extensions block; they are already encoded in the Basic Latin, Latin Extended-A, and Greek and Coptic blocks. Thus when a linguist transcribes an IPA <p> on a QWERTY keyboard, it is valid Unicode IPA. However, keyboard <g> and <!> are not. These symbols require insertion of "special" characters <g> and <!> because they belong to the IPA Extensions block. I discuss the problems and challenges of adhering to Unicode IPA in detail in Section 2.3.5.

Unicode defines a set of characters that are abstractions of graphemes, but it does not

[16] Unicode defines *writing system* as, "A set of rules for using one or more scripts to write a particular language. Examples include the American English writing system, the British English writing system, the French writing system, and the Japanese writing system."

[17] `http://www.macchiato.com/unicode/globalization-gotchas`

[18] In Unicode a *block* is a grouping of related characters. A block typically contain characters from a single script, but some scripts are encoded in different blocks.

[19] Glyph variants of different characters may result in **homoglyphs**, i.e. a set of glyphs with shapes that are either identical or are beyond differentiation by swift visual inspection, as illustrated in these examples.

provide visualizations for these characters. A **glyph** is a graphical representation of a character as it appears when rendered (or rasterized) and displayed on an electronic device. Each character can be displayed by a glyph in a font that supports that character. A **font** is comprised of a repertoire of glyphs.

A glyph's rendering is dependent on its font and its context within in a word. For example, the Unicode character LATIN SMALL LETTER G is rendered with the glyphs <g> and <g> in the Computer Modern and Courier fonts because their typefaces are designed differently. Characters in writing systems like Hebrew and Arabic have different glyphs depending on where they appear in a word. For example, some letters in Hebrew change their form at the end of the word, and in Arabic, primary letters have four contextually-sensitive variants (isolated, word initial, medial and final). In Unicode these different glyphs are encoded by a single character and it is the font that determines how they look when displayed.

Technologically, we must distinguish between characters and glyphs because:

1. There is not always a one-to-one mapping between characters and glyphs.

2. The logical order of a sequence of characters may not be the same as the visual order of their glyphs.

As noted above, a single character may have different contextually determined glyphs. However, a single character may also result in a sequence of multiple glyphs. For example, in Tamil one Unicode character may result in a combination of a consonant and vowel, which are rendered as two adjacent glyphs by a font that supports Tamil. A multiple character sequence may also result in a single glyph. For example in this thesis I use LaTeX, a typesetting system that by default combines the two characters <f> and <i> into a single glyph <fi> through a process called glyph substitution. When two or more glyphs are conjoined into a single glyph, the result is called a **ligature**.

Characters are stored in a computer's memory and must be mapped to glyphs to render text. The order in which characters are stored in memory is called logical order. In Unicode the visual order of glyphs may not be the same as the logical order of their characters, i.e. contiguous display is not indicative of contiguous text. Although in some cases this difference is encoded in the Unicode standard, in others it may be due to the order in which users have inserted a sequence of characters. For example, phonetic characters with certain combinations of diacritics may be homoglyphs, even while the logical order of their character sequences are non-equivalent.[20] Thus some type of standardization, or what Unicode calls *normalization*, of the logical ordering of characters is required to make sure that all data are logically consistent and therefore comparable and equally searchable. Standardization is a step towards data interoperability.

In this work I use the term **standardization** to refer to the process of transforming some object so that it conforms to a particular standard. For example, adherents of the Americanist Phonetic Alphabet (APA)[21] transcription system use symbols such as <y> and <č> to represent the palatal glide and voiceless alveopalatal affricate, respectively. In the International Phonetic Alphabet (IPA), <č> has no defined meaning and the symbol <ʧ> is used instead for the voiceless alveopalatal affricate. The symbol <y> is also used in IPA, but it represents a high front rounded vowel. Different standards are simply followed by different communities. My point here is that each standard serves the same purpose: to provide a standardized system for phonetic transcription, which allows the transcriptions of various languages in the same system to be easily understood and compared. All systems provide a mechanism to make data sets interoperable, or in other words, mutually intelligible. In this work I have standardized all transcriptions into IPA and into a set of distinctive features, so that all symbols from all sources adhere to one standard and can be easily compared by using that standard or an ontological mapping to that standard.[22] Another example of standardization used in this work is mapping language names used in language descriptions to ISO 639-3 unique language name identifiers. This allows data from different resources that describe the same language with different language names to be identified as different descriptions of the same language. I discuss issues regarding standardization in Sections 2.3.4 and 2.3.5.

[20] One example is a vowel that is both nasalized and creaky voice, e.g. <ḛ̃>. See discussion below.

[21] APA goes by various names; I have simply chosen one.

[22] For a mapping of APA to IPA symbols, see Odden 2005, 34-37.

The aim of standardization is to attain interoperability of data. By **interoperability** I mean the ability to ubiquitously exchange and merge disparate data sets, and data encoding formats, to facilitate data sharing and to "effortlessly" undertake comparison and analysis. Interoperable data should be integrated, shared and exchanged in a transparent way. Attaining interoperability in this work requires standardizing segments at both the linguistic and technological levels. For example, interoperability of linguistic data at the transcription level requires standardizing segments from different transcriptions, especially idiosyncratic ones, into one explicit standard transcription system. To attain interoperability of linguistic data at the technological level, segments must adhere to a set of Unicode characters, the code points of which must adhere to a standardized logical order.

Normalization has two distinct and mutually exclusive meanings in this work.[23] First, normalization is a term used by The Unicode Consortium (2007) to refer to:

> "A process of removing alternate representations of equivalent sequences from textual data, to convert the data into a form that can be binary-compared for equivalence. In the Unicode Standard, normalization refers specifically to processing to ensure that canonical-equivalent (and/or compatibility-equivalent) strings have unique representations."[24]

In other words, there are equivalent sequences of Unicode characters that can be normalized, i.e. transformed, into a unique Unicode-sanctioned representation of a character sequence called a *normalization form*.[25] Data preprocessing to achieve interoperability requires strings of characters to be normalized. There are different types of normalization forms in Unicode. Consider the characters in 1-3:

1. <Å> LATIN CAPITAL LETTER A WITH RING ABOVE (U+00C5)

[23] Sometimes the term *normalization* (or *to normalize*) is also used to mean standardization. This sense is co-opted from statistics, where it means to remove statistical error from a measured data set, to refer to the process of standardizing disparate data. Note also that sometimes the term *normalize* is used to mean *standardize* (cf. Hyman 2008, 85). In this work I will stick to **standardize** for transforming objects into a standardized form, unless I am referring specifically to Unicode normalization or database normalization.

[24] `http://unicode.org/glossary/`

[25] See discussion and examples in Sections 2.3.5 and 4.3

2. <Å> angstrom sign (U+212B)

3. <Å> latin capital letter a (U+0041) + < ˚ > combining ring above (U+030A)

The character <Å> is represented in Unicode in the first two examples by single-character sequences and in the third example by a multiple-character sequence. All three sequences are canonically equivalent, i.e. they have the same appearance when displayed. However, they are logically different. If one were to search a text for angstrom sign (U+212B), instances of latin capital letter a with ring above (U+00C5) would not be returned.

The first of the three <Å> characters is considered the Normalization Form C (NFC), where "C" stands for composition. When the process of NFC normalization is applied to the character sequences in 2 & 3, both sequences are normalized into the character sequence in 1. Thus all three canonical character sequences are standardized into one composition form in NFC. Another Unicode normalization form is the Normalization Form D (NFD), where "D" stands for decomposition. When NFD is applied to the three examples above, all three, including importantly the single-character sequences in 1 & 2, are normalized into the decomposed multiple-sequence of characters in 3. Again, all three are then logically equivalent and therefore syntactically interoperable.

In this work I normalize all strings into NFD because each character in a segment has phonetic value and by using NFD all characters are decomposed into a standardized order. For example, a vowel that is both nasalized and creaky looks like <ṽ̰> in IPA. Although visually the same, a nasalized and creaky vowel can be composed of several different character sequences, as illustrated with <õ̰> in 1-3:

1. latin small letter o + combining tilde + combining tilde below
   U+006F + U+0303 + U+0330

2. latin small letter o + combining tilde below + combining tilde
   U+006F + U+0330 + U+0303

3. LATIN SMALL LETTER O WITH TILDE + COMBINING TILDE BELOW
   U+00F5 + U+0330

Applying NFD to these three character sequences results in one standard sequence; in this case the character sequence given in 1. NFD makes different sequences of input interoperable and it retains all of the phonetic information captured by the separate characters that combine to form a vowel with nasalization and creaky voice phonation. Regardless of how someone may have entered the segment on a computer, all three are treated equivalently after normalization and each part of the phonetic transcription signal is analyzable and queryable.

The second sense of **normalization** refers to a specific aspect of relational database design. In the broadest sense, a **database** is simply a mechanism that stores data, e.g. an address book or library catalog. The term database is now primarily used to refer to a set of data, often a collection of related data, stored electronically in a computer. A **relational database** is a set of tables joined, or related, in a standardized way (Codd, 1970). A **table** is a two dimensional data representation that consists of columns and rows. Data are stored in cells in the table. A row represents a particular entry and column represents a data type shared by those rows.

Database normalization encompasses the design principles for organizing data into tables to minimize duplication of data across related tables. It is a modeling technique used to optimize database performance by reducing data redundancy. The database's design can be evaluated by whether or not it adheres to one of several *normalization forms*.[26] Another important process is called **denormalization**, which means to remove normalization forms. This process typically reduces the number of tables in the database and it intentionally introduces data redundancy that often results in much simpler database queries, but at the cost of performance.

A **database schema** is a description of the structure of a database in a formal language that is supported by a database management system (DBMS). A DBMS is software that performs database functions such as storing, accessing and modifying data. A relational

[26] See discussion in Section 3.2.1.

database management system (RDBMS) is a DBMS that is based on the relational model by Codd (1970). In Section 3.2.1 I describe the relational database that I created for the PHOIBLE data by using MySQL, an RDBMS. To illustrate my relational database design, I use an extended entity-relational model (EER) to diagram the entities and their relationships in my database schema. My EER diagrams use a notation called Crow's Foot, developed by Everest (1986).[27] A description of a relational database's schema allows users to formulate queries and operations on the database. The Structured Query Language (SQL) is a standardized language that is used to create, update and retrieve data in tables and databases. There are several implementations of SQL; each is dependent on the RDBMS that it uses.

A relational database is one information model for storing, accessing and manipulating a data set. A **data warehouse** is a copy, or in other words a *data dump* or *data export*, of transactional data restructured for query and analysis. Data warehousing is the process of creating and maintaining a data warehouse (Kimball, 1996). The distinction between a relational database and a data warehouse lies in their different purposes. A database is often designed for transactions, i.e. data are added, removed or updated. A data warehouse is a snapshot of data from the relational database. It contains a (sub)set of data structured for query and analysis for particular tasks. For example, in Chapter 3 I will explain how I designed a relational database to bring together different data sets into one resource. The design of my relational database, however, follows principles of database normalization to reduce data redundancy. This makes querying the relational database pretty complicated. To make the data more easily accessible, I create a data warehouse by denormalizing the relational database into a flat table that is easily queryable and human readable.

In addition to relational database technologies used in this work, I also use several Web standards developed by the World Wide Web Consortium (W3C).[28] One is the Extensible Markup Language (**XML**; Bray et al. 1998). XML is a text-based format for encoding documents for representing and transmitting machine-readable information. It is a markup

[27] See Section 3.1.2 for details.

[28] http://www.w3.org/

language like HTML, except that XML is designed for representing the structure of documents, not their appearance.[29] Like XML, the Resource Description Framework (**RDF**; Lassila and Swick 1999) is also a model for data interchange, but whereas XML models data in a tree structure, RDF encodes representations of knowledge in a graph data structure by using sets of triples (also called statements). For example the triple (German, hasPhoneme, a) represents a statement that indicates the object "German" is in a "hasPhoneme" relation with the object "a". RDF is a graph data model for specifying resource objects and the relations that hold between them. XML and RDF formalisms have different strengths and are used in different applications.[30] To confuse matters, RDF data models can be serialized in XML.[31] Whereas XML imposes no semantic constraints on the data it encodes, RDF was developed to represent knowledge so that information can be queried to extract "meaning" by inferring additional statements through implicit relationships that are encoded via logic statements encoded in predicates.[32]

RDF falls under the often misunderstood *Semantic Web* (Berners-Lee et al., 2001). The Semantic Web is a set of technologies, tools and standards that provide digital architecture to address complex data compatibility issues.[33] The term "semantic" often conjures up confusion because it is used to denote a range of ideas. Essentially the Semantic Web is a web of data that can be accessed using Web architecture and technologies in a range of application areas including data integration, resource discovery and sharing.[34] The goal is a common framework for sharing and reusing data that can be processed by both human inspection and by automated tools that leverage advances in knowledge representation. To accomplish these tasks, data (aka resources) need to be described and marked-up with logic

[29] XML is also used to encode arbitrary data structures in web services (application programming interfaces accessed through HTTP).

[30] For a comparison of the different RDF and XML models, see `http://www.w3.org/DesignIssues/RDF-XML.html`.

[31] Serialization is the process of converting an object or data structure into a format, or sequence of bits, that can be later converted back to its original format with equivalent properties.

[32] I provide more detail about data modeling in Section 3.1 and knowledge representation in Section 3.3.

[33] There is much criticism of the Semantic Web, see for example Marshall & Shipman 2003.

[34] The W3C provides a growing list of Semantic Web case studies at: `http://www.w3.org/2001/sw/sweo/public/UseCases/`.

annotation. One component is the use of the Universal Resource Identifier (URI). A **URI** is a formatted string that provides a unique identifier for a resource. URIs identify physical or abstract resources and they are used for the subject, predicate and object of the triples encoded by RDF. URIs hold the key to addressability as they are unique namespace identifiers that eliminate naming conflicts. A URI can be further classified as a Uniform Resource Locator (URL), a reference to an Internet location, or as a Uniform Resource Name (URN), an abstract unique name that remains persistent and is used for identification of a resource even when it ceases to exist. URIs may or may not be dereferenceable.[35] A dereferenceable URI is a resource retrieval mechanism that uses an internet protocol to retrieve a representation of the resource it identifies. The type of representation is determined via content negotiation, a mechanism defined in the HTTP specification that determines which version of a document to serve, e.g. a human readable webpage or a machine readable format intended for computer processing, like RDF. In a non-dereferenceable context, such as when a namespace URI is used in an XML Schema, the URI is simply a unique identifier that is not intended to be dereferenceable via HTTP. RDF based vocabularies include RDF Schema (RDFS) and the different flavors of the Web Ontology Language (OWL). RDFS provides the specification of precise semantics for describing the basic elements of an ontology. **OWL** is a more expressive ontology language for processing information than RDFS. An **ontology** exactly describes information in a domain model and consists of statements about concepts (*resources* in Semantic Web speak), their relations and constraints on those relations. Like RDF, OWL is a W3C standard and can be serialized in XML, as well as other formats. It currently has three increasingly expressive sublanguages: OWL Lite, OWL DL and OWL full. **Description Logics** (DL) are a family of structured languages based on computationally tractable fragments of first-order logic (Baader et al., 2003). They provide the logic formalism for ontologies used in the Semantic Web. For example, OWL DL (literally "Web Ontology Language Description Logic") supports ontology development by providing the meaning representation language to formally specify the semantics of a domain of interest with the guarantee of computational completeness, i.e. all conclusions are computable and

[35] In computer science, a pointer references an address (location) in memory where a value is stored. Dereferencing refers to obtaining the value at that location that the pointer refers to.

decidable in a finite time (Smith et al., 2004).

2.1.5 Abbreviations

I refer to several projects throughout this work by abbreviated names. The Stanford Phonology Archive is referred to as **SPA** (Crothers et al., 1979). The UCLA Phonological Segment Inventory Database is referred to by the commonly used acronym **UPSID**. The original UPSID database contained a sample of 317 languages and is referred to as UPSID_{317} (Maddieson, 1984). Maddieson and Precoda's (1990) extended UPSID database with 451 languages is referred to as UPSID_{451}. Where the distinction is irrelevant, I simply use UPSID. For Hartell's (1993) *Alphabets des langues africaines* (Alphabets of Africa), I use the abbreviation **AA**. I also use AA to refer to Chanard's (2006) digitization and online implementation of Hartell's AA.[36] The cross-linguistic data set produced in this work is referred to as **PHOIBLE** for PHOnetics Information Base and LExicon. Each of these resources is described in detail in Chapter 4. Additional information about languages, such as genealogical and geographic data, comes from the World Atlas of Language Structures, commonly referred to as **WALS** (Haspelmath et al., 2008).

2.2 Linguistic theories

In phonetics and phonology, there is a long tradition of representing spoken language as strings of symbolic units. The roots of this theoretical framework are found in work of the ancient Sanksrit grammarian Pāṇini.[37] Pāṇini's descriptive grammar of Sanskrit uses a sophisticated system of rules and representations and it is regarded as the first work to describe the phoneme-allophone relationship. Pāṇini's work influenced structuralists (e.g. Bloomfield 1927) and their approach to segmental phonology that used alphabet-inspired symbols for encoding articulatory steady states. His work also influenced the development of generative phonology (Chomsky and Halle, 1968), in which segments are phonological representations that consist of distinctive features (Jakobson et al., 1952; Jakobson and

[36] `http://sumale.vjf.cnrs.fr/phono/`

[37] See discussions in Kiparsky 1979 and Anderson 1985.

Halle, 1956). In this section I provide a very brief overview of segmental phonology and distinctive feature theory, before discussing the challenges of modeling these theories in a typological database in Section 2.3.

2.2.1 Segmental phonology

Phoneticians have long used classification systems for describing speech sounds. In the late 19th century, speech sounds were modeled as discrete segments (e.g. Bell 1867, Sweet 1881 and Passy 1888). The advent of the kymograph, an instrument that records variations in pressure, and adoption of the scientific method led to the discovery that a sound's pronunciation varied greatly and that segment boundaries indeed do not appear in the continuous speech stream (Sievers, 1876; Rousselot, 1897; Scripture, 1902). However, in phonological theory, phonological units were to remain segmental, abstract, invariant and sequential.[38]

Segmental phonology is the study of speech sounds modeled as abstract segments that are discrete and serially ordered. It investigates the distribution of sounds and their patterning by means of a theoretical framework that strives to answer questions regarding the nature of phonetic alternations and contrastive sounds that trigger lexical or grammatical differences in languages.

Each spoken language can be described with a language variety-specific set of segments, which it uses to form and differentiate words. The two types of relations, *paradigmatic* and *syntagmatic*, are concerned with the substitutability of a segment in a particular position in a word, and with the positioning of segments in a word, respectively.

The paradigmatic role of segmental phonology is to describe the vertical relations that hold between segments that appear in the same environment. For example, /dæd/ "dad" and /bæd/ "bad" are two words that contrast to form a minimal pair in English. These two words are contrastive by their first segments' place of articulation, a feature that causes /b/ and /d/ to be interpreted as distinct sounds by the listener.

Segmental phonology is also concerned with the language-specific relationship between an underlying and abstract symbolic phoneme, its set of its surface-level allophonic variants,

[38] For an overview, see Osterhout et al. 2007.

and the phonological and morphological environments that trigger these variations. This is the syntagmatic role of segmental phonology, i.e. to investigate the horizontal relations between segments. For example, in Western Sisaala [ssl] the first person pronoun *n* assimilates to the place of articulation of the following morpheme's initial consonant phoneme (Moran, 2008). The underlying first person pronoun is posited as /n/ because it occurs on the surface level in the most environments, which includes [n] before vowels. This process is captured in the phonological rule in 2.1 and examples are given in 2.2-2.5.

(2.1) $[\mathrm{N}] \rightarrow [\alpha \mathrm{N}] \; / \_ \; [\alpha \text{ place of articulation}]$

(2.2)

n	*tummi*	*sɪnkan*
1S	chew	groundnuts

"I chewed groundnuts."

(2.3)

m	*ballo*
1S	hunt

"I hunt."

(2.4)

ŋ	*kiɛrɛn*
1S	sit

"I sit."

(2.5)

n	*e-o*	*pa*	*koʤo*
1S	made-3S	for	Kojo

"I made it for Kojo."

The study of the paradigmatic and syntagmatic relations between segments of a language allows the linguist to posit a segment inventory that describes (and is used to describe) aspects of that language's phonological system. Cross-linguistic comparisons of segment inventories provide insights into the phonetic factors that shape the range of all languages' phonological systems. It has long been noted that not just any set of consonants and vowels can make up a segment inventory (Sapir, 1925). Certain sounds and certain combinations of sounds also occur more frequently than others in the languages of the world (Maddieson, 1984). Where similarities occur across unrelated languages, this suggests there are factors that cause segment inventories to be similar in ways other than shared descent, such as language contact via areal proximity. The frequency and distribution of segments may also reflect non-linguistic factors such as violent and non-violent human interaction that has affected which languages have survived and in which language families (Mielke, 2009). There is also a growing body of research investigating other non-linguistic factors, such as ecology, climate, demography and genetics, and their possible effects on phonological systems and their structure.[39]

Note however that since their creation, segments (ergo segment inventories) and the use of segments as a theoretical construct have faced controversy. Even after advances in technology showed that segment boundaries do not exist and that each instance of a pronunciation differs measurably, phonological theory continued to model phonological systems with segments. Mielke (2009, 700) notes that, "Just about every aspect of defining a segment inventory for a language is controversial, from whether it is appropriate to divide words into segments in the first place, to how segments should be represented, to what they represent."[40] Nevertheless, research in segmental phonology led to modeling segments with sets of features, which has provided linguists with a theoretical framework that allows them to elegantly describe many of the phonological processes that appear in the world's languages. Segmental phonology became a serious avenue of research for phonological theory and was integral in the development of distinctive feature theory and Generative Phonology (Chomsky and Halle, 1968).

Although there are non-segmental formalisms of phonological theory, e.g. Articulatory Phonology (Browman and Goldstein, 1986, 1989, 1992) and Firthian Prosodic Analysis (Firth, 1957; Palmer, 1968), in this work I limit my investigation to the computational modeling of segmental phonology. Segments offer a finite set of phonological representations and are used in linguistic descriptions to document the contrastive sounds employed by languages. Segments are phonetically defined by the IPA and are represented in the Unicode standard. Therefore, there exists a standard for transcribing segments (researchers'

[39] See Chapter 7.

[40] See Section 2.3.4.

idiosyncratic transcription systems can be mapped to the standard to achieve interoperability), a standard for encoding this set of segments computationally, and a standard for comparing the different sounds in different languages typologically because of the internal structure of the IPA system. IPA symbols are also convenient abbreviations for the set of distinctive features that constitute a segment.

2.2.2 Distinctive feature theory

Even as X-ray analysis of speech gestures and spectrographic analysis of acoustic patterns in the speech signal emerged in the 1940s and early 1950s, distinctive feature theory was becoming a serious avenue of research for phonological theory. Distinctive feature theory emerged and defined the features (or parameters) for labeling sets of sounds, e.g. "the set of voiceless sounds" or "the set of voiceless velars". This formalism allows linguists to generalize about regularly occurring phonological patterns and to describe the behaviors of sets of sounds with predictive power, thus informing phonological theory (e.g. "in German all voiced obstruents devoice in syllable final-position").

Distinctive feature theory is considered one of the most important contributions to linguistics in the 20th century because of the explanatory power that it provides. It has a long tradition in linguistics, in such works as Trubetzkoy 1939, Jakobson 1949, Jakobson et al 1952 and Jakobson & Halle 1956.[41] By building on the work of members of the Prague Linguistic Circle (or Prague School) and the American structuralists in the early to mid 20th century, Noam Chomsky and Morris Halle created generative phonology.[42] Although several of their works led to its development (e.g. Halle 1962; Chomsky 1964; Chomsky and Halle 1965), *The Sound Pattern of English* (SPE) is the first full systematic exposition and magnum opus of generative phonology (Chomsky and Halle, 1968). In generative grammar, phonological representations were modeled as sequences of segments composed of distinctive features. This provided a framework for phonologists to describe phonological rules and derivations, and levels of phonological representations through fully explicit algorithms

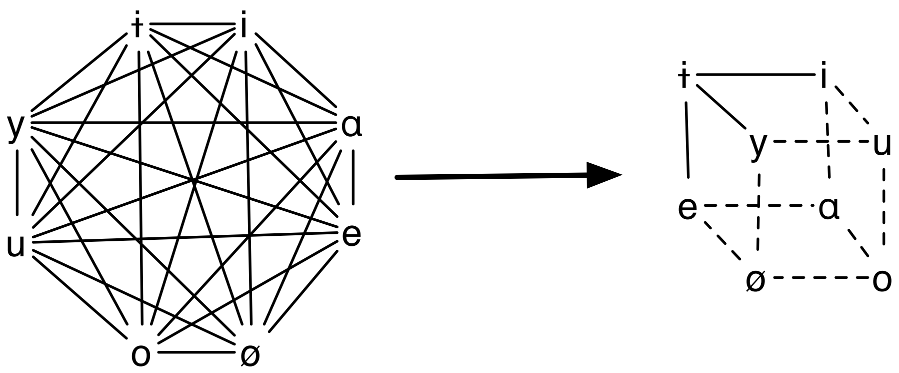
[41] See Baltaxe 1978 for an account of the development of distinctive feature theory as a conceptual framework.

[42] See Goldsmith and Laks, to appear, for a historical review of generative phonology.

using linear sequences of matrices of feature values.

Distinctive features represent abstract properties of speech sounds, typically modeled on phonetic correlates rooted in human anatomy. The mental representation of a speech sound was originally modeled as an unorganized set of feature values.[43] Two speech sounds contrast if they differ by at least one distinctive feature. Jakobson's approach was to keep the number of distinctive features at a minimum (e.g. Jakobson 1949). For example, an eight vowel system requires 28 binary relations if each vowel opposes every other vowel. These 28 binary oppositions can be expressed in terms of three distinctive features (e.g. [high], [back] and [round] in SPE), resulting in only three oppositions, as illustrated in Figure 2.2. This approach reduces entropy, so that there is less functional load involved in the storing and processing of language for the speaker and listener.

Figure 2.2: Reduction of oppositions with distinctive features (Mielke and Hume, 2006, 723)



In the work of Jakobson et al. (1952), distinctive features were almost exclusively acoustically defined. However, in the years following the feature set proposed by Chomsky and Halle (1968), articulatory features have come to predominate. More recently, distinctive features include both articulatory and acoustic features. On the one hand, the features

[43] Although features started off in distinctive feature theory without a notion of distance, much research has shown the value of viewing segments as made up of hierarchically structured features. For example, Clements (1985) formulated features into constituent structures with internal organization, much like syntactic trees. This tree model was in part motivated by groupings of features that commonly pattern together, especially in rules of partial assimilation.

[bilabial], [dental], [plosive], [fricative], [round], etc., are grounded in articulatory phonetic factors that involve forming constrictions in the human vocal tract with speech organs like the lips, tongue, teeth, etc. On the other hand, vowel features including [high] and [back] are better defined in the acoustic perceptual realm. For example, taken together the three features of [high], [back] and [round] describe the tongue's position within the acoustic space of the mouth cavity and the articulatory constraint of lip rounding. A feature matrix for an eight vowel system using these three binary distinctive features is given in Table 2.1.

Table 2.1: Feature matrix

	i	y	ɨ	u	e	ø	o	ɑ
high	+	+	+	+	−	−	−	−
back	−	−	+	+	−	−	+	+
round	−	+	−	+	−	+	+	−

The feature matrix expresses the contrasts between speech sounds by their distinctive features. The matrix can be used to calculate how much two segments differ by summing up the oppositions of their features. The complexity (and plausibility) of a phonological change is formalized as the modification of the values of a (set of) distinctive feature(s). Another critical function of distinctive features is that they make possible the formal study of natural classes, i.e. sets of sounds that have certain phonetic features in common. Natural classes form groups of sounds that share a set of one or more features to the exclusion of all other sounds in a particular language.[44] Sounds in a natural class behave the same way in the same environment and they affect other sounds that share the same environment in the same way. Natural classes also tend to participate in phonological processes that often pattern similarly across languages. For example, it is widely attested in languages that the

[44] The specificity of a class is related to the number of features used to define that class (or inversely, the generality of a class is related to the inverse number of features used to define that class). For example, in Table 2.1 the natural class of high vowels includes the set { i, y, ɨ, u }. The class of high back vowels is { ɨ, u } and the class of high back round vowels includes only { u }.

natural class of voiced obstruents devoice at the end of a word (obstruents are a natural class made up of the natural classes of stops and fricatives). This phonological pattern seems to be rooted in articulatory effort; it requires more effort to maintain voicing when a voiced obstruent is not followed by a vowel.

Like segments, distinctive features play both a paradigmatic and a syntagmatic role in a language's phonology by defining the contrasts in a language's sound inventory and by formalizing its phonotactics, i.e. rules governing the possible combinations of phonemes.[45]

From a paradigmatic perspective, distinctive features play a role in governing and structuring the structure of speech sound inventories. As outlined in Clements 2009, there are several feature-based principles that constrain the structure of contrastive speech sound inventories. For example, the Feature Bounding principle states that given a set of n binary distinctive features, a language may have a maximum of 2^n distinctive sounds. In the example in Figure 2.2, a distinctive feature set using 3 binary features may have maximally 8 sounds (2^3). This feature-based principle constrains the upper limit on the number of contrastive sounds in a language, based on its number of distinctive features. This principle also claims that the upper limit on the number of possible contrasts (C) is set by the number of features, as given by the equation C = (S * (S -1)) / 2 (Clements, 2009, 25). Since the maximum number of sounds (S) is 2^n, the maximum number of contrasts is (2^n * (2^n - 1)) / 2. Thus, the Feature Bounding principle constrains a sound inventory with two features to a maximum of four sounds and six contrasts.[46]

From a syntagmatic perspective, words in a language are made up of a string of segments with each segment consisting of a set of features, as shown in Table 2.2.[47] In English the contrast in the place of articulation feature in these two words, here referred to as labial,

[45] For example, many languages, like Russian [rus], permit clusters of consonants only if they all have the same feature for voicing, while other languages, such as Tsou [tsu], permit combinations of voiced and voiceless elements in the same cluster (Wright, 1996).

[46] Other feature-based principles examined in Clements 2009 include: Feature Economy (tendency to maximize feature combinations; see de Groot 1931, Martinet 1955; 1968 and Clements 2003a; 2003b), Marked Feature Avoidance (tendency to avoid marked feature values), Robustness (in a universal hierarchy of features, languages draw higher-ranked features before lower-ranked features) and Phonological Enhancement (increasing the acoustic difference between contrasts).

[47] The features used here are a subset of those defined in Hayes 2009 and include zero as a value for features that aren't relevant to a particular sound.

triggers a meaningful lexical contrast.

Table 2.2: Feature representation of the words “bad” and “dad”

	b	æ	d	d	æ	d
voice	+	+	+	+	+	+
labial	+	−	−	−	−	−
consonantal	+	−	+	+	−	+
high	0	−	0	0	−	0
back	0	−	0	0	−	0

Distinctive feature theory expresses the architecture of phonological segment inventories. Therefore a distinctive feature set should characterize all contrastive sounds in all languages.[48] The number of distinctive features is specified by the distinctive feature theory that employs them, but in general theories that have been proposed have around two dozen features (Mielke and Hume, 2006). This small number of distinctions has proven useful and has allowed linguists to make predictions about sound structures, sound patterns and the cognitive organization of sounds in languages. Several distinctive feature sets, or portions of sets, exist and they differ in their classification and descriptions of segments. These works include, but are not limited to: Chomsky and Halle 1968, Sagey 1990, Goldsmith 1990, Clements and Hume 1995, Ladefoged and Maddieson 1996, Ladefoged 1997 and Hayes 2009.

2.2.3 Summary

To summarize, speech sounds have long been modeled as abstract segments. The analysis of phonological segments as sets of features is considered one of the great advances of linguistic research in the 20th century. The premise of distinctive feature theory is that each

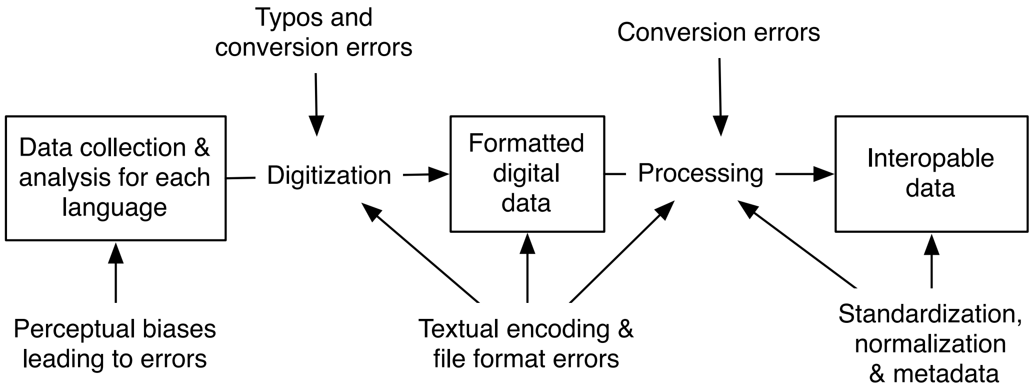
[48] See Section 6.3.

phoneme is composed of a matrix of (binary) features that can be used to encode similarities, differences and classes of sounds. Distinctive feature theory provides a framework for modeling features, segments and phonological patterns. In the next section I describe the challenges involved in creating a cross-linguistic data set situated in segmental phonology and distinctive feature theory.

2.3 Challenges

In this work I have faced both theoretical and technological challenges in developing a cross-linguistic segment inventory data set that is accessible through different technologies in order to investigate questions of phonological typology. Within linguistic theory, there are arguments about what constitutes typological categories and how they can be compared. These are non-trivial issues that typologists will continue to debate far into the future. In my work these issues revolve mainly around the notion of phoneme and the assumption that segments and distinctive features are linguistic entities that can be compared cross-linguistically. At the technological level, there are many challenges involved in creating an interoperable digital resource to store and access descriptive linguistic data. Both types of challenges are present in the workflow illustrated in Figure 2.3.

Figure 2.3: Conversion workflow


Typos and conversion errors
Conversion errors
Data collection & analysis for each language
Digitization
Formatted digital data
Processing
Interopable data
Perceptual biases leading to errors
Textual encoding & file format errors
Standardization, normalization & metadata

The workflow begins with the field linguist's collection and analysis of language data.

Typically the linguist makes an impressionistic study through transcription and phonemic analysis (as opposed to an in-depth acoustic analysis of the speech stream). This is an area of theoretical debate. Can impressionistic data be trusted? Can these data, coming from many different linguists, be typologized (cf. Sherman and Vihman 1972; Haspelmath 2010)?

Moving a step further through the workflow, for the data to be made widely available, the field linguist's data and analysis needs to be digitized. Digitization is another point where errors can be introduced into the data. The digitization process may include not only typos and misinterpretations by the digitizer (who may or may not be the original author), it also introduces computationally complex issues of character encodings, such as segment homoglyphy, which can affect any results or conclusions reached when querying and analyzing the data. For example, although two segments may be visually indistinguishable, they might in fact be encoded as two different characters computationally.[49] Finally, for the data from disparate resources to be made interoperable in the sense that queries can be made across the entire data set, the transcription and analysis of many idiosyncratic language descriptions must be standardized. Again this is a theoretical issue – to do typology, standardization of a linguistic data type is necessary if different language descriptions are to be compared. Transcription systems must also be standardized; segments must be resolved to equivalent characters within the same character encoding or they will not be computationally equivalent. Taken together, at the linguistic level the workflow is fraught with theoretical issues that are not easily resolvable, such as, do phonemes exist and can they be compared across languages? At the technological level, the workflow can propagate errors from the initial data collection stage, through the digitization and processing phases, and into a final data access and storage format. Lastly, there are issues at the intersection of linguistic theory and technology, such as using statistical sampling to address various biases inherent in the available typological data. In the following sections I discuss criticisms of cross-linguistic typological databases, statistical sampling, and the linguistic and computational issues involved in creating a data set for phonological typology.

[49] See Section 4.3.

2.3.1 Typological databases

Typological databases provide a tool to access and characterize the distribution of linguistic phenomena in the world's languages. However, there are at least two fundamental problems with making these characterizations. The first, raised by Sherman and Vihman (1972, 163), is the question of what constitutes adequate descriptive categories for linguistic phenomena and how can they be compared?[50] It is addressed in this section. The second problem involves statistical sampling and how to estimate the relative frequency of a linguistic type in light of typological biases like shared genealogical descent,[51] areal diffusion, and a lack of linguistic data for many of the world's languages. This second problem is discussed in Section 2.3.2.

Language documentation varies in its descriptive adequacy. In order to make cross-linguistic comparisons, linguistic analyses must be extracted from language descriptions. However, the comparative linguist should not typologize on the basis of descriptive linguists' analytical preferences (Hyman, 2008). Hyman argues that there is a paradox in using linguistic theory to describe languages because of the necessity in abstracting away from different linguistic theories to undertake typological comparisons. Therefore, criteria to normalize data need to be formulated to make cross-description categories comparable. But what constitutes adequate descriptive categories?

Instead of a set of universal cross-linguistic categories used for both language description and comparison, Haspelmath (2010) distinguishes between descriptive categories and comparative concepts. Descriptive categories are language-specific categories established by the linguist to describe phenomena in a particular language. These descriptive formal categories cannot be equated across languages because the criteria for their language-specific category assignment is different in each language.[52] Comparative concepts, on the other hand, are

[50] Sherman and Vihman (1972) may be the first to ask what are appropriate formats for storing and accessing descriptive linguistic data. This issue is discussed in Chapter 3.

[51] Throughout this work I will refer to the "genealogical" relationships between languages instead of "genetic" relationships, although the latter has been used quite frequently in the literature. This dichotomy makes clearer the split between research on the relatedness of languages versus research on the genetic relationships between human populations, which some claim affects language structure (cf. Dediu and Ladd 2007; Nettle 2007).

[52] For a rebuttal, see Newmeyer 2010.

categories created by typologists for undertaking cross-linguistic comparison. They are created by evaluating which descriptive categories from a set of languages can be compared. Haspelmath notes that in practice many linguists implicitly collapse the distinction between descriptive categories and comparative concepts.

In phonetics and phonology, language-particular descriptive categories are required to describe languages' phonological systems (Haspelmath, 2010). Port and Leary (2005, 927) argue that phonologies differ incommensurably and that the description of speech sounds cannot be tied to a universally fixed phonetic alphabet, noting that "decades of phonetics research demonstrate that there exists no universal inventory of phonetic objects". Their conclusion is that there is no discrete universal phonetic inventory with an a priori inventory of phonetic atoms. They are not the only researchers to position themselves against a Universal Grammar (UG) of phonological atoms. At the featural level, Mielke (2008) argues against an innate set of universal features and for an emergent distinctive feature theory. He claims phonological patterns are not reliant upon a fixed set of universally available features, but can emerge from language particular features and constraints.[53] Mohanan et al. (2009) take the argument against inherent features a step further and ask if all feature-based cross-linguistic comparison must be abandoned if UG does not contain predefined features. In their approach, to undertake phonological typology what is needed is "a theory of feature emergence that expresses the family resemblances of features, connecting the concrete aspects of the articulation and perception of speech to a cross-linguistically shared set of features" (Mohanan et al., 2009, 151). A cross-linguistically valid "currency of distinctive features" can be obtained without UG stipulating a universally pre-defined set. Whether speakers are born with a pre-determined set of defined features, or those features are emergent, or some type of hybrid of both, segment inventories nevertheless show symmetric regularities that can be described in terms of an economy theory of feature-based principles (e.g. Clements 2003a,b, 2009). To undertake phonological typology on segments and features, comparative concepts must be established.

For UPSID, Maddieson (1984) created comparative concepts for cross-linguistic compar-

[53] Emergent theories explain synchronic properties and observations in diachronic terms. See Blevins 2004.

ison of segment inventories by reinterpreting, where necessary, phonemes in phonological descriptions into (basically) IPA symbols.[54] In terms of comparative concepts, the IPA is a useful tool for cross-linguistic comparison, but not as a universal set for representing all possible sounds of the world's languages (Haspelmath, 2010). A database of segment inventories, like UPSID, can be used to answer questions about which contrastive consonants and vowels appear in which languages, or with what frequency a segment type occurs across languages in the sample.

Segment databases make several assumptions that have not gone without criticism. One assumption is the phoneme.[55] The basic principle of the phonemic method is that of contrast; two sounds contrast if they do not occur in complementary distribution. However, phonologists do not necessarily agree on how to do phonemic analysis and establish phonemic representations. The phoneme is an analysis of the set of allophones that minimally distinguish it from other phonemes, and is therefore a language-particular descriptive category. On the other hand, to create concepts for comparison purposes, the typologist has to take a stance on how contrastive segments are encoded. For example, Maddieson had to either go with the original phonemic analysis (in the resource descriptions from which he extracted segment inventories) or reinterpret those linguists' analyses according to some consistent standard to achieve uniform comparability across segment inventories.

Another assumption is the uniform comparability of segments. Simpson (1999) criticizes UPSID's interpretation of phonemes as abstract and contrastive segments. The problem boils down to choosing a single allophone to represent a phoneme, which is the typical methodology employed in positing a phonemic inventory. Simpson takes issue with this process, arguing that the comparison of phonemic inventories is of little use for qualitative and quantitative comparison and that "the phonetic interpretation of phonemic inventories may make them comparable, but tells us little about the languages they claim to be representing" (Simpson, 1999, 352). He argues that UPSID (and therefore inventories of contrastive segments like UPSID) fail to "recognize the abstract nature of even the most

[54] See Section 4.3.2 for a description of UPSID.

[55] For an early overview of different definitions of the term *phoneme*, see Twaddell 1935.

phonetically based definition of a phonemic system" (Simpson, 1999, 349). As such, Simpson suggests that phonological comparison is based on an arbitrary selection of the phonetic contrasts of languages in the database. He argues that this comparison misrepresents the abstract relational nature of the phonological system, thus "grossly oversimplifying the complex phonetic patterns employed in languages to bring about differences in meaning" (Simpson, 1999, 349).

These arguments have consequences for comparative and typological statements. Simpson asserts that "we still have no way of identifying sameness and difference in two phonological systems, a problem which is only apparently overcome by casting phonological contrasts in terms of a selection of features from a universal inventory" (Simpson, 1999, 349). An example supporting his point is Maddieson's categorization of a wide range of phonetically disparate sounds that are symbolized by "r-sounds" in UPSID.[56]

Simpson argues for a clear demarcation of levels, with each level requiring different types of analyses. Thus, "the unprincipled reduction of the complexity of linguistic sound systems severely weakens any qualitative or quantitative statements made using them" (Simpson, 1999, 352). Finally, he also takes argument with the use of features as specifications of contrasts (Simpson, 1999, 352):

> "Casting the phonological contrasts in a language in terms of universal feature specifications does not solve the problem any more than UPSID's system of phonetic classification. As there are no criteria for assigning the same feature to different phonetic patterns in two languages or even to assigning them to different sets of phonetics in the same language, the inventory of features becomes little more than a list of possible contrasts which must simply be large enough to capture the number of contrasts in a particular language. Stating that two languages have the feature [ATR] or [labial] is as trivial as stating that phonemes in two languages are symbolized with k or r."

Simpson concludes that comparative analyses using phonetic interpretations, such as

[56] See Section 2.3.4.

those undertaken with SPA and UPSID, are flawed and of little use in answering questions in phonetics and phonology (outside of its application as a reference for identifying languages that have a sound type or for calculating phonological complexity based on phoneme count).

However, I do not agree that using abstractions is of little use in doing phonological typology (or doing linguistics in general). Simpson's argument expands to any abstract analysis of language; the same argument can be leveled at phonemes, allophones and features because no two person's pronunciation is identical, nor does anyone say the same sound in exactly the same way twice in his or her lifetime. As scientists we must acknowledge the limitations of our analysis and interpret data with an appropriate level of coarseness. For example, with the PHOIBLE database we cannot say anything about language-specific factors relating to typology, such as the relative acoustic height of an /u/. There are phonetic details that get missed; this is a detail problem. How can someone characterize something as changing and variable as speech sounds?[57] Many acoustic and articulatory phoneticians believe that one cannot characterize speech sounds with discrete and invariant symbolic representations. However, note that even those researchers measuring individual muscle fibers must nevertheless employ some form of data reduction. On the other hand, from a quantitative perspective there is a problem of overfitting the model, i.e. putting so much detail into the model that it is modeling the detail and not the generalizations. As described elegantly in *Tao Te Ching* and also by Borges (1935) in "On Exactitude in Science": in making an observation, the medium used to describe the observation necessarily shapes and limits the observation.

In more recent criticism, Vaux (2009) disapproves of using UPSID as the empirical basis for phonological typological studies. He describes general problems with the UPSID data, including the use of "relatively arbitrary old grammars and articles", reported database coding errors including the omission of segments in certain languages, and "unwittingly imported phonetic and phonological errors from the source materials" (Vaux, 2009, 77-78).

From a phonetician's perspective, Vaux asserts that UPSID contains several significant phonetic mischaracterizations, which affect typological studies undertaken with UPSID. He

[57] And at which level should the speech sounds be characterized: individual dialect, sociolect, individual person, individual word, individual instance (token) of a particular word? If so, which instance then?

suggests that “UPSID in fact generally fails to capture the actual phonetics of vowel systems, which unfortunately facilitates claims about dispersion patterns in vowel systems by, for example, Liljencrants and Lindblom (1972) and Flemming (2004), though careful phonetic study of a representative range of vowel systems has shown these claims to be unjustified (Disner, 1983)” (Vaux, 2009, 79). An example contrasting Khalkha Mongolian [khk] in UPSID and a phonetic study by Rialland and Djamouri (1984) is provided. Vaux shows that UPSID fails to include more than one high front unrounded vowel and instead organizes the vowel system in terms of backness and roundness. The point that many grammars and phonological descriptions do not contain a phonetic study is a straightforward criticism of collecting segment inventories from the available literature (it is an unfortunate truth that much language documentation does not include in-depth acoustic phonetic studies). This fact is exemplified by UPSID incorrectly representing “many languages with aspirated stops as not aspirating these stops”, as shown in phonetics literature published after UPSID$_{451}$ (Vaux, 2009, 79). Vaux suggests that flawed results from grammar writers that fail to indicate aspiration in their transcriptions, even if they are aware of it, ultimately leads successive researchers like Maddieson (1984) and Clements (2009) to conclude things like non-aspiration as the unmarked state for voiceless stops.[58] This is part of the larger issue of transcription/orthographic effects that are due to the extraction of segments from language descriptions, i.e. distinctions that are not conveyed in the transcription or orthographic systems may be lost even if they are noted elsewhere in the grammar. Vaux (2009, 79) cites some examples in UPSID:

- “Sinhalese implosive stops are nowhere to be found in the inventory of page 272 of Maddieson 1984, presumably because they are not written as such in the orthographic systems”

- “the famously rounded Farsi [ɒ] is rendered as <a> (1984:268)”

- “the Turkish [æ] allophone of /e/ that occurs before {r, l, m, n} is omitted from

[58] Vaux and Samuels (2005) argue against the generalization of non-aspiration as the unmarked state of voiceless stops.

> the Osmanli inventory on page 277, presumably because it is not conveyed in the orthography"

The first two examples seem like errors.[59] The third seems irrelevant – why include an allophone in UPSID when it is specifically designed to be a database of phonemically contrastive segments?

In addition, transcription is an impressionistic analysis and its use in phonetic generalizations requires caution because it reflects the linguist's perceptional biases. As an example, Vaux (2009, 80) cites a generalization from Clements 2009, based on UPSID, that "having one voiced fricative makes it more likely that another will occur in the same inventory can follow directly from whether or not the individuals who did the original transcriptions were able to hear voicing in obstruents successfully". However, as he notes, "This is no trivial matter, as shown by the fact that only the most observant phoneticians and phonologists are aware that speakers of English generally devoice word-initial and word-final obstruents (e.g., Haggard 1978, Pierrehumbert and Talkin 1992, 109)." (Vaux, 2009, 80).

Another criticism from Vaux is that segment inventory databases like UPSID do not contain idiolectal and dialectal variation, which he asserts is crucial in formulating accurate typological generalizations. An example is provided of the variation found in English between speakers who oppose unaspirated fully voiced and voiceless series (Lisker and Abramson, 1964; Scobbie, 2002) and speakers who oppose plain and aspirated series (Vaux, 2009, 79). This is perhaps an extreme example, considering the variation among the myriad of English speakers in the world.

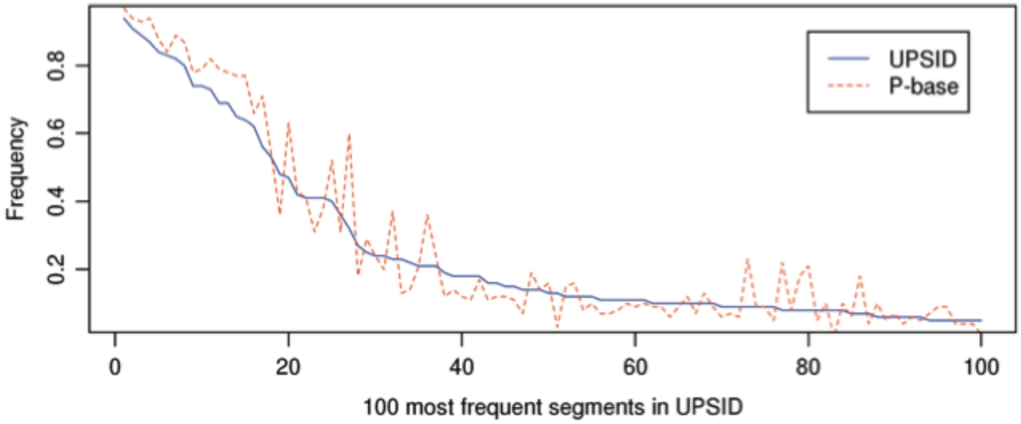
Typological databases like UPSID are also criticized from a phonologist's perspective. Vaux asserts that UPSID is inconsistent in its level of phonological representation because it sometimes seems to describe allophonic representations, and other times phonemic ones (perhaps these were just mistakes, as mentioned earlier). He provides a list of confusions that he says exemplifies the conflicting levels of surface and underlying representations found in UPSID. One example is UPSID's Turkish segment inventory, which allophonically, "is described as having a glottal stop (p. 277), which to the best of my knowledge appears only

[59] See Section 4.3.2.

allophonically in word-initial position", and phonemically, "is listed as not having /ŋ/, which is true phonemically but not allophonically" (Vaux, 2009, 80-81).[60] The basic problem is the collapsing of the surface and underlying levels of phonological representation.[61]

Vaux (2009, 82) summarizes UPSID's database flaws by concluding that it "should not be used as a basis for typological phonological analyses". Regarding Vaux's criticisms, there will undoubtedly be errors and inconsistencies in UPSID and other databases.[62] What is the alternative? No databases? Selecting language descriptions that agree with one's point of view? Or perhaps typological observations are not useful because they necessarily involve disagreements, errors and inconsistencies? Mielke (2009, 714) notes that "an alternative to dismissing inventory databases as useless is to look carefully at the factors that intervene between the language data and the database". Figures 2.4 and 2.5 show comparisons of the typological distribution of segment frequencies and inventory sizes in UPSID$_{451}$ and P-base.

Figure 2.4: Comparison of most frequent segments in UPSID$_{451}$ and P-base (Mielke, 2009, 702)



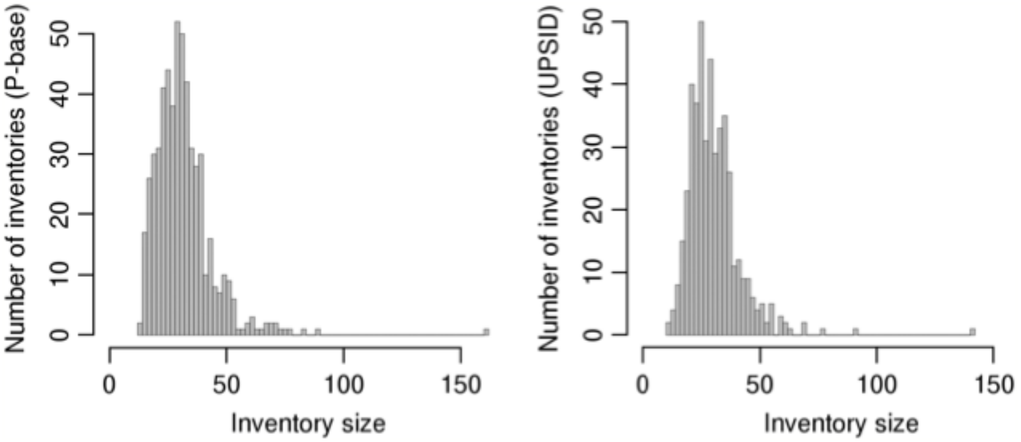
Mielke's P-base is a database of 549 languages that encodes several thousand sound patterns, which he used in his work on emergent feature theory (Mielke, 2004, 2008). Although

[60] These observations remain in UPSID$_{451}$.

[61] In Section 2.3.3 I discuss these issues further.

[62] Errors and inconsistencies ultimately need to be corrected. A nice feature of PHOIBLE is that it is extensible and its inventories are easily correctable.

Figure 2.5: Comparison of inventory sizes in UPSID_{451} and P-base (Mielke, 2009, 703)



P-base was not explicitly built for studying segment inventories, comparisons of its inventories against UPSID_{451} shows that although there is a difference in their contents, they "nonetheless reflect properties of human language" and "underneath the effects of methodology, there is a core of truth" because "both [databases] nonetheless reflect properties of human language" (Mielke, 2009, 714). This occurs despite the fact that P-base's sampling method did not exclude languages in an attempt to create a genealogically balanced sample, whereas UPSID attempts to create one via a quota sample. Additionally, Clements (2009, 24) insists that generalizations "supported at a high level of significance by large numbers of genetically diverse languages are unlikely to be far off the mark" and that problems with typological databases like UPSID are "to a considerable extent [...] alleviated by the sheer size of the sample". UPSID_{451} and P-base represent roughly 6-7% of the world's known languages.

In the end, there seems to be an underlying truth present in the phonological inventories of languages. The notion of a segment inventory is an abstraction over the set of segments as defined by the distinctive features employed by a language (Clements, 2009). It is clear that phonemes are chosen in groups based on their features. In this work I move beyond segments and create models that allow researchers to investigate inventories and lexical items, encoded with segments, at the level of distinctive features. Lastly, just because we cannot make a perfect database that is free of all kinds of bias, this does not mean a database built out of the current information is not worthwhile. It does mean that the research using the database has to be informed by what its limitations are and that a principled approach to data collection and analysis should be undertaken. Hyman (2008, 88) points out that “All of the above is, of course, well-known and unsatisfyingly general: We would like to establish that all languages have specific consonants and/or vowels. However [...] the study of universals is fraught with difficulties.” Clearly the question of what constitutes adequate descriptive categories for linguistic phenomena, particularly in its application to typological databases, is an area of ongoing research and debate. To add fuel to the fire, extrapolating statistically valid results across a typological database with incomplete genealogical coverage is also an area in typology that has been intensely debated. This is the topic of the next section.

2.3.2 Sampling

The second problem that arises from using typological databases to characterize the distribution of linguistic phenomena is due to the challenges involved in creating a reliable data sample for undertaking statistical inference. The challenge of deriving a cross-linguistic language sample that captures genealogical, areal and typological diversity was raised as early as Sherman 1975. Later, statistical methods based on classical sampling theory were described as not tenable for most typological data (Janssen et al., 2006). The foundation of many of these methods requires a population from which a random sample can be drawn and one that fits a normal distribution.[63] However, language data are a skewed population of data points due to factors including the diffusion of typological features through shared descent and geographic proximity. Of course one can draw a random sample from the population, but it might not be representative for the question being asked. Thus, the question of how to establish an ideal sample for purposes of statistical evaluation is central to typological methodology.

[63] I use the term *sample* to mean a set of languages under study and the term *population* to mean the set of all languages from which a sample is drawn.

The nature of linguistic data presents several confounding factors, or biases, that distort the ability to draw a random sample of languages from a population of all languages. The first is the bibliographic bias which stems from the fact that as many as 2/3rds of all languages have no grammar or grammatical sketch (Bakker, 2011).[64,65] This restricts samples to languages that are (well) documented. The bibliographic bias is one factor that causes the genealogical bias. Sampling randomly of the available linguistic documentation risks oversampling widespread well-documented language families. However, the genealogical bias is also reflected in the unequal distribution of languages into language families. Of the 118 language families listed in the Ethnologue 16th edition, over 1/3rd (45) are language isolates.[66,67] By choosing a random sample from a population of unequally dispersed languages, there is a greater chance that large language families will be better represented than isolates or small language families. Additionally, we might assume that isolates or small language families have potentially unique typological features. Inferences on a sample that does not take into account a genealogical weighting, or *stratification*, are likely to be biased towards the features of the larger language groups. Bakker (2011) also mentions the possibility of population size as a cultural parameter that affects the speech community. He likens it to the principle of genetic drift, i.e. a change in genetic variation that causes unlikely gene combinations to be successful due to random sampling in small populations (cf. Kimura 1968, 1983), to linguistic drift. In small populations of speakers then, the likelihood of encountering more exotic (or rare) linguistic phenomena may be greater. An example is

[64] This figure might be a bit too high. Hammarström's most current estimate is that of 7622 languages (living and extinct), there are minimally 2600 languages with grammars and an additional 1310 with grammatical sketches.

[65] Bakker (2011) points out that the bibliographic bias can also be inflicted by the linguistic theory used in language documentation, i.e. creating a sample not only requires language documentation, but also comparable analyses.

[66] For visualization, see Figure 7.6 on page 302.

[67] These 118 language families do not include the categories for pidgins, creoles, unclassified languages, constructed languages and deaf sign languages. In addition to the 45 isolates listed in the language isolates category, there are seven language families listed with one language: Alacalufan, Basque, Chimakuan, Lule-Vilela, Mura, North Brazil and Peba-Yaguan. It is not stated why these single-language language families are not listed in the isolates category. Further, the Chimakuan family had at least two languages. Chimakuan has been extinct since about 1920 and Quileute is also likely extinct at this point (Sharon Hargus, p.c.).

given by Nettle (1999a), in which object initial word order most often appears in languages with under 3000 speakers (Bakker, 2011). Taking the possible effects of genetic influences on language even further, research undertaken by Dediu and Ladd (2007) shows a correlation between a linguistic feature (tone) and two alleles (alternative forms of a gene) when testing 26 typological features in 49 populations on 983 alleles. This correlation appears although most linguistic features and genes investigated show no correlation.

Confounding biases have typically been dealt with through methods for statistical stratification, in which the population is divided into strata (e.g. genealogical units like language families) from which a random sample is drawn equally from each stratum. Yet it is not only linguistic genealogical factors that play a role in the divergence and convergence of typological variables. The linguistic diffusion of areal features caused by language contact may also require stratification to create an unbiased data sample. Additionally, a sample may contain a typological bias in which languages with the same linguistic feature are by coincidence disproportionately represented[68] or a cultural bias because of a disproportionate number of languages from the same cultural area (Perkins, 1992).[69] It is important to note that the confluence of these factors is not independent of each bias. The diffusion of typological variables are the combined result of shared descent, areal diffusion and universal structural principles (Bickel, 2008). Furthermore, many genealogical and areal classifications are not well established[70] and the effects of language contact are not completely understood. To boot, the outcomes of statistical approaches change drastically depending on the genealogical classification used for stratified samples (Rijkhoff and Bakker 1998, 277-292; Cysouw 2005, 556).

There are four types of sampling used in typological studies: convenience, random, variety and probability (Bakker, 2011). The type of sampling used in a study is driven by the question that is intended to be answered. In general, there are two types of studies. The

[68] A typological feature shared by a group of languages need not be caused by genealogical or areal diffusion; it may have developed independently in different languages.

[69] Cultural bias stratification is useful for investigating correlations between linguistic structures and cultural complexity. See Perkins 1980.

[70] For visual comparisons of competing genealogical hypotheses, see `http://multitree.linguistlist.org/`.

first aims to establish the probability that a language has a specific feature. For example, what is the chance that a language has /ŋ/ or that a language is of a specific word order type? For these question types, random or probability samples are used. The second type of study is to simply explore the range of variation of a particular linguistic feature or language type (e.g. what is the range of attested vowel harmony?). For these studies, the convenience and variety samples are used.

A convenience sample is simply that – a set of languages chosen with no restrictions on the basis that data are readily available. Convenience samples are typical of exploratory investigations, but must be refined when testing proposed hypotheses.

A random sample ignores any genealogical, typological, geographic or cultural stratification (Bickel, 2008). Based on their research investigating sampling and stratification techniques with a sample of 4375 languages' numeral systems, Widmann and Bakker (2006) show that capturing diversity is more dependent on stratification than sample size. They also show that a random sample fares well against stratification methods when the sample size is very large. At this time, however, the large size and typological coverage of their sample is currently atypical of most typological databases.

A variety sample is used for explorative research and its aim is to maximize linguistic variety and the likelihood that different values are attested for the typological variable under investigation (Rijkhoff et al., 1993; Rijkhoff and Bakker, 1998). It aims at producing a reliable snapshot of current genealogical and areal distributions, and is therefore opposite of genealogically balanced samples that control for these biases (Bickel, 2008). Variety samples tend to be large and are designed to be diverse. Shosted (2006) uses a variety sample to investigate the language complexity problem, i.e. the historical linguistics truism that simplifying language structure in one place is likely to complicate the language elsewhere. Shosted calls this the negative correlation hypothesis and shows that there is no evidence of a trade-off in complexity between potential syllables and verbal inflection markers in a variety sample of thirty-two geographically and genealogically diverse languages. The maps used in WALS are another example of a variety sample aimed at typological diversity (Haspelmath et al., 2005). However, any summary statistics based on a sample that contains a higher number of languages than known language families, like several chapters in WALS, should be controlled for genealogical bias (Bickel, 2008). Whereas variety samples are suitable for exploratory research and for illustrating the range of linguistic diversity, a probability sample that strives to be free from bias should be used in studies that investigate the probability of occurrence of a specific phenomenon or the correlations between the occurrence of phenomena.

Bell (1978) is the first to discuss in detail sampling techniques and sources of bias in typology, and proposes a stratified probability sample, which is also the most widespread technique used in the social sciences (Cysouw, 2005). This type of sampling is preferred when deriving conclusions about the distribution of some phenomenon over a population because probability samples control for biases through stratification. Bell's proposal for genealogical stratification is to sample languages from the same stock proportionally to the number of genera per stock.[71] Since Bell's proposal there has been much work undertaken in attempt to perfect sampling. Perkins (1980, 1988, 1992) introduces cultural independence by stratifying Bell's genealogical sampling method by including only one language from each world cultural area, as formulated by Murock (1967). Tomlin (1986) uses a combination of genealogical and areal stratification and bases his sample on the number of languages per genus, instead of stock. These genera divide the world into 26 linguistic areas. Dryer (1989, 1991, 1992) introduces 322 language genera and proposes ignoring any classification above the level of genus, which introduces caps at 3500-4000 years (although many genera are much younger than this), a reportedly reasonable time depth for exploring correlations of shared descent.[72] Additionally, variable values are established per genus and each genealogical group is put into an areal grid, thus addressing the areal bias to an extent. Also, by moving the level of sampling up from language to language genus (Dryer, 1989, 2000), the problem of exhaustive sampling of languages is avoided (Janssen et al., 2006). Each author's method provides a degree of independence between sampled families (Bakker, 2011).

[71] I follow the terminology used in Cysouw 2005, 555. The term *genus* (also *family* in Nichols 1992, 24) refers to a genealogical group along the lines of subfamilies like Germanic or Romance (Bell, 1978, 147) (Dryer, 1989, 267). The term *stock* (also *phylum* in Perkins 1992, 128 denotes the highest node in a genealogical tree, e.g. Indo-European or Niger-Congo (Bell 1978, 148; Nichols 1992, 25). I use *language family* when the distinction between stock or genus does not matter.

[72] However, is there any basis for time-depth when there is no (or very little) physical record?

A fully formalized general sampling technique and algorithm that produces genealogical stratification is introduced by Rijkhoff et al. (1993) and refined in Rijkhoff and Bakker 1998. Their method has become standard in typology for controlling for genealogical factors (Bickel, 2008). The sampling technique uses a language classification as input and is designed to generate a sample with the maximum genealogical diversity. For each stock, the structure of the genealogical tree is used to compute a diversity value to insure that the sample is proportional and that rare types are represented. This stratification method can be used to produce a probability sample. In a probability sample, typological values are represented by genealogical units instead of individual languages. Languages cannot be drawn from the same genealogical origin, since that is equivalent to counting the same language twice. One datapoint per genealogical branch is included so as not to skew the sample.

Unfortunately there are several problems with probability sampling. A general problem with all sampling is that the world's (current) languages do not represent all possible languages (Maslova, 2000; Cysouw, 2005; Newmeyer, 2005). Any sample then, represents actual languages, but not all possible human languages, nor all languages that have ever been spoken due to extinction or diachronic change. Another problem, beyond the fact that any stratified probability sample depends on a particular language classification, is the paradox in constructing probability samples (Rijkhoff and Bakker, 1998). If the sample is too small, it will lack the linguistic diversity found in the world's languages. If the sample is too large, it is not possible to exclude genealogically related or areally related languages. The fact is that ideally we would like to include as much data from the world's languages as possible when sampling. Consider for example what happens if one data point is taken per genus (or stock), but that particular genus happens to be radically diverse in regard to the typological variable under study. The data point chosen, then, cannot be the best representative of its particular genus. Furthermore, the heterogeneity of typological features in the genus may or may not be due to genealogical factors (Dryer, 1989; Bickel, 2008). Thus genealogical sampling does not ensure representativeness of the population. Nor is it ideal for investigating family-internal diversity.

Alternatively, Bickel suggests that language families should be sampled as densely as possible to overcome the genealogical stratification problem of all-or-nothing sampling, which leads to sole typological feature representation in diverse language families (Bickel, in press). This approach moves typological sampling away from the *one-language-per-family* stratification method and aims to unwind the confounding factors of shared descent, areal effects and universal structural principles. The problem is not only that taking one data point per genealogical group skews diversity present in those groups. It is also that genealogical sampling methods are not sensitive to the stability of typological variables (Bickel, 2008). Stability refers to the degree that a typological feature is resistant to change over time. The stability of typological variables differ. Moreover, stability for the same typological feature varies in different language families (cf. Nichols 2003). Bickel's *controlled genealogical sampling* algorithm tests for statistical skewing of typological variables by using a recursive sampling technique that tests for diversity at each level of the phylogenetic tree and reduces homogeneous language families to a single data point (Bickel, 2008). This method addresses the distribution of within-family typological features as a result of common descent and takes into account the inflationary effects of language family size on the distribution of features.[73]

Ultimately, sampling procedures impose constraints on hypothesis testing because they limit the already limited data on the world's languages. Another recent approach strives towards full coverage of the population of languages through use of transition probabilities to quantify linguistic change in investigating inter-language dependencies in establishing typological correlations. This work has been pioneered by Maslova (2000, 2002) and Maslova and Nikitina (2008) and adapted recently by Dunn et al. (2011) to investigate the lineage-specific evolutionary dependencies of word order universals. Michael Cysouw refers to these procedures as "dynamic typology" because they attempt to integrate historical factors into synchronic typological data sets by addressing the historical stability of genealogical factors. These approaches move quantitative methods in typology away from a one-language-per-family approach and towards methods that incorporate the full population of languages by developing approaches that do not require classic statistical assumptions.

[73] Open source R code that implements the controlled genealogical sampling algorithm is available at: `http://www.uni-leipzig.de/~bickel/research/software.html`.

To summarize, in this section and the previous one I have explored two problematic issues with using typological databases to characterize the distribution of phenomena in the world's languages. The first is the question of what constitutes comparable typological categories. The second is how to establish samples for purposes of statistical evaluation. Both questions are central to linguistic typology and are relevant in light of building and using typological databases. In particular, I have described some of the specific criticisms against segment inventory databases as a tool for phonological typological studies. I address issues of typological comparison in Section 4.3 in which I describe the implementation of the PHOIBLE data set. In Chapter 5, I revisit the conclusions from typological studies on the distribution of segments in the world's languages and I present a basic stratification technique to address the genealogical bias in the PHOIBLE data set. Accounting for bias is a central issue in linguistic typology and I think there is much more work to be done to explain distributional patterns using modern statistical approaches and typological databases. The following section explores in more depth issues involving the analysis of linguistic data from a phonological perspective.

2.3.3 Data and analysis

In the description of a language's phonological system, the first point for error is encountered during the data's collection. Linguists use a system of transcription to encode the phonetic details of the language they are documenting. Transcription is a scientific procedure that approximates speech by representing a particular researcher's perception of sounds as spoken by a particular speaker of a language. It is an impressionistic analysis that includes the field linguist's own perceptual biases. These biases are due to factors like their phonetic and linguistic training, their own language background, and their experience working with the target and related languages. Because transcriptions are not typically derived through a physical analysis of a speech stream's wave forms, they omit phonetic properties that are not contrastive in the language's phonological system. Thus human transcription encodes less detail than is actually produced in the speech stream.

Two utterances are never pronounced exactly the same way. Variants of speech sounds can occur at the non-contrastive phonetic level, so phonologically conditioned non-contrastive differences are not typically perceived by speakers. This variation is found not just within, but also across languages. In fact the speech signals for the "same" sound in different languages, such as English [i] and German [i], show a difference that is physically measurable, even if an untrained ear has difficulty discerning the difference (Odden, 2005). The linguist's ability to perceive and transcribe sounds directly constrains the input that he or she uses to undertake phonological analysis. Furthermore, an analysis often involves considerations of whether phonetic distinctions are contrastive and these decisions may lead scholars to different conclusions (Maddieson, 2008c). Rather than a given, the number and set of phonemes in a language is a matter of analysis. Conflicting descriptions of the same language's phonemic inventory illustrate this point and there are many examples.[74] Also, the problem is actually more complex than just two conflicting analyses of the same language. It can involve different interpretations of the same analysis, as well as reinterpretations of interpretations of the analysis.[75]

It is common practice for linguists to begin by establishing phonologically contrastive segments when describing a language's phonological system because some system is required to collect and record data (and phonemically contrastive segments are often used to develop a practical orthography for speakers of the language, which provides the mechanism for developing a dictionary and written materials). The procedures that linguists use to determine contrastive segments involves postulating the phonetic characteristics of an underlying contrastive segment, the phoneme, from a series of non-contrastive phonetic surface sounds, the allophones (e.g. Bloomfield 1926; Bloch 1948; Jones 1967). As one example, Jones (1967) establishes phonetic and distributional criteria for positing a phoneme from a set of allophones. The phoneme is:

1. An articulatorily central allophone.

2. The most frequent allophone.

[74] See Table 2.3 and discussion on page 52.

[75] See Section 2.3.6 on data provenance.

3. The allophone least affected by its context.

4. An allophone which can occur in isolation.

These criteria, as interpreted and applied by different linguists, can lead to differences between descriptions of phonemic inventories of the same language.[76] Additionally, the distinguishing criteria may be drawn from different theories that treat the level of phonological representation differently, further allowing linguists to draw different conclusions. In a recent investigation reviewing the current state of phonological universals, Hyman (2008, 85) discusses issues involved in establishing criteria for the cross-linguistic analysis of segment inventories:

> "Consider, for example, the possible claim that all languages have voiceless stops. Is this a claim about the input consonants ("underlying representations") of morphemes, surface ("phonemic") contrasts derived from the comparison of words in isolation, or allophonic ("phonetic") realizations of the input segments anywhere within the phrase level? If the claim does not concern the phonetic level, but a more abstract level of representation, a second question concerns the latitude a phonologist can take in (re-)analyzing a system to fit an alleged universal. Phonologists adhering to different theories will certainly draw different conclusions."

Hyman (2008, 99) illustrates a striking example of four different analyses of the vowel system of Kabardian [kbd], reproduced in Table 2.3. This example illustrates the description of contrastive vowel qualities in abstract models that delineate series of sounds by features. In this case, the height dimension is used to describe the various vertical vowel systems proposed for Kabardian. In UPSID, vertical vowel systems were reanalyzed to "normalize" the different theoretical analyses across different phonological descriptions (Hyman, 2008, 98). To attain interoperability in a cross-linguistic resource that draws from so many different language descriptions, various standardizations are required.[77] Thus as Hyman points

[76] Examples are given in Sections 2.3.4, 2.3.6 & 5.4.1.

[77] See Section 2.3.5.

out, there is a paradox between the need for linguistic theory to describe languages and the abstraction away from individual linguistic theories to undertake cross-linguistic research (Hyman, 2008, 85).

Table 2.3: Analyses of vertical vowel system of Kabardian (Hyman, 2008, 99)

Ladefoged & Maddieson 1996	/ɨ ə a/
Halle 1970	/ə a/
Anderson 1978	/a/
Kuipers 1960	No vowels

Consider another example of the different functions of phonology frameworks and their phonological representations, reproduced in Table 2.4. The function of each framework directly affects how a linguistic universal is stated because of the inherent nature of that framework's phonological representation. This in turn determines the methods in which the linguistic universal can be evaluated across a cross-linguistic data set because that data set's contents must all adhere to a given framework's level of representation to make valid generalizations. For example, the claim that all languages have voiceless stops must be evaluated at a different level in each framework. Note also that theoretical frameworks are affected by trends in phonology (as pointed out in Hyman 2008; Clements 2009; Vaux 2009 and others), which have shifted from features, rules and abstract underlying representations (or "symbolic categories and operations in human linguistic cognition" (Vaux, 2009, 75)) towards phonetic reductionism. Thus current trends are pushing phonology towards surface realizations without underlying representations. Hyman (2008, 86) attributes the shift away from underlying representations to 1) Optimality Theory (Prince and Smolensky, 1993) and 2) technological approaches (phonology studied through experimental, computational and statistical methods). To undertake linguistic universals research requires standardization within a particular framework of linguistic theory (or more ambitiously across frameworks) and in each framework some set of issues must be addressed.

Table 2.4: Comparison of four phonology frameworks and their positions (Hyman, 2008, 85)

Framework	*Representations*	*in terms of*
Structuralist phonology	contrastive	phonemes, allophones
Generative phonology	morphophonemic	URs, (ordered) rules
Non-linear phonology	syntagmatic, geometric	tiers, trees, grids, domains
Optimality theory	n/a (?)[78]	ranked, universal, violable constraints

In this work I adhere to principles of what has been termed basic linguistic theory (Dixon, 1997, 2009a,b) and typological theory (Nichols, 2007), i.e. framework neutral approaches used in language description and for the analysis and comparison of different languages. The focus in basic linguistic theory is to describe each language in its own terms.[79] It is in a sense a general theory-neutral framework used by many linguists and typologists that has been influenced by pre-generative structuralist traditions and by early generative grammar.[80] The structuralist and generative phonology frameworks have been integral in the development of contrastive segment inventories and distinctive feature theory. In the following sections, I discuss the issues in segment analysis and standardization for creating a cross-linguistic data set to undertake phonological typology.

2.3.4 Segments

There are four particularly problematic areas in postulating segments. The first is determining whether a segment is a single unit or a sequence of segments (Maddieson, 1984, 6). This case is illustrated by many different segment types, such as diphthongs, long vowels,

[78] This question mark appears in Hyman 2008, 85.

[79] Compare with *descriptive categories* in Haspelmath 2010.

[80] For a description of basic linguistic theory, see Dryer 2006 and `http://linguistics.buffalo.edu/people/faculty/dryer/dryer/blt`.

geminate consonants, affricates, clicks, and segments that are nasalized, labialized, palatalized, velarized, etc. The second problematic area is determining whether suprasegmentals like stress and tone add to the total number of phonemes in a language's segment inventory. To this list we can add a third problematic choice: whether to include marginal phonemes in the total number of segments in an inventory (Maddieson, 2008a). A final consideration involves what to do with homorganic segments and underspecified segments. Each area is discussed in this section.

Let us start by examining more closely the first issue, determining whether a segment is a single unit or a sequence of more elementary segments. Miret (1998, 27) identifies this question as one of mono- vs biphonematicity and points out that it has long been a controversial topic in structuralist phonology, e.g. "suspicious sounds" in Pike 1947, 251 and "suspect" complex phonetic events in Maddieson 1984, 161. This issue of whether a complex segment type should be considered contrastive or not can drastically change the total number of segments in a language. For example, if non-quality vowel distinctions like length, nasalization or phonation type are taken into account, the total number of vowel segments in a language may double or even triple. This in turn affects claims made about the range or mean number of segments across languages. As analyzed by Migliazza (1998a, 56), Table 2.5 shows contrastive length and breathy voice in So [thm], a Mon-Kher language spoken in Northeastern Thailand.

Migliazza (1998a, 55) states, "There are 22 single vowels (11 basic vowels that can be short or long)... These can occur in either register which gives a total of 44 vowels". That is, there are 22 vowels when the 11 basic vowels are considered short and long. According to the Migilazza's analysis, there can be an additional 22 vowels because both short and long vowels can be contrastive in breathy voice. On the other hand, Nuchanart (1998a, 39) posits "twenty single vowels and three diphthongs", where single vowels include short and long counterparts of /i, e, ɛ, ɨ, ə, ʌ, a, ɔ, o, u/. Vowel register is mentioned as clear voice, clear glottalized voice and breathy voice (Nuchanart, 1998a, iv).

Another example of the difficulty in analyzing the number of distinctive segments comes from Holton's description of Tanacross [tcb]. The difficulty lies in determining phonemic length, which is morphologically conditioned and determined, as stated, by his choice of

Table 2.5: Contrastive non-quality vowel distinctions in So [thm]

Form	Meaning
pu	“pregnant”
puː	“grandfather (Thai)”
pṳ	“blow gum”
kom	“to grab”
ko̤ːm	“to bump into”
ko̤m	“to be sharp”
ɲa	“with”
ɲaː	“the head of spirits”
ɲa̤	“to divide”
ɲa̤ː	“grandmother (Thai)”

analysis (Holton, 2000a, 66-67):

> “Beyond these morphologically conditioned length contrasts there is little evidence for a phonemic length contrast in stem vowels. However, I should stress that this conclusion relies crucially on my analysis of the Tanacross vowel system as consisting of six phonemic vowels. Many of the phonemic distinctions in stem vowels which I have analyzed in terms of vowel quality have been previously analyzed in terms of length. For example, Leer analyzes Tanacross as having a five-vowel system and interprets the distinction between my [teɬ] ‘crane’ and [tɛɬ] ‘blood’ as a length distinction between [teˑɬ] and [teɬ], respectively (1982b: 6).”

Maddieson (2005, 14) asserts that lengthened and nasalized vowels that are listed as separate phonemes, e.g. [õ] vs [o], are not reliable because the considerations that linguists use to determine if their distinction is phonemic can lead different scholars to different conclusions. Therefore, Maddieson (2005) excludes length and nasalized forms from his analysis (Hay and Bauer, 2007, 389). This approach favors treating complex segments as combinations of elementary units. Additionally, in cases like diphthongs, it is often difficult to tell from a language description whether the author intended a diphthong or a sequence of vowels. This is apparent in the fact that basic monophthongs are more consistent across analyses of the same language (Bauer, 2007). However it should be noted that these quality distinctions are included in the segment inventories databases of Maddieson 1984 and Maddieson and Precoda 1990. In studies both approaches have been pursued.

Another approach is to list complex segments separately. Hay and Bauer 2007 distinguish between basic monophthongs, extra monophthongs and diphthongs.[81] This can help alleviate the non-trivial issue exemplified by descriptions of languages like English, in which phonemic contrasts may be lost in statistical or typological studies that throw out diphthongs because their analysis cannot be necessarily relied upon (cf. Maddieson 2005). For example, a description of American English may not contain a separate /ɔ/ phoneme, because it is described in a diphthong (e.g. Ladefoged 1999).[82] Diphthongs like those in American English can be analyzed as having two complex types of nuclei (Miret, 1998). Lehiste and Peterson (1961) distinguish between diphthongs as two target positions, such as [aɪ, aʊ, ɔɪ] in words like “tight”, “loud” and “voice”, and single target position complex segments that should not be classified as diphthongs, including [eɪ, oʊ, ɝ] in “fate”, “lope” and “hurt”.[83] Simply throwing out diphthongs like /ɔɪ/ can artificially decrease the total number of contrastive segments in the language because a description may not posit the nucleus of the diphthong as phonemically contrastive. The approach I have taken in the development of PHOIBLE is to include all complex segment types, but I kept track of the

[81] Extra monophthongs consist of non-quality distinctions such as length and nasalization (Hay and Bauer, 2007, 389). See Chapter 7.

[82] This is a bit of a simplification because there are many different varieties of English spoken and their segment inventories vary quite a bit. For example compare Ladefoged 1999, Hillenbrand 2003, Cox and Palethorpe 2007, Roach 2004, Watson 2007, Bauer et al. 2007, and Watt and Allen 2003.

[83] See Miret 1998 for an overview diphthongs, a discussion of the problems of their analysis, and the different dichotomies proposed for classifying them.

type of each segment, so that researchers can exclude complex segments like diphthongs from their analyses, if they wish.

The second problematic issue is whether suprasegmentals like stress and tone add to the total number of phonemes in a language's segment inventory. In the SPA database, the compilers included tones as contrastive segments (Crothers et al., 1979). In UPSID, suprasegmentals were not included in the total number of distinctive contrasts in segment inventories. Maddieson (1984, 6-7) states, "Stress and tone have always been treated as suprasegmental; this is, tonal and stress contrasts do not by themselves add to the number of distinct segments in the inventory of a language, but if differences in segments are found which accompany stress or tone differences, these may be regarded as segmental contrasts if the association does not seem a particularly natural one". Perhaps not coincidentally, this shift in opinion of prosodic features as contrastive segments occurred around the time of Autosegmental Phonology (Goldsmith, 1976); work on SPA came to an end around 1976 and Maddieson published his UPSID$_{317}$ database in 1984. In the PHOIBLE data set,[84] I decided to include tones as separate segments in segment inventories, so that they can be used in queries and in statistical analyses.[85] Tone segments, however, are also labelled so they can be excluded from analyses as well.

The third problematic issue is whether or not to include marginal phonemes in segment inventories. Marginal phonemes encompass the less "prototypical" segments found only typically in few linguistic forms in a language, such as borrowings, onomatopoeia or rare grammatical functions.[86] Maddieson (2005) excludes marginal phonemes that have been borrowed through the spread of world languages, generally within the last few generations.

[84] See Chapter 4.

[85] Another method to include tones in segment inventories is to mark them as features on vowels, e.g. high tone /á/. This information is inferable from treating tones as separate segments and keeping track of which segments are vowels.

[86] Jelaska and Machata (2005) examine principles of phoneme categorization. Using Croatian as an example, they show that the "prototypicality" of a phoneme varies, with marginal phonemes lying on the periphery of phonemes. To this we can add that within a certain type of marginal phoneme, for instance marginal phonemes found in loanwords, there can also be a continuum, such as "degree of nativeness". For example, Bowden (1997a, 30) notes that in Taba [mky]: "loan phonemes range from highly marginal /ʔ/, through the increasingly less marginal /ʤ/ and /ʧ/ to the almost nativised /f/... Any dividing line that could be drawn between phonemes that are 'native' and phonemes which are not would by necessity be somewhat arbitrary."

In the PHOIBLE data set, I decided to include marginal phonemes from the segment inventories that I extracted from grammar and phonological descriptions. However, I have taken the additional step to mark these phonemes as marginal, so that users can include or exclude them from their queries and statistical analyses.[87]

The fourth and final problematic issue is what to do with homorganic or underspecified segments. A homorganic segment is a type of “proto-” or “archi-” phoneme. Because of an author’s analysis, the segment is determined to be underlyingly underspecified. An example is provided by a description of the Baule [bci] nasal segment in Table 2.6 (Timyan, 1976, 13). The homorganic segment assimilates in place of articulation with the following consonant; only voiced stops occur following the homorganic nasal. Additionally, the homorganic nasal is syllabic and tone bearing when it appears word initially before a consonant. Nasals do not appear in onset position before vowels; they may appear in coda position.

Table 2.6: Homorganic nasal segment in Baule

Segment	Environment
/N/	Homorganic nasal underlyingly
[m]	preceding /b/ and /m/
[ɱ]	preceding /f/
[n]	preceding /d/, /l/ and /s/
[ɲ]	preceding /ɟ/ and /j/
[ŋ]	preceding /ɡ/
[ɲm]	preceding /ɡb/

Homorganic segments typically appear in nasals, rhotics and laterals. According to phonetic and distributional criteria in a structuralist analysis, it is often difficult to establish a phoneme from the set of allophones that appear in the language. This is due to the fact that, on the surface level, the contrastive underlying phoneme sound assimilates in place of

[87] Note that marginal status is only available when that information was described in the original resource.

articulation with the preceding or following segment and there seems to be no most frequent sound.

Another type of underspecified segment is simply an unspecified sound in a language description. In SPA the symbols “r” and “r-retroflex” are used in segment inventories “when the manner of articulation cannot be determined” from the resource in which the inventory was taken (Crothers et al., 1979, 13). In UPSID Maddieson (1984) encountered this phenomenon in language descriptions with rhotics and labeled these “r-sounds”. UPSID examples and PHOIBLE interpretations are provided in Table 2.7. In PHOIBLE I have marked these segment types with an asterisk. The table also shows the underspecification of a segment’s place of articulation, which not only occurs in rhotics in UPSID, but across segments in the dental/alveolar space.

Table 2.7: Unspecified “r-sounds” in UPSID

UPSID description	UPSID$_{317}$	UPSID$_{451}$	PHOIBLE
voiced alveolar r-sound	rr	rr	*R
voiced dental r-sound	r̪r̪	rrD	*R̪
voiced dental/alveolar r-sound	“rr”	‘‘rr	*R̪|*R
voiceless dental/alveolar r-sound	N/A	‘‘hrr	*R̪̣|*Ṛ
laryngealized voiced dental/alveolar r-sound	“r̰r̰”	‘‘rr*	*R̪̰|*R̰
palatalized voiced alveolar r-sound	rr^j	rrJ	*R^j

In the overall development of a segment inventory database, each language description from which an inventory is extracted needs to be examined in detail and the segments determined from the author’s description. In the UPSID inventories Maddieson sometimes agrees with the interpretation of the original source, e.g. Rotokas [roo] (Firchow and Firchow, 1969a), and other times does not, Maxakali [mbl] (Gudschinsky et al., 1970), as noted in Hyman 2008. Maddieson (1984, 6) explains, “Our decisions on phonemic status and phonetic description do not always coincide with the decisions reached by the compilers of the

SPA, and we have sometimes examined additional or alternative sources". Furthermore, in the UPSID database, "each segment which is considered phonemic is represented by its most characteristic allophone, specified in terms of a set of 58 phonetic attributes".[88] As explained in Section 2.3.1, this method has drawn much criticism, including Simpson (1999, 350), who states, "It is little wonder then that both Maddieson and Crothers use the term 'characteristic' without defining it". Because transcription systems approximate speech, they are limited by necessity to a small number of segments, represented with alphabetic symbols. Mielke suggests that it is possible to deal with some of these issues, like using characteristic allophones as contrastive segments, by reducing segments into important phonetic distinctions. A general statement of the type "Language X contrasts labial and coronal sounds... is less likely to be corrupted by description issues" than a more specified statement like "Language X has /p/ and /t/" (Mielke, 2009, 715). This broadening of the phonological claim then relies less on an author's thesis of what a particular phoneme for a particular set of allophones is.[89]

The development of a segment inventory data set faces the problems of establishing inventories that can be compared and should ideally document the procedures taken. Some of the theoretical linguistic issues regarding segments have been discussed in this section. The general strategy in the development of PHOIBLE has been to encode as much information as possible from the original resources, in such a way that users can query based on their views of these issues. In the next section, I investigate how disparate data in the PHOIBLE data set have been standardized to make segment inventory data interoperable.

2.3.5 Standardization

The observation, "The nice thing about standards is that you have so many to choose from", is spot on (Tanenbaum, 2003, 235). Choosing and following standards is a complicated task.

[88] The term *phonetic attributes* presumably covers the distinctive features specified in UPSID (e.g. high, front, etc.) as well as categories for vowel, diphthong, etc.

[89] It is also practical for a segment inventory database to allow users to query not only on segments, but on features and combinations of features as well. The PHOIBLE knowledge base provides this functionality, as discussed in Section 3.2.3. In Section 6.5 I use this functionality to investigate descriptive universals in phonological systems.

Some other observations about standards include: they cause their adopters more work; in general most people don't follow standards or they tend to cut corners when they can; standards are often difficult to understand and adhere to; and many (or maybe most) people simply have their own methods that they believe to be superior to an established standard. Without standardization, however, different parties face the *coordination problem*, i.e. only when all parties make mutually consistent decisions can all parties realize mutual gains. In the scope of technological infrastructure for linguistic data, choices of technical standards are required to make disparate data sets interoperable. In this work, standardization is the process of establishing or adhering to already existing technical standards to attain interoperability. This section discusses standards for transcription, digital encoding of data and metadata.

Like many standards, the IPA receives its fair share of criticism.[90] Therefore, it is likely to be a point of criticism of the PHOIBLE data set, which uses the IPA as the standard of transcription for its contents. I used IPA in PHOIBLE because it is the most commonly used transcription system for linguistics and it will be into the foreseeable future. For the most part, IPA's segments are also digitally encoded in the Unicode Standard.

The IPA underwent a major revision at the 1989 Kiel convention, resolving long historical debates like the transcriptions of tone in Africanist and Sinological conventions.[91] Ladefoged (1990b) urges linguists to abandon idiosyncratic transcription in favor of the revised chart (even though there was consensus by the convention attendees that it wasn't the best possible chart, nor were attendees in agreement on all aspects of the chart). However, in the spirit of standardization, Ladefoged offers three points of encouragement. First, the chart is intended to represent all possible sounds in all languages. Second, although not actually defined by the IPA, the segments in the IPA chart can be taken to represent a bundle of distinctive features, e.g. the symbol <b> is shorthand for the features

[90] For example, see discussions in Ladefoged and Roach 1986; Bruce 1989; Ladefoged 1990a; Pullum and Ladusaw 1996; Beckman and Venditti 2010 and Sally Thomason's Language Log post, "Why I don't love the International Phonetic Alphabet", at: `http://itre.cis.upenn.edu/~myl/languagelog/archives/005287.html`.

[91] The IPA was revised to include both systems for tagging pitch patterns in African and Asian languages: diacritics above vowels and numerals after each syllable, respectively.

[+*voice*, +*bilabial*, +*plosive*]. Third, the chart presents an agreed upon description of phonetic knowledge on a single page. Those who use symbols diverging from the chart, it was hoped, would feel compelled to provide a mapping from their transcription element(s) to the IPA when one is possible. Additionally, from time to time the International Phonetic Association will update the IPA chart. This was done for example in 2005 with the inclusion of the voiced labiodental flap, which was later added to the Unicode Standard in version 5.1.0. The International Phonetic Association also removes symbols, as it did at the Kiel convention for the Japanese-specific syllabic nasal symbol. Although the Japanese syllabic nasal is unusual among the world's language phonologically, the International Phonetic Association decided from a phonetic point of view that the sound was not unusual among syllabic nasals (Ladefoged, 1996). Therefore the IPA, like many standards, continues to evolve. Although this may cause problems for its users, it is good for the standard in general because it is continuously refined towards a general phonetic theory based on our increased understanding of sounds, which adheres to the International Phonetic Association's goal to represent all distinctive sounds in the world's spoken languages.

During the development of PHOIBLE, one major issue was what to do with phonetic and phonemic distinctions that appear in linguistic descriptions, but that are not sanctioned by any IPA symbols or diacritics, e.g. "half-voice" or "weak aspiration/nasalization" in SPA (Crothers et al., 1979). Another more commonly encountered example is the IPA chart's lack of distinct symbols for voiceless implosives (visually voiceless stops with hook top). These distinct symbols were were added in 1989 at the Kiel convention and then subsequently retracted in 1993 because voiceless implosives were considered to only occur as allophones of voiced implosives (Pullum and Ladusaw, 1996). Following the principles of the International Phonetic Association, diacritics should be used for allophonic distinctions, and wherever possible, differently shaped letters should be used to distinguish phonemes (The International Phonetic Association, 1999). The absence of distinct symbols for voiceless implosives in the IPA chart, however, does not change the fact they are used in many language descriptions. This leads to a conundrum. Whereas the International Phonetic Alphabet does not sanction the use of voiceless consonants with hook top to indicate voiceless implosives, they are nevertheless used regularly and interchangeably to indicate allophones (which is wrong) and phonemes (which is not sanctioned) because the Unicode Standard includes characters that visually represent the voiceless implosive series.[92] On the other hand, instead of using these distinct symbols to indicate phonemic contrasts, the voiceless diacritic is used in conjunction with the voiced implosive symbol to indicate phonemic voiceless implosives in a description of Seereer-Siin [srr] (Mc Laughlin, 2005, 203). This use goes against the International Phonetic Association's principles, nevertheless the article adheres to the current standard. Thus I also followed the current approach used by the Journal of the International Phonetic Association.[93]

Although the IPA is easily adhered to with pen and paper, to encode IPA characters electronically, a character encoding system is needed. Early work addressing the need for a universal computing environment for writing systems and their computational complexity is discussed in Simons 1989. For a long time, linguists were limited to ASCII-encoded 7-bit characters, which only includes Latin characters, numbers and some punctuation and symbols. Restricted to these standard character sets that lacked IPA support or other language-specific graphemes that they needed, some linguists made their own solutions (Bird and Simons, 2003). For example, some chose to represent unavailable graphemes with substitutes, e.g. the combination of <ng> to represent <ŋ>. Tech-savvy linguists redefined selected characters from a character encoding to map their own fonts to. However, one linguist's redefined character set would not render properly on another linguist's computer if they did not share the same font. If two character encodings defined two character sets differently, then data could not be reliably and correctly displayed. This is a common example of failure of data inoperability.

To alleviate this problem, during the late 1980s, SAMPA (Speech Assessment Methods Phonetic Alphabet) was designed to represent IPA by uniquely mapping IPA symbols to ASCII characters; thus providing linguists with a standardized electronic character encoding system for sharing data (Wells, nd). However, SAMPA does not encode the entire

[92] Voiceless consonants with hook top are used in many phonological descriptions and orthographies of African languages, e.g. *Systeèmes alphabétiques des langues africaines* (Chanard, 2006), an online digitization of *Alphabets of Africa* (Hartell, 1993). See Section 4.3.3.

[93] In cases where phonetic symbols were needed that are not in the IPA, I added those symbols to the list of "Unicode IPA" characters used in PHOIBLE. See Appendix D.

IPA. SAMPA was derived from phonemes appearing in several European languages and an individual table was created for each language. Therefore, SAMPA was a collection of tables to be compared, instead of a large universal table representing all languages. An extended version of SAMPA, called X-SAMPA, set out to include every symbol in the IPA chart including all diacritics (Wells, nd). X-SAMPA was considered more universally applicable because it consisted of one table that encoded the set of characters that represented phonemes in IPA across languages. SAMPA and X-SAMPA have been widely used for speech technology and computational linguistics encoding. Eventually, ASCII-encoding of the IPA became depreciated through the advent of the Unicode Standard.[94]

The Unicode Standard is now the standard character encoding for the Web (The Unicode Consortium, 2007) and for encoding linguistic data (Anderson, 2003). It aims to provide a unique number for every character in all the world's written languages and it was invented to solve the inoperability problem of different encoding systems.[95] There are hundreds of different encoding systems that were invented independently to capture orthographic diversity as different nations adopted and developed computer systems. These different encoding systems were problematic and in conflict with one another because different standards were formalized differently and for different purposes by different standards committees in different countries. No unified encoding scheme contained enough code points to encode all characters, so two different encoding schemes possibly used the same code point for different characters, or used different code points to represent the same character. Because computers support multiple character encoding schemes, data risked being corrupted when handled by different applications and encodings. The Unicode Standard was devised to alleviate these problems.

IPA, as encoded in the Unicode Standard, is also not without its criticisms. The Unicode Standard encodes *characters*, not glyphs, in scripts and it treats a character as the smallest component of a writing system that has semantic value (Anderson, 2003). It therefore some-

[94] Note, however, that many software packages still require ASCII encoding, e.g. RuG/L04 (`http://www.let.rug.nl/kleiweg/L04/`) and SplitsTree4 (`http://www.splitstree.org/`).

[95] For discussion see Moran 2009.

times unifies duplicate characters across multiple scripts.[96] For example, IPA characters of Greek and Latin origin, such as <β> and <k> are not given a distinct position within the Unicode Standard's IPA Extensions block. The Unicode code space is subdivided into character blocks, which generally encode characters from a single script, but as is illustrated by the IPA, characters may be dispersed across several different character blocks. This poses a challenge for interoperation, particularly with regard to homoglyphs. Why shouldn't a speaker of Russian use the <а> CYRILLIC SMALL LETTER A at code point U+0430 for IPA transcription, instead of <a> LATIN SMALL LETTER A at code point U+0061, when visually they are indistinguishable and it is easily typed on a Cyrillic keyboard? Furthermore, homoglyphs come in two flavors, linguistic and non-linguistic. On one hand, linguists are unlikely to distinguish between the <ə> LATIN SMALL LETTER SCHWA at code point U+0259 and <ǝ> LATIN SMALL LETTER TURNED E at U+01DD. On the other hand, non-linguists are unlikely to distinguish any semantic difference between an open back unrounded vowel <ɑ>, the LATIN SMALL LETTER ALPHA at U+0251, and the open front unrounded vowel <a>, LATIN SMALL LETTER A at U+0061. In fact, this distinction in different "a" characters is another area of criticism for the current version of the IPA.[97] As noted earlier, measurements of formants in language descriptions are quite rare. Mielke (2009) points out that 75% of languages have a five-vowel system in Maddieson 1984. This leads one to ask if transcribed characters are prone to *transcription effects*. For example the common use of "a" in transcriptions could be in part due to the ease of typing the letter on an English keyboard (or for older descriptions, the typewriter). In my work with electronic resources, it is exceedingly rare that a linguist uses <ɑ> for the low back unrounded vowel. Authors simply use <a>.[98] Another example I have commonly encountered is the use of <g> LATIN SMALL LETTER G at U+0067, instead of the correct Unicode IPA character for the voiced velar stop <ɡ> LATIN SMALL LETTER SCRIPT G at U+0261. One begins to

[96] See Section 2.1.4.

[97] For example, see `http://itre.cis.upenn.edu/~myl/languagelog/archives/005287.html`.

[98] One example is *Pilagá Grammar*, in which Vidal (2001a, 75) notes: "The definition of Pilagá /a/ as [+back] results from its behavior in certain phonological contexts. For instance, uvular and pharyngeal consonants only occur around /a/ and /o/. Hence, the characterization of /a/ and /o/ as a natural class of (i.e., [+back] vowels), as opposed to /i/ and /e/."

question whether this issue is at all apparent to the linguist, or if they simply use the former <g> because it is easily keyboarded and saves him or her time, whereas the latter must be inserted as a special symbol. Lastly, the use of the apostrophe is even more confusing and has led to long discussions on the Unicode Standard email list. An English keyboard inputs <'> apostrophe at U+0027, although the "preferred" Unicode apostrophe is the <’> right single quotation mark at U+2019. Yet the glottal stop/glottalization/ejective marker is another completely different character, the <ʼ> modifier letter apostrophe at U+02BC. There is also the ambiguous encoding of IPA segments within Unicode. An example is the U+02C1 modifier letter reversed glottal stop <ˁ> vs the U+02E4 modifier letter small reversed glottal stop <ˤ>. Both are denoted in Unicode as the pharyngealized diacritic and both appear in various resources representing phonetic data online.[99] Lastly, there is at least one case in which the character name assigned by the Unicode Consortium does not match the IPA's description: in the Unicode Standard <ǃ> at U+01C3 is labeled latin letter retroflex click, but in IPA <ǃ> is an alveolar or postalveolar click.

Each of these issues in itself is perhaps enough for the ordinary working linguist to throw in the towel on adhering to Unicode IPA standards.[100] However, it gets better. Computationally, two sequences of characters that are rendered visually identical, e.g. a creaky voice nasalized close front unrounded vowel <ḭ̃>, are in fact different characters depending on the sequence in which the user inputted them.[101] This issue requires using Unicode normalization forms and is discussed in detail in Section 4.3.

An additional problem with the IPA is the lack of symbols for certain distinctions that have permeated the literature. One such example in SPA is the "tense" and "lax" distinction that is found phonemically in languages like Lak [lbe], Pima [ood] and Modern Hebrew [heb].

[99] I chose to go with the latter, U+02E4, in line with both online IPA keyboard implementations from Weston Ruter (`http://weston.ruter.net/projects/ipa-chart/view/keyboard/`) and Richard Ishida of W3C (`http://people.w3.org/rishida/scripts/pickers/ipa/`). The digital implementation of *Alphabets of Africa* by Chanard (2006) uses the former.

[100] For a list of Unicode confusables, checkout `http://unicode.org/Public/security/revision-02/confusables.txt`. John C. Wells also provides a list of easily confusable phonetic symbols at `http://www.phon.ucl.ac.uk/home/wells/confusables.htm`.

[101] U+0069 <i> + U+0330 <ḭ> + U+0303 <ḭ̃> vs U+0069 <i> + U+0303 <ĩ> + U+0330 <ḭ̃>.

At first I chose to represent tense consonants as voiceless and lax consonants as voiced, but this led to the problem of ambiguous segments in the data.[102] For example, Sa'ban [snv] has the phonemically contrastive segments, as given in SPA, in Table 2.8.[103]

Table 2.8: Segments in Sa'ban

SPA	Initial conversion	Final
p	p	p
p-tense	p	p͈
b	b	b
b-tense	b	b͈
t	t	t
t-tense	t	t͈
d	d	d
d-tense	d	d͈
k	k	k
k-tense	k	k͈
g	g	g

Because the IPA does not have sanctioned diacritics for tense and lax, I made an executive decision to take the "strong articulation marker", the combining double vertical line below U+0348 character from the "Extensions of to the IPA" to represent tense. This character has been used in the literature and seems to be the best choice at present. Laxness was a bit more problematic. The combining three dots below character at

[102] The terms "tense" and "lax" are sometimes used to describe a state of the vocal folds in languages that contrast consonants by greater glottal tension. A gross simplification is to equate the feature "tense" to "voiceless" because there is a simultaneous oral closure and a glottal stop. Korean is a well-known example of a language with this distinction, although this contrast is also often referred to as "fortis" and "lenis". Ultimately I decided to include these features in PHOIBLE as they were described by various linguists.

[103] In SPA, Sa'ban has reportedly 46 phonemes (38 consonants and 8 vowels). In UPSID$_{451}$ this figure is much lower; Sa'ban is reported to have 26 phonemes (19 consonants and 7 vowels). Both cite the same bibliographic source: Clayre 1973.

U+20E8 has visually a nice analogy to the breathy voice diacritic, but it is not represented in many fonts, is from an entirely different Unicode block than most of the IPA diacritics, and unfortunately does not seem to combine well when visually displayed. Therefore the COMBINING LEFT ANGLE BELOW character at U+0349 in the "Extensions of the IPA" was chosen to represent "weak articulation" and lax consonants. All decisions that I reached regarding character assignments are documented in Appendix C.

A final issue in character encodings is when a character is supported by a phonetic font, like Doulos SIL, but the font encodes the glyph as a code point in the Unicode Standard Private Use Area (PUA).[104] This occurs when a character is needed, but not supported by the current version of the Unicode Standard. These assignments are problematic because the character may be accepted into the Unicode Standard, at which time the font will depreciate its use of the PUA code point and update the font accordingly. This leaves the onus on the developer to continue to monitor and update changes to their data. Two examples from an earlier version of Doulos SIL are U+F174 COMBINING ACUTE MACRON and U+F171 COMBINING MACRON ACUTE, which have now been depreciated and assigned to code points U+1DC7 and U+1DC4 in the Unicode Standard version 5.0.0.

So far in this section I have highlighted some of the standardization issues involved in phonetic transcription and digitally encoding the IPA. Another issue of standardization is the use of metadata to identify linguistic resources with bibliographic information and to identify which language(s) the author(s) are describing.[105] Metadata is essential in the development of a cross-linguistic data set because for each data point its original source should be identified to allow third party verification of the data in the data set.

For cataloging and describing physical resources and digital materials, the Dublin Core Metadata Initiative (DCMI) has become the standard in the fields of library science and computer science. DCMI aims to create interoperable metadata standards and is defined by the ISO standard 15836. The DCMI metadata set was adopted and expanded by the Open

[104] `http://scripts.sil.org/PUA_FAQ`

[105] Metadata is structured data about data. For an overview of metadata for linguists, see Jeff Good's "A Gentle Introduction to Metadata", at `http://www.language-archives.org/documents/gentle-intro.html`.

Language Archives Community (OLAC)[106] for describing language resources like grammars, field notes, recordings, etc. OLAC expands the set of DCMI metadata categories to include information pertinent to linguistic data to create a standard way to document all types of language resources, by adding metadata elements like subject language and linguistic data type to enhance greater discovery of language resources.[107] For example, the OLAC subject language uses ISO 639-3 three-letter language identifiers to identify a language resource's subject language, i.e. the language being described in a grammar, etc.

ISO 639-3 is an international standard for uniquely identifying language names with three-letter codes. These three-letter codes are commonly referred to as language codes, though they do not uniformly identify languages. The scope of ISO 639-3 codes includes individual languages and macrolanguages.[108]

Why are unique identifiers important and how do they foster interoperability? Now that language codes are available to the community as a standard, researchers and projects that have language data can share that information with a unique, interpretable code that identifies a particular language or language variety. If you know the language's code, searching online databases becomes more accurate and faster because languages tend to have many names and completely unrelated languages may share the same name.[109] For example, consider searching on the language name "Mono". Mono is a language spoken in the Democratic Republic of the Congo by an estimated 36,000 people. Mono, however, is also a language spoken by a few remaining speakers in California, in the United States. The use of ISO 639-3 codes lets us uniquely distinguish these two languages. Mono [mnh] is a Niger-Congo language and Mono [mnr] belongs to the Uto-Aztecan family. This may sound like a one off case, but it is more common than one might think. Consider Mende [men] (Sierra Leone) and Mende [sim] (Papua New Guinea), Kamba [kam] (Kenya) and Kamba [xba] (Brazil), Nama [naq] (Namibia) and Nama [nmx] (Papua New Guinea), and Saliba (Papua New

[106] `http://www.language-archives.org/`

[107] The OLAC Metadata set can be accessed at: `http://www.language-archives.org/OLAC/metadata.html`.

[108] See Section 4.3.1.

[109] The Ethnologue currents lists over 47,000 alternative language names for roughly 7000 unique languages.

Guinea) and Sáliba (Colombia), to name a few.[110]

Language codes are also used to distinguish between closely related languages like Tukang Besi North [khc] and Tukang Besi South [bhq], both of which are referred to as Buton. It is often the case that a canonoical language name is used when in fact there are numerous distinct languages under the umbrella of that language name. Consider for example some of the languages listed in Hay and Bauer 2007 and Bauer 2007: "Berber" (25 distinct languages); "Fula" (9); "Ijo" (9); "Cree" (6); "Mam" (5); "Erromangan" (3); "Friesian" (3); "Gaelic" (3); "Miwok" (3); "Oromo" (3); "Panjabi" (2); Romany (2); "Sotho" (2); "Sorbian" (2).[111] By using language names and not including language codes, it is difficult to retest other researchers' analyses.[112] Following metadata standards like using ISO 639-3 language code identifiers is therefore an important step in validating cross-linguistic research.

To summarize, using standards allows different parties to realize mutual gains by addressing the coordination problem; only when all parties make mutually consistent decisions can all parties realize mutual gains. This allows for greater discovery and access to all kinds of linguistic information, from the identification of language resources to the unambiguous encoding of phonetic data. Bird and Simons (2003) call for community consensus for describing language resources and for identifying suitable data structures for linguistic data types. By adhering to standards, language researchers take a step towards overcoming the coordination problem. In the next section I take a closer look at data provenance, a difficult problem in regard to identifying the source(s) of linguistic data, and in particular, for collecting, extracting and properly citing data from disparate linguistic documents.

2.3.6 Data provenance

From the French word *provenir* "come or stem from", provenance pertains to the evidence of origin and history of something. Its roots are in art attribution, but the notion of

[110] This example does not touch on the even messier situation of ambiguity among language names *and* alternative language names, as they are listed in the Ethnologue. An example is given in Section 3.1.

[111] The number of distinct languages given here is based on Ethnologue 16 (Lewis, 2009).

[112] See Chapter 7.

provenance affects most fields in some way.[113] Addressing provenance of documents has occupied historians, scholars and textual critics for centuries. However, since the emergence of the Web and the ability to easily copy and transform data, a new set of issues in tracking data provenance has emerged as a critical challenge in the Digital Age.

In this work we have gathered segment data from over a thousand different language descriptions. Hundreds are through manual inspection of grammars and phonological descriptions, yet the rest are through the extraction of inventories from databases from projects that have already extracted segment inventories from linguistic descriptions. To provide accountability for a data set's contents, the obvious initial step is to identify and list each source from which data was taken. However, this process is problematic when a segment inventory has been reanalyzed from its original resource by a third party. Furthermore, this process can chain so that a segment inventory that has been reanalyzed is again reanalyzed for the purpose of digitization and online publishing. Let's take a look at some examples.

In Section 2.3.3, I pointed out that rather than a given, the number and set of phonemic segments in a language depends on the linguist's analysis. Thus two linguists' analyses of the same language may contain different segment inventories. Therefore, if researchers wish to collect segment inventories for cross-linguistic analysis, they are faced with several choices. They can include one representative sample of a segment inventory, they can include multiple segment inventories, or they can make their own analysis of a segment inventory based on one or more resources.

One example is the different interpretations of the Ocaina [oca] phoneme segment inventory described in Agnew and Pike 1957. In this work Ocaina is described as having "twenty six consonant phonemes", "five contrastive tongue positions in the vowels", "oral vowels contrast with nasalized vowels, except /e/ which has no nasalized counterpart; it is a very infrequently occurring vowel", and "two contrastive tone levels" (Agnew and Pike, 1957, 24-26). According to my calculation, this indicates a total of 37 segments (26 consonants, 9 vowels and 2 tones). In SPA, 38 phonemic segments are listed, including the two pitch

[113] For example in business, provenance is used to judge the value of something. In archaeology, evidence of provenance is needed to determine an artifact's location of excavation and its history. In law, chain of custody is equivalent to provenance.

accents high and low (Crothers et al., 1979, 495). If we throw out the prosodic features as Maddieson does in UPSID, one would expect there to be 36 segments based on SPA's calculations. However, UPSID lists a total of 34 segments for Ocaina, differing from SPA by the two phonemes /w/ and /ʤ/. In SPA /w/ is labelled "transitional" with the note, "[w] is a transitional sound which occurs after /o-mid/, /h/, and labial consonants, when they occur before /i-trema/" (Crothers et al., 1979, 496). But this is not stated in Agnew and Pike 1957, leading one to question if inclusion of the /w/ is a compilation error that was later caught by Maddieson. On the other hand, UPSID does not include the segment /ʤ/, which Agnew and Pike (1957, 25) list among the "Voiceless Assibilants ¢ č and Voiced Assibilants z̧ ǰ (alveolar, alveopalatal)".[114] If this segment has been reanalyzed in UPSID, no documentation of why is provided (all four affricates are listed in SPA). These different analyses of the same segment inventory provide one example of why data provenance is important for validation in the creation of cross-linguistic data sets.

Data provenance is also an issue of documentation of the reliability of the data and its source. This is particularly important for data on the Web. For example, data extracted from a Web-accessible database may have been originally extracted from another database (and so on), or from another resource that may or may not be publicly available. An example that I encountered is Chanard 2006. This online database is a digitization of segment inventories that were originally collected in an edited volume listing the phonemic and orthographic systems of African languages.[115] These phoneme inventories were each gathered and analyzed from one or more publications, or provided by various language specialists. The digitization of the volume introduced another level of interpretation, one that sometimes differs from my own. Although Chanard's changes are not documented on the website, they can be gleaned in a comparison of the original resource and the digitized version. For example, Hartell (1993) uses Africanist transcription conventions, the IPA symbols of which have changed since the 1989 Kiel convention.[116] These changes have

[114] According to Pullum and Ladusaw (1996, 29), <¢> typically means [ts], so we can infer that "assibilant" means "affricate". Translated into modern terminology, we have "voiceless affricates ts and ʧ and voiced affricates dz and ʤ".

[115] See Hartell 1993 and references therein.

[116] For example, [ɩ] is now [ɪ] and [ʋ] is [ʊ].

been made in Chanard's online version. However, Chanard does not always follow the IPA guidelines, nor do all the digitized segments adhere to the Unicode IPA standard.[117] To adhere to best practices concerning data provenance, this chain of interpretations should be documented from the original publication, to the edited volume, to the online database. Unfortunately this is not often done, nor is it always possible as an outside observer and data consumer to track these changes after the fact.

Dealing with data provenance also means establishing a kind of metadata that documents the data's original source and its history and derivation. Lewis et al. (2006) provide interesting examples of the same snippet of interlinear glossed text being reused and reanalyzed across publications.[118] Their article provides a broad overview of linguistic data use in the internet age and discusses issues of fair use of data. Of course these problems are not new to editors of linguistics journals, who have long faced the challenge of publishing articles that may contain an analysis of data from a secondary source. Such cases are difficult to identify, putting a journal editor in the position of either vetting the examples or trusting that an author's analysis is based on a primary resource. If the primary data source is available, a researcher should not rely solely on a secondary resource (Thomason, 1994). An example is provided by an investigation of vowel length in Haida.

The UPSID database contains a segment inventory for Haida [hai] with a three vowel system ("high front unrounded vowel" /i/, "low central unrounded vowel" /a/, "lowered high back rounded vowel" /ʊ/) taken from Sapir 1923. However, Bauer (2007, 222) writes that Haida might have a six vowel system:

> "For example, Maddieson (1984) states that Haida has three vowels, while Mithun (1999) states that it has six. This does not appear to be a matter of how to analyse long vowels, though it might well be a matter of dialect. The outsider cannot judge."

Although the point that it is difficult to analyze vowel length holds, under closer inspection

[117] For details see Section 4.3.3.

[118] This was discovered with a Web crawler designed to extract interlinear glossed text data from online documents. For details see `http://odin.linguistlist.org`.

Bauer has misquoted Mithun 1999, 415:

> "The general structure of the language [Haida] is illustrated here with Skidegate material from Levine 1977a. The consonant inventory includes... Vowels are i, e, a, ʌ, u... A distinction between high and low tone is easily perceived. Enrico notes that in Kaigani the system is one of pitch accent, so that at most one syllable in a word bears high tone (1991: 103). In Masset, tone contrasts only in heavy syllables, but it is otherwise predictable from syllable structure. Skidegate tone is essentially like that of Masset except that extra length (which has disappeared from Kaigani) has different effects in the two dialects."

If a researcher were to rely on the second hand account of Haida having long vowels, his or her analysis would be based on incorrect data.

Data provenance is a difficult problem and there is much current research which aims to simply clarify and identify the issues involved.[119] Avenues towards a solution are being investigated and they tend to include recording provenance as some type of annotation. This annotation could be attached to components of a database, but because of its rigid structures it is not always easy to attach amorphous metadata. Loosely structured forms of data like graphs may act as a substrate for tracking provenance. This is currently a hot topic in the digital library sciences.

In the OLAC Metadata Usage Guidelines,[120] under "other elements" there exists a metadata definition for "Provenance" that reads: "A statement of any changes in ownership and custody of the resource since its creation that are significant for its authenticity, integrity and interpretation."[121] OLAC models this element after the DCMI, which is actively investigating data provenance.

In this work I have tried to be as transparent as possible with regard to data provenance. A guide to all references from which segment inventories were extracted is provided in

[119] `http://db.cis.upenn.edu/research/provenance.html`

[120] `http://www.language-archives.org/NOTE/usage.html`

[121] `http://www.language-archives.org/NOTE/usage.html#Provenance`

Appendix B. Because in some cases our work with language resources has also involved interpretations of phonetic descriptions into IPA, I list the segment conventions that we developed and use in Appendix C. I provide these data while knowing that data extracted from other databases may contain undocumented errors, reinterpretations and reanalyses.

2.3.7 Summary

In this section I have discussed the linguistic and technological challenges involved in developing a cross-linguistic data set to compare and characterize the distribution of linguistic phenomena. Although my focus is on data from segmental phonology and distinctive feature theory, the broader challenges that I face are applicable to developers of other typological databases. One issue is whether typology can be undertaken with language-specific analyses or if separate over-arching cross-linguistic comparative concepts are needed.[122] This problem is highlighted by typological databases that can bring together a wide range of different descriptions of languages. Large samples of diverse data also raise the issue of how statistical sampling should be used to account for the various types of bias that are inherent in linguistic data sets. Another problem related to typological comparison involves the analysis of data; the problem is captured by the paradox of using linguistic theory to document and describe languages, but the need to abstract away from theory to undertake cross-linguistic comparison (Hyman, 2008). Keeping track of different analyses from different authors is also an issue of data provenance. New analyses may involve the reinterpretation of older analyses, particularly when one wants to standardize across descriptions to create comparative concepts. Lastly, the practical implementation of a cross-linguistic data set to undertake phonological typology requires the standardization of segments at both the linguistic and technological levels. Once these issues have been addressed, the next question involves asking what type of questions can be asked of the data set given the model(s) in which the data are encoded. In the next chapter I contrast three different ways of modeling data and I describe in detail knowledge representation in computational theory and how it can be used to query the PHOIBLE data set from different perspectives. In

[122] This is an area of an ongoing debate. For recent discussions see Lazard 2006; Haspelmath 2007, 2010; Newmeyer 2010; Bickel 2010.

Chapter 4 I discuss how I bring together several different segment inventory databases into one large and interoperable cross-linguistic data set and in later chapters I use the different data models that I implement to ask questions of the segment inventories.

Chapter 3
DATA MODELING

There are many ways to model data. Some methods are well researched, considered mature and are used in all kinds of applications across many different industries. Other methods represent the state-of-the-art in data structures and algorithm design and are being researched and developed at the peripheries of computer science. While there are many different ways to think about and model data, different methods have different strengths and weaknesses for different purposes. Therefore it is necessary to model different data types with appropriate data structures to enable the desired questions to be answered. In this chapter I give a brief overview of some data modeling basics in Section 3.1. In Section 3.2 I describe the PHOIBLE data models in detail. Lastly, in Section 3.3 I discuss the details of knowledge representation and their implementation in RDF graph models as it pertains to modeling segments and distinctive features.

3.1 Data modeling basics

3.1.1 Table

Tabular data is a simple data set represented in a table such as a delimiter-separated text file, spreadsheet or HTML table. The table (or flat file) model is simple to read and easy to manipulate. It consists of a two-dimensional array of data elements. The placement of data in rows and columns provides the data with structure, and thus, meaning. A table's columns and rows specify relationships among the cells in the table, some of which are implicit. For example in Table 3.1 the `LangID` column identifies a set of three-letter ISO 639-3 language codes that are used to uniquely identify the set of languages in the current Ethnologue database (Lewis, 2009).[1,2] The language ID “dts”, or [dts], identifies the language name

[1] The Ethnologue language codes table is available online at: `http://www.ethnologue.com/codes/`.

[2] The full set of ISO 639-3 codes from SIL International are at: `http://www.sil.org/iso639-3/`.

"Dogon, Toro So", which is spoken in country ID "ML" (Mali) and has the language status "L" (living). This is one way to model data that associates unique language name identifiers with language names and information about where those languages are spoken and their status of endangerment.

Table 3.1: Language codes table

LangID	CountryID	LangStatus	Name
dgs	BF	L	Dogoso
dtm	ML	L	Dogon, Tomo Kan
dts	ML	L	Dogon, Toro So
dtt	ML	L	Dogon, Toro Tegu
dtu	ML	L	Dogon, Tebul Ure

Modeling data in a table has limitations. Consider the tabular data in Table 3.2, which is an expansion of Table 3.1 with an additional column for specifying alternative language names; they are separated by commas. The data in the table cannot be easily sorted to discover that "Dogon, Toro So" [dts] spoken in Mali has an alternative language name "Dogoso", which is the same name as a different language spoken in Burkina Faso, also called "Dogoso" [dgs].

To illustrate a more complicated example, let's add a column to specify each language's genealogical affiliation. A fine example is provided by Dogon, a language family whose position relative to other African language families is unclear.[3] Adding the language family and its citation forces too much data into the table as shown in Table 3.4. Individual fields now store different values. The situation is hopeless if the user wants to compare competing

[3] In comparison to many other language families in West Africa, Dogon is lexically and structurally different. Dogon languages have an unusual combination of agglutinating verbal morphology and isolating nominal morphology. They have SOV word order and do not have noun classes that are associated with Niger-Congo languages (Heath, 2008). See Hochstetler et al. 2004 for a historical overview of the genealogical classifications of Dogon. See the Dogon Languages Project for our current understanding of Dogon languages: `http://dogonlanguages.org/`.

Table 3.2: Language codes table augmented with alternative language names

LangID	CountryID	LangStatus	Name	AltLangName
dgs	BF	L	**Dogoso**	Bambadion-Dogoso, Bambadion-Dokhosié, Black-Dogose, Dorhosié-Finng, Dorhosié-Noirs, Dorossié-Fing
dtm	ML	L	Dogon, Tomo Kan	Tomo-Kan
dts	ML	L	Dogon, Toro So	Bomu Tegu, **Dogoso**, Toro So
dtt	ML	L	Dogon, Toro Tegu	Tandam
dtu	ML	L	Dogon, Tebul Ure	

trees for a particular language family or to compare two or more resources' descriptions of different families. More sophisticated data models are needed to access the relationships encoded in the data. Rather than tightly packed table data, our data model needs to be broken out into multiple tables, each of which reference the same data.

Table 3.3: Abbreviated history of the classification of Dogon

Year	Classification	Authors
1981	Voltaic (English: “Gur”)	Manessy (1981)
1981	Volta-Congo	Bendor-Samuel and Hartell (1989)
1994	Unresolved; non-classified	Plungian and Tembiné (1994)
2000	Ijo-Congo	Williamson and Blench (2000)
2005	Volta-Congo	Gordon (2005)
2009	Volta-Congo	Lewis (2009)

Table 3.4: Language codes table with proposed language families

LangID	Name	LangFamily
dgs	Dogoso	Gur (Gordon, 2005), Gur (Lewis, 2009)
dts	Dogon, Toro So	Voltaic (Manessy, 1981), Volta-Congo (Bendor-Samuel and Hartell, 1989), Unresolved; non-classified (Plungian and Tembiné, 1994), Ijo-Congo (Williamson and Blench, 2000), Volta-Congo (Gordon, 2005), Volta-Congo (Lewis, 2009)

3.1.2 Database

A database is a mechanism that stores data and it can be modeled, or structured, in different ways. A relational database is a particular type of database model that consists of a set of tables that are joined, or related, in a standardized way. The relational database model was introduced in Codd 1970 and is based on set theory and predicate logic. Relational databases are the mature product of decades of research, optimization and the financial backing or open source development of products like Oracle DB or MySQL (Hebeler et al., 2009). They are fast and powerful tools and their data models and design patterns are well researched and understood. A relational database is typically what one is talking about when the term *database* is used. The structure of a database is defined in a formal language, the product of which is called a database schema. The database schema defines the logical grouping of tables (and other database elements like views, procedures, etc.) and is essentially a blueprint of the database's structure.

Relational tables provide two basic operations: retrieving a set of columns and retrieving a set of rows. The SQL query in Example 3.1 retrieves the columns and results displayed in Table 3.5 from the data in `LanguageCodes` table that was given in Table 3.1 on page 77.

(3.1)
```
SELECT LangID, Name
FROM LanguageCodes
```

Table 3.5: Results from a basic operation to retrieve database columns

LangID	Name
dgs	Dogoso
dtm	Dogon, Tomo Kan
dts	Dogon, Toro So
dtt	Dogon, Toro Tegu
dtu	Dogon, Tebul Ure

The query in Example 3.2 retrieves the set of rows displayed in Table 3.6 from the data in the `LanguageCodesAlternativeNames` table that was given in Table 3.2 on page 78.

(3.2)
```
SELECT *
FROM LanguageCodesAlternativeNames
WHERE CountryID = “ML”
```

Table 3.6: Results from a basic operation to retrieve database rows

LangID	CountryID	LangStatus	Name	AltLangName
dtm	ML	L	Dogon, Tomo Kan	Tomo-Kan
dts	ML	L	Dogon, Toro So	Bomu Tegu, Dogoso, Toro So
dtt	ML	L	Dogon, Toro Tegu	Tandam
dtu	ML	L	Dogon, Tebul Ure	

Fundamentally, a relational database is a set of tables, which themselves are made up of sets of rows and sets of columns. Therefore, set operations on tables can be performed on two or more tables, allowing users to perform operations like intersection, cartesian product, adding or subtracting tables from each other, etc. The real power of relational databases becomes apparent when operations are made on sets of tables that are not the same, but that share at least one column.

Let’s look at the data in Table 3.2 for language codes and alternative language names again. One possible way to model these data in relational database tables is shown in the database schema in Figure 3.1, where the two tables `LanguageCodes` and `AltLangNames` are joined by a one-to-many relationship on the `LangID` fields, which contain the ISO 639-3 three-letter language identifiers.[4] In the `LanguageCodes` table, the `LangID` is the

[4] In Section 3.2.1 I provide an overview of how to read and interpret the relational database schema notation used in this work.

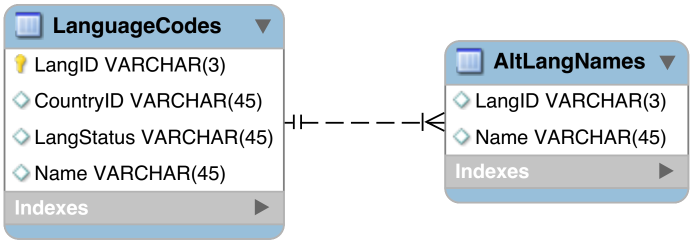


Figure 3.1: Language codes and alternative language names schema

primary key, i.e. a key that uniquely identifies each row in the table. It cannot be NULL.[5] The LangID column in the AltLangNames table represents a foreign key, i.e. a referential constraint that matches to the LangID primary key in the LanguageCodes table. This relationship, visualized with the dotted arrow in crow's feet notation, indicates that the LangIDs in AltLangNames are in a many-to-one relationship with the LangIDs in the LanguageCodes table. Thus, the foreign key cross-references the data in these two tables. Examples of the tables with data are shown in Tables 3.7 and 3.8.

Table 3.7: LanguageCodes table

LangID	CountryID	LangStatus	Name
dgs	BF	L	Dogoso
dts	ML	L	Dogon, Toro So

Instead of sorting or filtering the initial language codes and alternative language names in Table 3.2 on a single column, the relational model allows more sophisticated queries. For example, a query to find language names that are identical to alternative language names is given in 3.3, which returns the result data set in Table 3.9.

[5] NULL is a special value that indicates a value does not exist. NULL represents missing or inapplicable information.

Table 3.8: AltLangNames table

LangID	Name
dgs	Dogoso
dgs	Bambadion-Dogoso
dgs	Bambadion-Dokhosié
dgs	Black Dogose
dgs	Dorhosié-Finng
dgs	Dorhosié-Noirs
dgs	Dorossié-Fing
dts	Bomu Tegu
dts	Dogoso
dts	Toro So

(3.3)

```
SELECT LanguageCodes.LangID,
  LanguageCodes.Name,
  AltLangNames.Name as AltName,
  AltLangNames.LangID as AltLangID
FROM LanguageCodes
JOIN AltLangNames
ON LanguageCodes.Name = AltLangNames.Name
```

In this relational model the meaning, or *semantics*, of the data are more explicitly stated. The database schema describes the meanings of the values and specifies there is a relationship between `LanguageCodes` and `AltLangNames`. The database does not know what these entities are, but they can be structured and joined in ways to be queried. Our example of language classifications would also require a schema that describes the relationships between languages' proposed genealogical classifications and the citations for those theories.

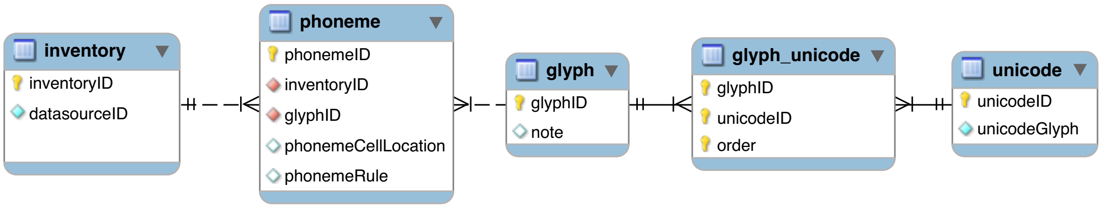
Table 3.9: Query result

LangID	Name	AltName	AltLangID
dgs	Dogoso	Dogoso	dgs
dgs	Dogoso	Dogoso	dts

3.1.3 Graph

As shown, the table and relational database structures have different methods for information retrieval. In this work I describe how information can also be modeled in a graph data structure, and *knowledge* through logical statements, can be added to it to create a *knowledge base.* To the programmer, a graph is a fundamentally different data structure than a relational database model.[6] Interacting with graphs requires different programming approaches. Contrast the representations of a portion of PHOIBLE's relational database schema in Figure 3.2 and its graph implementation in Figure 3.3.[7]

Figure 3.2: PHOIBLE database schema for segments

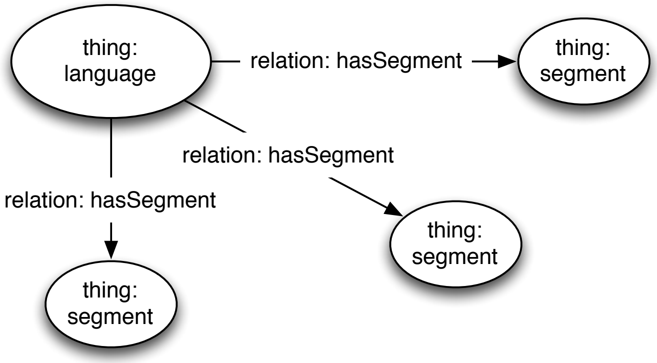


They illustrate two different ways of representing information or knowledge about a language and its segment inventory. The relational database model in Figure 3.2 is designed to query

[6] The term *graph* is polysemous. In its data visualization sense, a graph is a diagram showing a relation between variables on a pair of axes. This type of graph is also called a *plot.* In its stricter mathematical sense, a graph is a collection of objects connected by links. In its computer science sense, a graph is an abstract data type (or structure) that implements the mathematical concept of graph.

[7] See Section 3.2 for a detailed explanation of PHOIBLE's relational database and graph models.

Figure 3.3: PHOIBLE graph model for segments



languages' segments, e.g. by joining the `inventory`, `phoneme`, `glyph`, `glyph_unicode` and `unicode` tables and executing a SQL query. The query in Example 3.4 selects the phonemes and inventory ID for the inventory identified with *N*, where *N* is an integer in the range of inventory IDs.

(3.4)
```
SELECT phoneme.phonemeID, inventory.inventoryID
FROM phoneme
JOIN inventory
ON phoneme.inventoryID = inventory.inventoryID
WHERE inventory.inventoryID = N
```

In the relational database model in Figure 3.2, an inventory has one or more phonemes, which map in a many-to-one relation to a glyph and so on. The mechanics and reasoning of this model are explained in Section 3.2.1.

Now compare the relational database model with the graph model in Figure 3.3, which illustrates how one might model language and segment objects and the relation between these objects.[8] This graph data structure represents a collection of *statements*, sometimes called *facts*, about knowledge that we have. In this simple model, each node is a *concept*,

[8] Figures in this work use a simple graph visualization for illustrating concepts and their relations. Other methods can be used to visualize these relationships, such as UML diagrams.

or *entity*, representing languages and segments (here prefixed with "thing:"). Each link between concepts encodes a relationship between nodes. Both relational database and graph models can be queried when these designs are implemented in tools like a MySQL relational database or an RDF/OWL knowledge base (Lassila and Swick, 1999; Smith et al., 2004). The knowledge base can be queried with SPARQL (Prud'Hommeaux and Seaborne, 2006),[9] an RDF query language. SPARQL queries consist of triple patterns that match concepts and their relations by binding variables to match graph patterns. A SPARQL query to retrieve a list of segments from the PHOIBLE graph models for segments is given in Example 3.5.

(3.5)

```
SELECT ?segments
WHERE {
thing:language relation:hasSegment ?segments
}
```

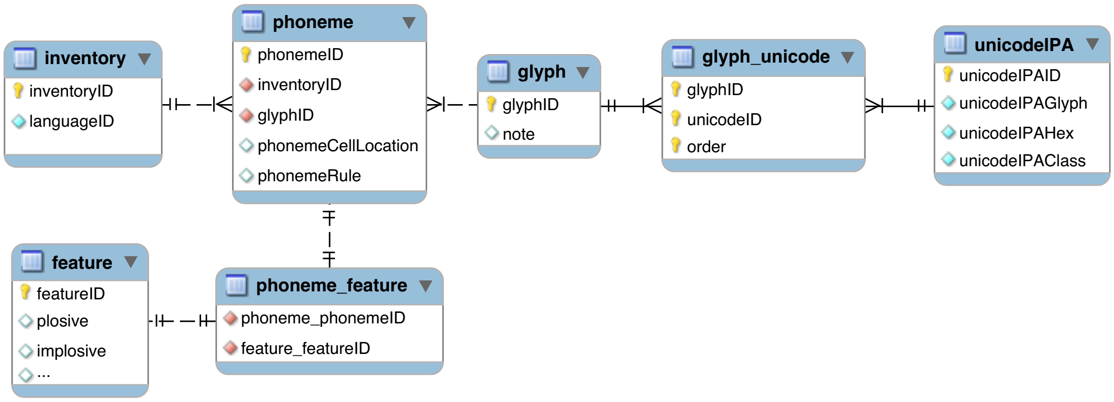
The SPARQL query matches sets of triples that contain thing:language as the subject and relation:hasSegment as the predicate. Because of the loose structure of graphs and the ability to define any type of relationship, the knowledge base approach enables higher levels of information expressiveness. For example, a database may constrain a data type (e.g. text with length of 3 characters), but not its use (e.g. values between aaa-zzz, exclusive of the range qaa-qqq). The programming application that uses the data must deal with the lack of expressiveness, causing the knowledge to be distributed between programming instructions and data storage. In the knowledge base implementation, relationships take on the primary role. Whereas in an object oriented approach relationships are dependent on an object class definition and do not exist outside of its associated class, in the knowledge base approach relationships can join to any collection of statements. They are not permanently bound to any class, can assign multiple classes to any given instance and provide information that is independent from object class definitions.

There are several differences between the relational database model and the graph data structure to point out. First, the relational database in Figure 3.2 on page 84 depends on

[9] SPARQL Protocol and RDF Query Language

a schema for structure. This schema is provided in a different language than the relational database's implementation. In a graph knowledge representation model, as in Figure 3.3 on page 85, the same knowledge representation language can be used to form the knowledge base's structure and data instances because the knowledge base depends on ontological statements to define its structure. This is an advantage because the schema is not decoupled or defined in a different language than the model. Second, the relational database is limited to one kind of relationship – the foreign key. The foreign key relates a set of columns that link information, such as the `LangID` column in Tables 3.7 and 3.8 on page 82. On the other hand, the structure in Figure 3.3 depends on ontological statements, also called *commitments*. Importantly, these statements offer multidimensional relationships, including logical relationships and constraints. And third, adding new knowledge to a relational database is more challenging than adding it to a knowledge base. Relational databases can easily include new data in rows, adding to the database's contents. However, to add new knowledge, the schema needs to be adapted and updated, and new tables or columns must be added or updated. On the other hand, the knowledge base's statements define its schema, individuals and instances. The self-describing structure of the knowledge base supports a model of open and shared data. Let us take for example the problem of updating the data models in Figures 3.2 and 3.3 to include new knowledge about distinctive features.

Figure 3.4: PHOIBLE database schema for segments and features

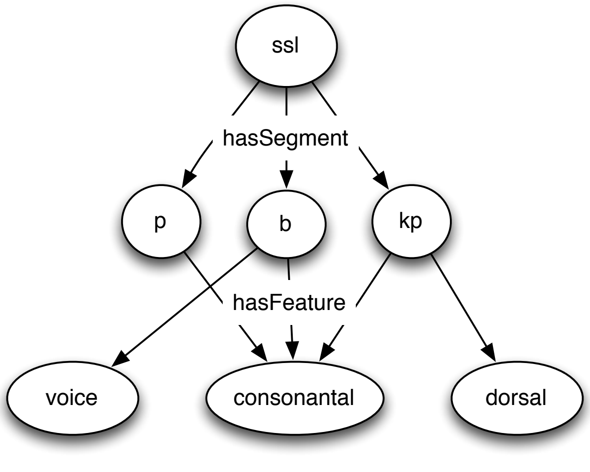


In Figure 3.4, new relational database tables are added to incorporate one set of distinctive features (if multiple distinctive feature sets were added, a decision between incorporating them all into one table or adding additional tables per feature set would be made). This is a simple example where the table's cells might look like that given in Table 3.10. If you want to search the database for languages that contain a natural class of sounds based on certain features, your SQL queries become more complicated and now include JOIN clauses to combine records from multiple tables. Additionally, to encode competing distinctive feature sets the schema needs to be extended.

Table 3.10: Example features table

phoneme_id	plosive	implosive	ejective_stop
1	FALSE	FALSE	FALSE
2	TRUE	FALSE	FALSE

Figure 3.5: PHOIBLE knowledge base for segments and features



In Figure 3.5, features are added to the graph by linking them from each segment via a "hasFeature" predicate that we defined with a URI.[10] To query the knowledge base

[10] In the following examples I use an ISO 639-3 three letter language name identifier to symbolize a

with SPARQL, the user can specify multiple graph patterns within a query. Example 3.6 shows how to use SPARQL to query for all languages in the knowledge base that contain segments that are voiced consonants. First we use SPARQL to query the segments of [ssl] and then query the returned graph that contains the segments and their features matching, for example, some natural class constraint. With SPARQL query solution sets can also be used to construct new graphs with the CONSTRUCT command. By explicitly encoding the relationships between concepts logically, new triples that contain implicit knowledge can be inferred and then added back to the graph, thereby increasing the graph's representation of knowledge. In general it is much simpler to add new knowledge to the graph data structure than the relational database model. For example, in Figure 3.6 we can associate different distinctive feature sets with the set of features that may or may not have the same extension. Now the graph can be queried to compare the overlap between different distinctive feature sets.

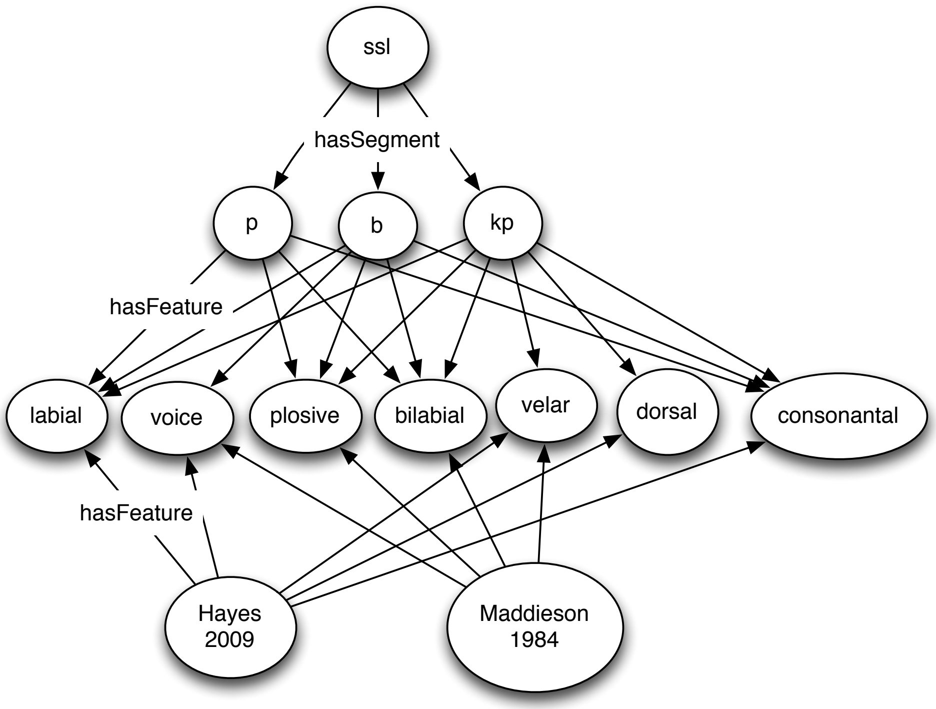
(3.6)
```
SELECT ?languages
WHERE {
?languages hasSegment ?segments
?segments hasFeature voice
?segments hasFeature consonantal
}
```

These examples illustrate a fundamental difference in the importance of data modeling. The knowledge base is a data-centric model. In comparison to individually devised relational databases, the knowledge base facilitates data sharing by publishing a self-describing data model according to explicitly encoded relationships found in the data. This model adheres to a set of design principles and enabling technologies developed by the World Wide Web Consortium (W3C) under the rubric of the often misunderstood "Semantic Web" (Berners-Lee et al., 2001).[11] Formal specifications in the Semantic Web, like the Resource

"thing:language" concept. To symbolize "thing:segment", I use letters like <p> to represent phonetic segments. I simply annotate predicate relations with camel-backed phrases, e.g. "hasSegment".

[11] See http://www.w3.org/2001/SW/ for a list of papers published by W3C on the Semantic Web.

Figure 3.6: PHOIBLE knowledge base for segments, features and feature sets



Description Framework (RDF) (Lassila and Swick, 1999; Beckett, 2004), the Web Ontology Language (OWL) (McGuinness and van Harmelen, 2004) and SPARQL (Prud'Hommeaux and Seaborne, 2006), provide a common framework for formally describing concepts, terms and relationships in a particular domain of knowledge. Linguists should care because this framework provides them with the opportunity to encode data in a way that is arguably more transparent than using a relational database schema. Therefore, data that are published become more easily reusable and they have the potential to reach a larger audience and have greater impact on research.

This graph data model is more dynamic and allows information to be added at any point. The ability to easily add, update and share data is attractive for resources capturing linguistic knowledge, e.g. data from the field that is undergoing analysis. Data are often collected, analyzed, reanalyzed and used in the development of linguistic theory in various subfields like syntax, morphology, semantics, and phonology, cf. Bender & Langendoen 2010 and Pericliev 2010. However, different annotation schemes, community or discipline-specific terminology, and different standards often prohibit easy data sharing within and across subfields. An added benefit of modeling knowledge in a Semantic Web framework is that it enables easy data sharing and data transformation. For example, ontologies have successfully been used in linguistics for tasks like terminology resolution and for interoperating over disparate transcription systems. An example of an ontology for morphosyntactic terminology is the General Ontology of Linguistic Description (GOLD) (Farrar and Langendoen, 2003). It is being used as a pivot to resolve different morphosyntactic annotations (that actually indicate the same morphosyntactic function) across lexicons from 16 different projects and several hundred languages.[12] Another example is an ontology for connecting a collection of languages' heterogeneous orthographies and their phoneme inventories to an interlingual pivot (Moran, 2009). The Ontology for Accessing Transcription Systems (OATS) provides users with a knowledge base that can answers questions of its content like, "How many languages contain the voiced palatal nasal /ɲ/ and how is it graphemically rendered in those languages?".

Each data structure has its tradeoffs, virtues and deficiencies.[13] A drawback of the RDF graph model, one that also gives it its flexibility, is that anyone can define any triple using his or her own naming conventions. Users can also use their own data modeling approaches. Allowing anyone to say anything about anything can obviously lead to miscommunications. OWL is the ontology language that logically marks up the RDF data structure to address this drawback. It can be used to restrict what can be logically stated about what. However, two features of OWL that also give it its flexibility can also be considered drawbacks. The first is commonly referred to as *the open world assumption.* The open world assumption, from formal logic, states that the truth value of a statement is independent of whether or

[12] For example, see the Lexicon Enhancement via the GOLD Ontology (LEGO) project: http://linguistlist.org/projects/lego.cfm.

[13] The tradeoffs in data structures can be thought of as analogous to the tradeoffs in different visualization designs in charts. The bar, line and bubble charts display different information. Each has its advantages and disadvantages for representing data visually and which is best depends on the task at hand.

not it is known to be true. This means that not knowing whether or not a statement is explicitly true, does not imply that the statement is false. This is the opposite of the *closed world assumption*, in which any presumption that is not known to be true is false. The second assumption is the *no unique names assumption*. This means a user cannot assume that any resources (concepts or relations) identified by different URIs are actually different.

As shown so far in this section, there is little semantic knowledge represented in the data structure of the relational databases. Their tables follow relational database design principles of normalization, but they do not describe the data in a meaningful way that applies fundamental concepts of an open world of data (there is always more information to be added) or of a non-unique naming convention (the same concept or entity can be known by more than one name). RDF graph data structures are more portable than standard relational databases, allowing anything to be said about anything. This generalized notion of a resource allows RDF statements to describe concrete or abstract concepts by using a single universal namespace built with URIs. URIs provide a foundation for data-sharing infrastructure because every statement unambiguously describes a particular resource, regardless of where that named resource resides in the graph. In this sense, any resource in an RDF graph can have any assertions made about it, even conflicting ones. Table 3.11 presents a comparison of the features relational databases and RDF/OWL knowledge bases.

In this work I have created several ways to access the PHOIBLE data set, including a relational database and an RDF graph.[14] These different models serve different purposes. In my opinion, the graph data structure uses a technology that embraces principles towards a cyberinfrastructure in linguistics, i.e. technological infrastructure for computational methods and research.[15] The main benefit with this data structure is that of data sharing. Because of global scope, the triple structure that makes up the graph allows for easy information integration. Two graphs from different sources that share a given URI can be merged without transforming the data. Figure 3.7 is one way to visualize RDF structure in which a point is defined by the intersection of its subject, predicate and object in a three

[14] See Section 3.2.

[15] See Section 8.4.6.

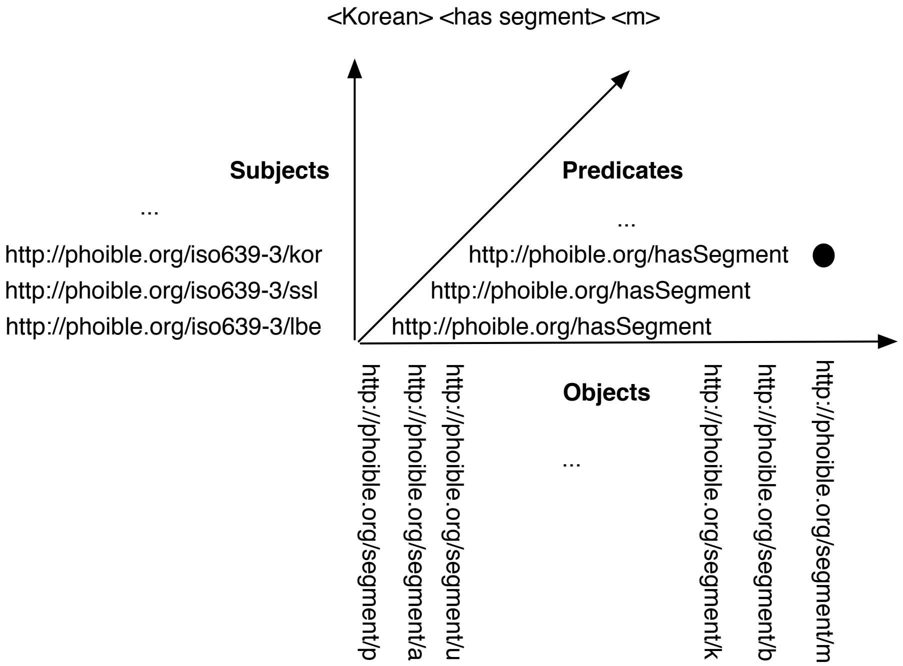
Table 3.11: Comparison of relational databases and knowledge bases (Hebeler et al., 2009, 9)

Feature	**Relational Database**	**Knowledge base**
Structure	Schema	Ontology statements
Data	Rows	Instance statements
Administration language	DDL	Ontology statements
Query language	SQL	SPARQL
Relationships	Foreign keys	Multidimensional
Logic	External of database/triggers	Formal logic statements
Uniqueness	Key for table	URI

dimensional space (Hebeler et al., 2009, 73). This visualization illustrates three principles of data sharing with RDF graphs: easy merging, no order and no duplicates. Two or more sets of points can easily be merged by overlaying them on top of each other; thereby forming a richer graph if two or more graphs are merged. Graph structures do not have root nodes like tree representations, such as XML. In an XML document for example, all nodes are in a hierarchical relationship with the root node. Thus the tree structure defines the orientation of elements. Merging two or more trees can be challenging because the merged tree requires a root, which must be determined from the roots and internal structure of the trees being merged. In other words, complementary information in two or more XML documents requires that the different elements be defined in their relationships to one and another. In RDF graph structures, there is no root node, so merging the graphs is trivial since there is no order to the elements. Also, as graphs are merged, if any statements with subjects, predicates and objects are identical, they will not be duplicated.

Compare these features with a traditional relational database approach that must join tables of data on IDs. Databases from different projects will have different schemas and

Figure 3.7: RDF statements as points


<Korean> <has segment> <m>
Subjects
...
http://phoible.org/iso639-3/kor
http://phoible.org/iso639-3/ssl
http://phoible.org/iso639-3/lbe
Predicates
...
http://phoible.org/hasSegment
http://phoible.org/hasSegment
http://phoible.org/hasSegment
●
Objects
...
http://phoible.org/segment/p
http://phoible.org/segment/a
http://phoible.org/segment/u
http://phoible.org/segment/k
http://phoible.org/segment/b
http://phoible.org/segment/m

different IDs. These IDs and relationships must be identified before data can be merged. An RDF graph is a data structure of self-contained assertions of information in a single global namespace, so it is easy to merge sets of points. Merging two graphs therefore makes a richer graph of information. To address the drawbacks of the open world and no unique naming assumption, which in actuality let linguists model their own data and use their own terminology, RDF graphs that model linguistic data can be given (or linked to) GOLD URIs (Farrar and Langendoen, 2003; Farrar and Lewis, 2005; Farrar and Langendoen, 2010).[16] Finally, knowledge representation in graphs is taken further by associating logical statements on links between resource nodes. These ontological commitments create a knowledge representation structure that allows logical inference to be made on the data.[17]

[16] http://linguistics-ontology.org/

[17] See Section 3.3.

3.2 PHOIBLE data models

In this work, a major challenge has been to standardize and merge data from different published resources and different databases, so that the data are interoperable at the linguistic and technological levels thereby providing linguists with cross-linguistic data to undertake typological analyses.[18] My discussion of the development of PHOIBLE's data models begins with the relational database because its structure allowed me to combine data from disparate sources in a modular fashion using flat files and ISO 639-3 codes as keys. This design allows different portions of the database to remain separate and easily updatable (e.g. PHOIBLE contains genealogical data from Ethnologue (Lewis, 2009) and WALS (Haspelmath et al., 2008), which are periodically updated). I then devised a procedure to aggregate the different data sources together and denormalize them into reporting data warehouse flat files, which are ideal for statistical software packages and for computer programs.

Initially I wanted to use just a graph data model, but unfortunately it does not capture all the information that I am interested in. For example, although a graph model is ideal for merging data sets, it is exactly this quality of removing duplicate data points that does not allow me to capture the distinction between the number of segment types and segment tokens in the combined PHOIBLE data set without having to write a specialized query to generate these data. I could have also combined the contents of different data sources into one large flat file table. However, due to the size and scale of the data involved, updating or changing a table of over 50k rows would be difficult and impractical. Therefore the relational database, which provides constraints that ensure referential integrity, is the data model that I use to combine separate resources. It is described first in Section 3.2.1.

The complexity of the PHOIBLE relational database model is not ideal for easily querying the data. As I will show, relational databases can be complex data models that require specialized training to understand and work with. One output format of a relational database is a denormalized data warehouse flat file table. Flat files can be queried using a set notation like SQL. In this work two flat files are created from a data warehouse SQL

[18] The individualized "extract, transform, load" processes for each database subsumed by PHOIBLE are discussed in Section 4.3.

script, discussed in Section 3.2.2. PHOBILE's flat files are also useful as input for developing other data models. In Section 3.2.3, I will discuss how I generate an RDF graph model from the flat file contents.

The PHOIBLE relational database, data warehouse flat files and RDF graph model and their path of development is depicted in Figure 3.8. The different data sources (SPA, UPSID_{451}, AA and PHOIBLE inventories) are merged into a relational database and data from the Ethnologue, Multitree, WALS, Unicode and the CIA World Factbook are added and connected to segment inventories via ISO 639-3 codes. The data warehouse procedure is then applied to create two flat files: an aggregated version of segment inventory data and a phoneme level version. Python scripts transform these flat files into RDF graph files, which can then be merged via RDF graph model technologies. Thus PHOIBLE is published in three formats: plain text flat files, a relational database and an RDF graph. The transformation process is automated so that when new data are added they can be processed and the three models can be updated.

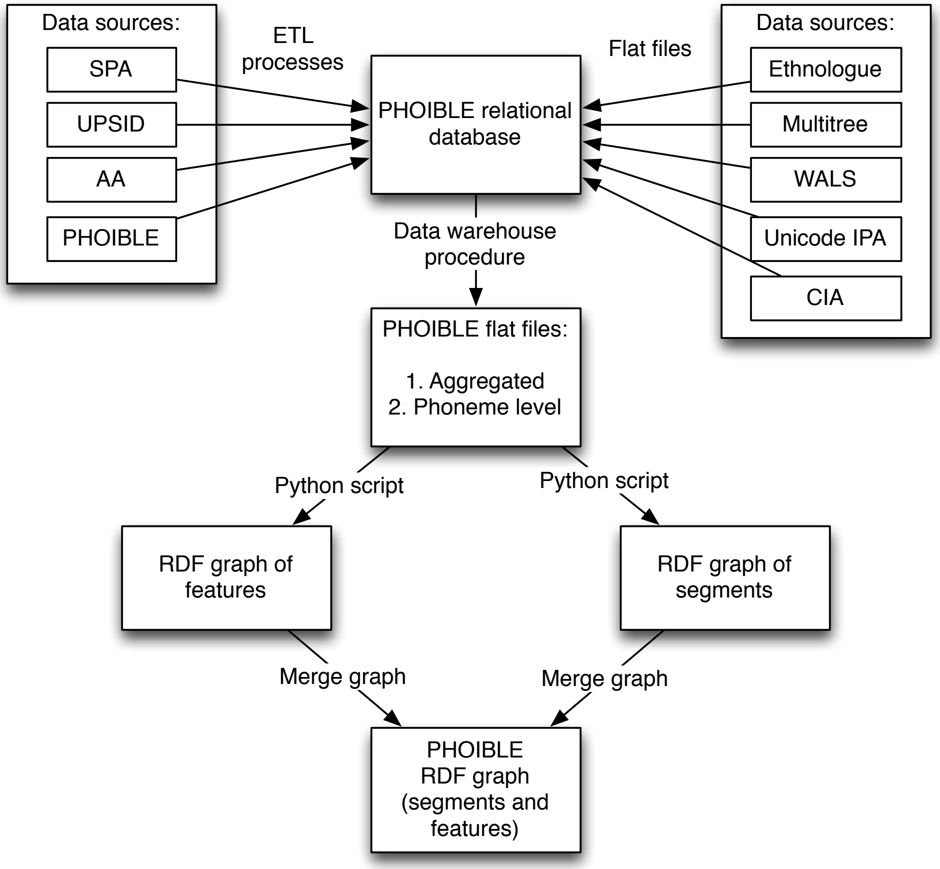
The design of a data model depends on the aims of the questions to be answered of the data being modeled. In my work there has been no one-model-fits-all-queries solution. I suspect this is the case for any large cross-linguistic resource. In the following sections I describe in detail the different PHOIBLE data models and how they can be used to undertake phonological typology.

3.2.1 Relational database

I developed the PHOIBLE database in MySQL,[19] a popular, free and open source relational database management system that has Unicode support and is easy to integrate into Web applications. The main reason to use a relational database model is to impose *referential integrity*. Referential integrity is a database concept that ensures that any data shared between tables remains consistent and synchronized. Since there are several distinct sources from which I take data to populate tables, referential integrity helps prevent inconsistent data from entering the database. When this property is satisfied, data quality issues such

[19] `http://www.mysql.com/`

Figure 3.8: Path of development for PHOIBLE data models



as spurious duplication are avoided because each foreign key value in a table must exist as a primary key in the referenced table. This will become clearer with some examples, below.

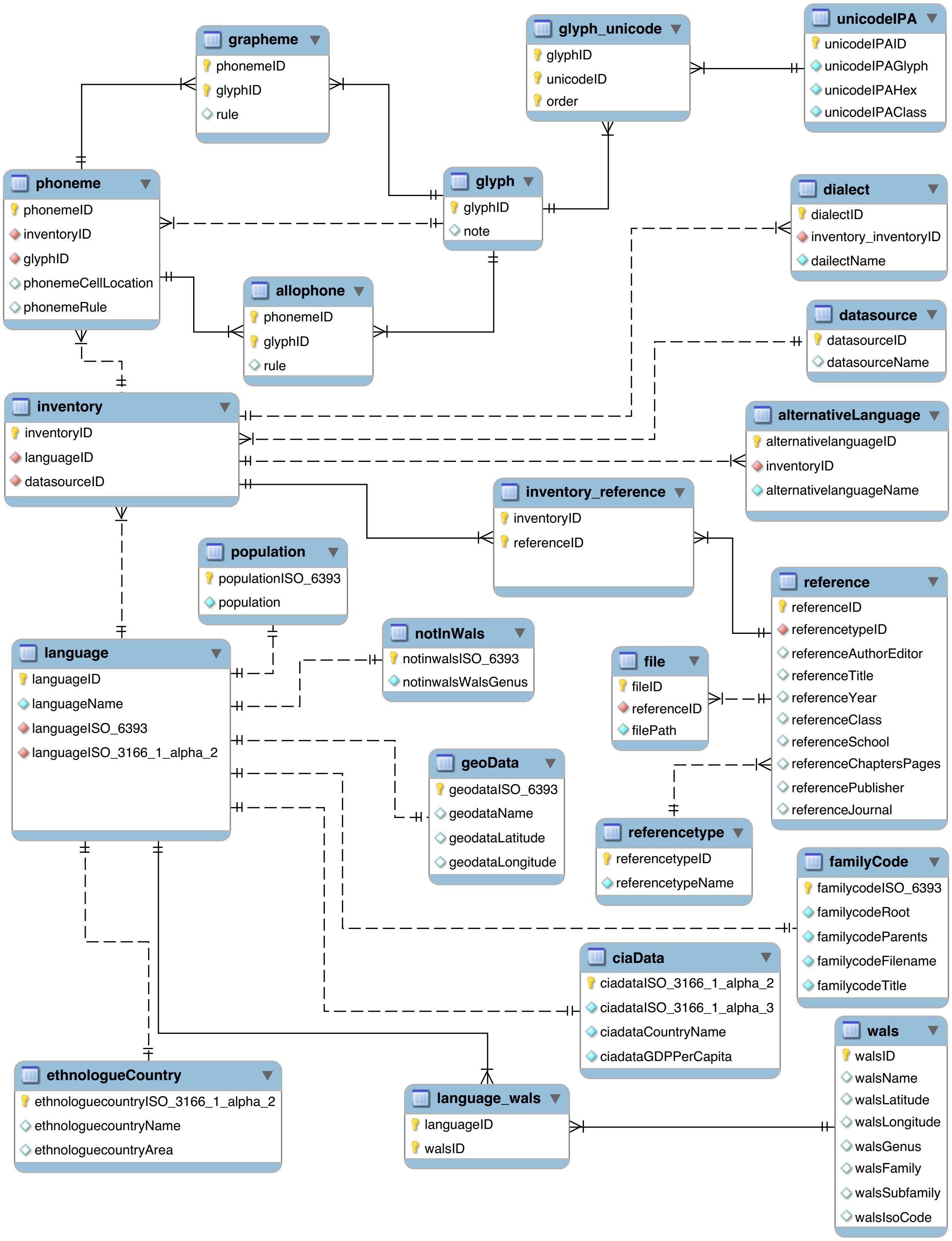
Another important part of database modeling is normalization. Normalization is the process of organizing data into tables to minimize duplication across tables. It is a modeling technique used to optimally design a database to minimize redundancy of data. The duplication of data in different tables should ideally be kept to a minimum. Instead of duplication, values in one column in a table may depend on values in a column in another table, the relationship of which is often controlled through the use of foreign and primary keys. Thus normalization supports data integrity and efficient modeling. Normalization forms are a series of conditions to ensure that a database is normalized (Codd, 1970). They are used to determine logical inconsistencies within the database's design. When designing a relational database, the ideal is to get the Third Normal Form (3NF). First Normal Form (1NF) means that all columns are atomic, i.e. there is a separate table for each set of attributed with a primary key, so that there are no repeating items in columns. In Second Normal Form (2NF), the database is in 1NF and every non-key column is dependent on a primary key. Third Normal Form adhering databases are in 2NF and every non-key column is mutually independent.

A large normalized database typically requires complex queries that involve joining multiple tables. A denormalized database allows more data redundancy, which makes querying the data simpler. In some areas my database does not conform to 3NF. This was done to allow certain frequently updated data sets, like the ISO 639-3 codes or language family data, to be more easily updated. Other areas of my database, such as the segment inventory and reference data, do conform to 3NF because these data are relatively static, e.g. a bibliographic resource and the segment inventory extracted from it are not a data source that is likely to change.

Section 3.1.2 provided some simple examples that illustrated the basic functionality of how data can be retrieved from a relational database. A driving factor for database normalization is that larger databases that include many tables that encode different sources of data are often much more complex than the simple examples I provided. One way to conceptualize and graphically represent a relational database model (aka database schema) is with an entity-relationship model, introduced by Chen (1976). There are many variants of the entity-relationship model. In this work I use the extended entity-relational model (EER). An EER is a logical diagram that is ideally self explaining, although deciphering it takes a bit of background knowledge.[20] Figure 3.9 is an EER diagram of the current PHOIBLE relational database schema.

[20] The EER diagrams in this work were produced with MySQL Workbench. See: `http://www.mysql.com/products/workbench/`.

Figure 3.9: PHOIBLE database schema



The components used in the EER diagram include entities and relationships. Each box in the diagram represents an entity, here specifically a table. Each table contains a number of data items (or fields). Data items can be in different formats, e.g. numeric values like integers, floats and decimals, various date and time formats, different varieties of strings and text, etc. Each data item in a table has a symbol to its left in the EER diagram. A primary key is a unique identifier in a table and is symbolized with a golden key.[21] A red diamond denotes a foreign key. A foreign key is the primary key from another table. A blue diamond represents a field that has to be populated, i.e. it cannot be NOT NULL. A clear blue diamond is the opposite; it is a field that can be NULL. Again, NULL is a special value that represents missing or inapplicable information. NULL differs from an empty cell, which indicates the absence of data (e.g. the `referenceSchool` field in the `reference` table is left empty when a bibliographic record is not a PhD dissertation or Master's thesis).

A relationship between two tables is represented by a connecting line. In this EER diagram, I am using Crow's Foot (also Crow's Feet) notation, developed by Everest (1986) (who originally used the term "inverted arrow"). A relationship illustrates an association between two tables. Two dashes, which look like a perpendicular equals sign on the line, indicates the "one" side of a relationship. A perpendicular line with "crow's foot" extended is the many side of a relationship. There are three basic types of relationships:

1. **one-to-one**: joining two key fields (generally one primary key to one foreign key)

2. **one-to-many**: one particular value on one side of the relationship can have many values

3. **many-to-many**: many values on one side of the relationship have the possibility of mapping to more than one relationship

A final note about about crow's feet notation has to do with whether a line between two tables is dotted or full. A dotted line represents a non-identifying relationship. In a non-identifying relation, one thing can exist without the other, i.e. a child table can be identified

[21] A primary key is the concept of uniquely identifying values in a table; note that a primary key can be composite, i.e. in the `allophone` table the primary key is the composite of a `phonemeID` and a `glyphID`.

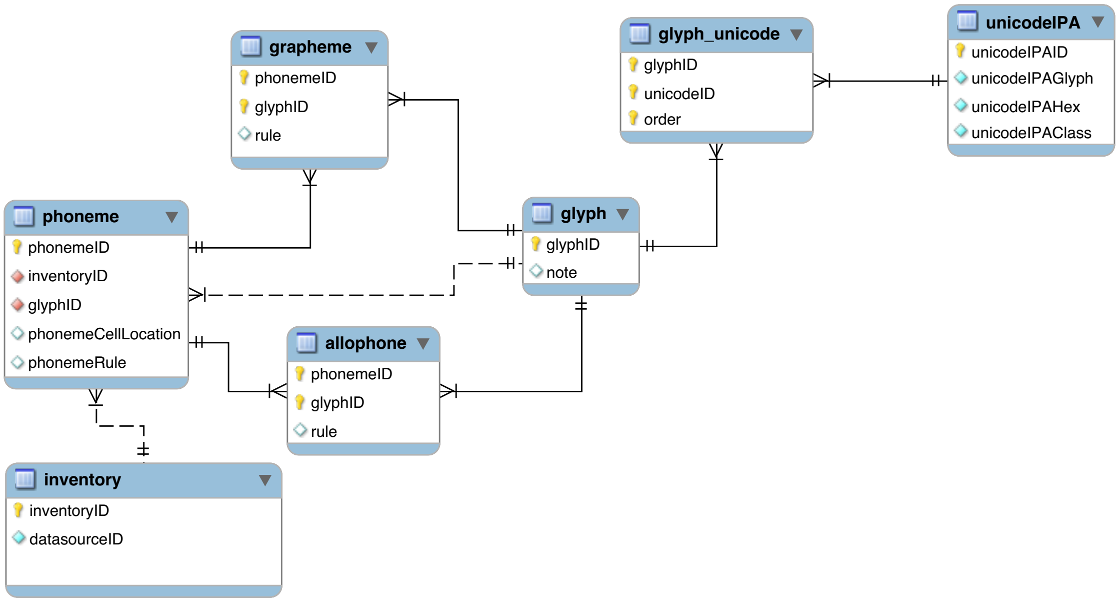
independently of the parent table. In an identifying relationship, the existence of a row in the child table is dependent on a row in the parent table. Putting this into a real world example, a grammar may be a book that belongs to a linguist. A linguist can also own multiple grammars. The grammar can change owners and the grammar can exist without an owner. The relationship between the grammar and linguist (as its possible owner) is a non-identifying relation. The grammar can exist without the owner. However, the grammar is also written by a linguist (who may or may not be the owner). The linguist may have also written more than one grammar. The grammars, of course, must be written by an author. The grammar would not exist without an author. Thus the relationship between the grammar and the author is an identifying relationship.

Using the EER diagram and crow's foot notation, in the rest of this section I will describe the PHOIBLE database schema presented in Figure 3.9 by starting from the inventory table. Following the inventory table upwards, I first describe the way in which I have modeled the relationship between segment inventories and their phonemes. I then explain how phonemes (and their allophones and graphemes) are modeled in respect to their Unicode representations. Second, moving from the inventory table downwards, I explain the design of additional language-specific data that is represented first by the relationship between the inventory and language tables and then with the many relationships between the language table and other tables to its right and below it. The language-specific information includes each language's population, geographic location, genealogical descent, etc. Lastly, by following the relationships from the inventory table to the right of the diagram, I describe the inventory-specific information including the bibliographic reference data for each inventory, as well as data extracted from source publications, such as author provided data on the dialect and any alternative language names.

I begin by describing how I modeled the relationship between an inventory, stored in the inventory table, and its phonemes that reside in the phoneme table. The relevant portion of the database schema is given in Figure 3.10.

The relationship between the inventory and phoneme tables is one-to-many because each inventoryID contains one or more phonemeIDs, i.e. each inventory has one or more phonemes. Although a phoneme is a theoretical concept that is language-dependent and

Figure 3.10: Inventories and segments



language-specific, in my database I have modeled a phoneme as if it can exist without an inventory. This is indicated by the non-identifying relation (dotted line) between the `inventory` and `phoneme` tables. I chose to model the relationship between an inventory and its possible phonemes as non-identifying because a phoneme is a contrastive sound that we can talk about independent of its occurrence in a given language. For example, when I refer to the sound /Θ/, another linguist familiar with IPA (or one who perhaps has an IPA chart handy) will know what kind of sound I am referring to, even if they cannot tell you what languages use that sound. The advantage of modeling the `phoneme` table in this fashion is that querying the number of rows in its table (on `phonemeID`) will return a number that is the total number of phonemes in all inventories represented in the database. Alternatively, if the shared data between the `inventory` and `phoneme` tables were more normalized, then all inventories that share the same phoneme, say /p/, would each map to the same `phoneme.phonemeID`.[22] In the case of /p/ occurring in *N*

[22] I use dot notation to indicate a data field within a table, e.g. `phoneme.inventoryID` refers to the

number of languages, there would be one and only one phonemeID that those different inventory.inventoryIDs would map to. This is precisely what an RDF graph model does; it eliminates redundancy by giving each object one unique identifier.

As I discussed in Section 2.3.3, the speech signal for the same sound in different languages shows a measurable difference, even if it is difficult to distinguish between the two. I felt it was important to model a phoneme in the database as a language-specific sound. This is what I call a segment token.[23] I also thought it was necessary to capture the relationship between different languages that contain the "same" sound. What I call the segment type relation is captured by the many-to-one relation from phoneme.phonemeID to glyph.glyphID. An illustration is given in Figure 3.11.[24]

Figure 3.11: Relations from inventory to glyph

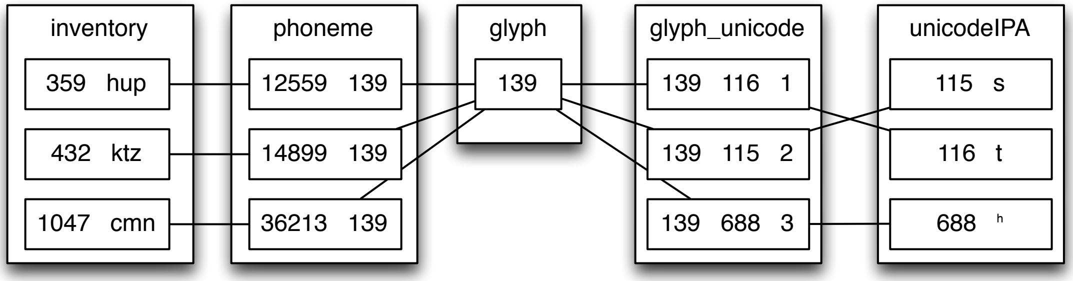


Figure 3.11 provides an example of how the three inventories (Hupi [hup], !Xu [ktz] and Mandarin Chinese [cmn]) each have a distinct phonemeID (12559, 14899, 36213) that maps to the same glyphID (139), which is a composite of Unicode characters that represent the segment /ts^h/. It also shows how I have mapped a glyphID to its Unicode

inventoryID field in the phoneme table and not the inventoryID in the inventory table.

[23] See Section 2.1.2.

[24] The inventory table in this illustration contains the inventoryID and the languageID data fields. The phoneme table contains phonemeID and glyphID. The glyph table contains the glyphID. The glyph_unicode table contains the glyphID, unicodeID and the order data fields. Finally, the unicodeIPA table contains the unicodeID and unicodeIPAGlyph fields. Data fields in the PHOIBLE schema that are not relevant to this example are excluded.

character components.[25] In this model, I capture both the segment token and segment type distinctions within the database's schema. Since every phoneme in an inventory is language-specific, this distinction is captured by the unique phonemeID in the phoneme table. These unique language-specific phonemes then map to the same glyphID. The glyphID appears in the phoneme table as a foreign key and as a primary key in the glyph table. The glyph.glyphID field is then in a one-to-many relation with the glyph_unicode table. In the glyph_unicode table, each row is represented with a combination of fields that include the glyphID, unicodeID and order data fields, respectively. Therefore the glyph_unicode table is a pivot table between the glyph and unicodeIPA tables. The glyph_unicode.order field is important because it stores the order in which a set of two or more Unicode characters combine to create a segment composed of more than one character. I have used the label "glyph" in table names and data fields loosely here. A glyph is a visual representation of a Unicode character. A character in Unicode is defined as the "smallest component of written language that has semantic value; refers to the abstract meaning and/or shape, rather than a specific shape".[26] Thus each character in Unicode is actually an abstraction of the different graphical forms (glyphs) of a grapheme.[27]

In my relational database schema, the phonemeID and glyphID fields are primary keys that are uniquely generated as the data from segment inventories are inserted into the database.[28] A unicodeID, however, is the decimal point that Unicode uses to uniquely encode a code point.[29] This decimal point makes for a practical and transparent unicodeIPAID in the unicodeIPA table. In this table, I also store each Unicode character's corresponding hexadecimal number in the unicodeIPAHex field and a graphical representation of each Unicode character in the unicodeIPAGlyph field. For each Unicode

[25] In Section 3.2.2, I show how I use additional information in the unicodeIPA table to provide a compositional break-down of each segment in the PHOIBLE database, as well as additional information about each segment's class, i.e. consonant, vowel or tone.

[26] http://unicode.org/glossary/#character

[27] See Section 2.1.4.

[28] See the descriptions of ETL processes in Section 4.3.

[29] From the unicodeIPA table I could easily link from each unicodeID to its Unicode character attributes, such as its name, canonical combining class, etc., via the public Unicode data: http://unicode.org/Public/UNIDATA/UnicodeData.txt.

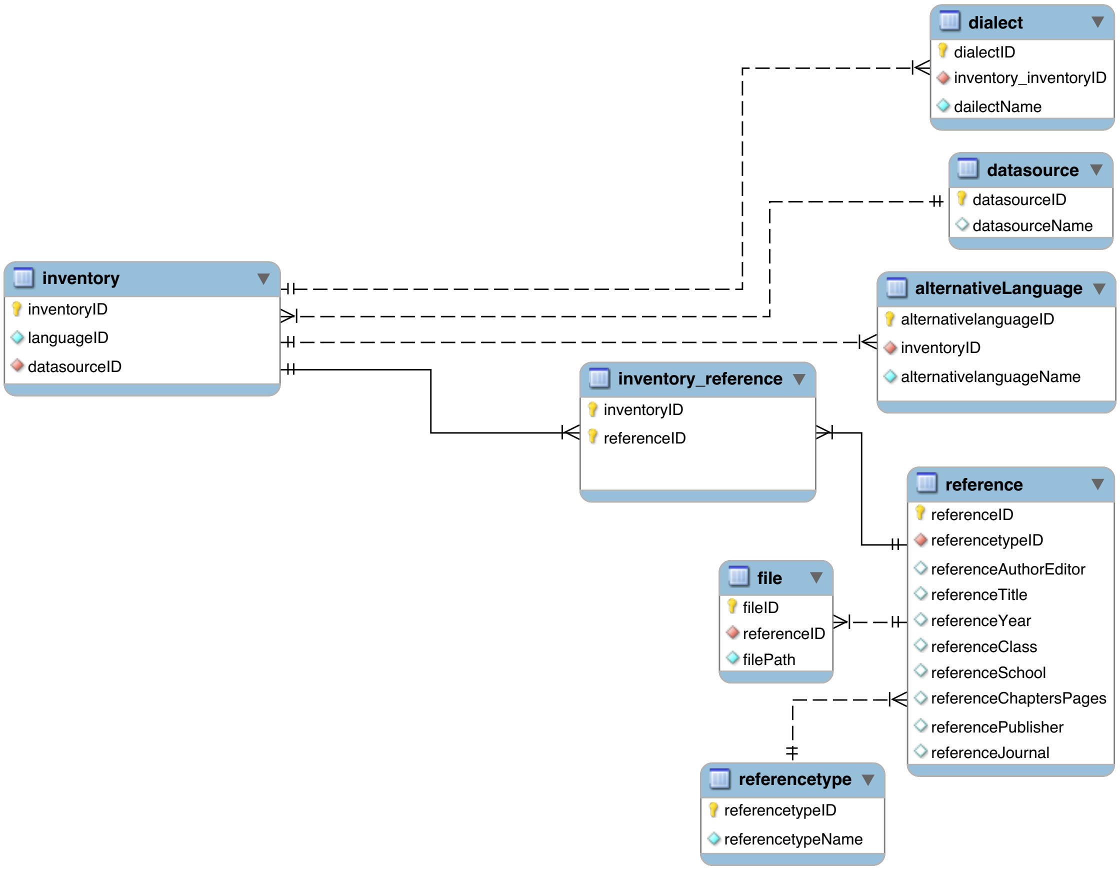
IPA character I also hand-coded whether it represented a consonant, vowel, tone or diacritic in the unicodeIPAClass field. This allows users to search on a particular segment class. For example, a user can retrieve all languages that contain a tonal segment (see Section 3.2.2).

Stepping back to the phoneme table in the relational database schema in Figure 3.10 on page 3.10, the relationship between the phoneme and glyph tables is modeled as a non-identifying relation. In an abstract sense, a glyph can exist without a phoneme (and vice versa). For example, there are many glyphs in undeciphered writing systems that exist, even though we do not know (and may never know) what phoneme, syllable, logogram or other thing that they represent. This is in contrast to the relationship between the phoneme and grapheme tables, which I have modeled as an identifying relationship. A grapheme requires both an auditory and a graphical representation to exist. Likewise, in my model an allophone also requires both a phoneme and a graphical representation. This may seem a bit backwards, since it is the phoneme that is derived from one or more allophones through linguistic analysis. However, many linguistic descriptions only list a language's contrastive segments (UPSID$_{451}$ is an example of a resource that only lists phonemes). Therefore, the phoneme table must be modeled in relation to the glyph table without an intermediate allophone table. The relationship between phoneme and glyph could be normalized even further. However, we would lose the notion that phonemes are language-dependent, i.e. each occurrence of a phoneme is distinct in each language. For example, if one queries the number of unique phonemes via phonemeID in the phoneme table, there are over 50,000 distinct phonemes across more than 1000 inventories.

Next I turn to the inventory table and the bibliographic and other metadata tables that link from its right side in the schema. The relevant portion of the database schema is reproduced in Figure 3.12. I will first focus on the relation between the inventory and reference tables in the lower half of the schema.

Relationships across tables can combine to create pivot tables. For example, there is a one-to-many-to-one relationship between the inventory table and the reference table through the inventory_reference pivot table as shown in the EER diagram. The inventory table holds information regarding inventories, including an inventory ID, lan-

Figure 3.12: Inventories and bibliographic reference data



guage ID and a data source ID. The `reference` table contains information regarding bibliographic citations (in “BibTeX” format, hence the fields in the table pertinent to BibTeX entry types).[30] I modeled this one-to-many-to-one relationship between inventories and references through a pivot table because one inventory can have one reference, one reference can be associated with many inventories, or one inventory can have multiple references.

Currently, the most typical situation is that there is one segment inventory for a given language referenced by one publication. For example, there is one record for Tanacross

[30] BibTeX fields can be easily expanded to include OLAC metadata extensions, such as olac:code for ISO 639-3 language name identifiers or WALS language codes that citations in WALS use.

[tcb], which is extracted from *The Phonology and Morphology of the Tanacross Athabaskan Language* (Holton, 2000a). At some point in the future, however, I could add another segment inventory from a different phonemic analysis of Tanacross, such as Leer 1982. Thus there would be two different inventories of the same language, but each would be referenced with its own `inventoryID` and each associated with the publication from which its segment inventory description was extracted.[31]

One reference publication may also be associated with many inventories (each inventory has a unique `inventoryID`). This association is captured by the crow's foot relation between the `reference` and `inventory` tables. Currently, the most extreme cases in PHOIBLE are the inventories that document the dialects of Kigiryama (aka Mijikenda [nyf]; Kenya; Bantoid) in Volk 2011 and Sebat Bet Gurage [sgw] (Ethiopia; Semitic) in Hetzron 1977.[32,33]

Lastly, one inventory can have multiple references and this ties in with the challenges of documenting data provenance described in Section 2.3.6. A simple example is that a researcher undertaking an analysis with inventories in PHOIBLE should cite both the data source and the relevant bibliographic citations from particular segment inventories. A more complex example is illustrated by the four inventories for Akan [aka] in PHOIBLE.

[31] See short discussion in Section 2.3.4 with regard to the difference in the descriptions of the segment inventories of Tanacross.

[32] Information regarding the dialect described in a given publication is recorded, when available, for each inventory in the `dialect` table, described below.

[33] An interesting note here is that on the one hand, each of the six dialects of Kigiryama (Giryama, Jiβana, Kambe, Kauma, Raβai and Reβe) contains the same set of segments. So any inventory is representative of the language. On the other hand, the six dialects of Sebat Bet Gurage (Chaha, Ezha, Gumer, Gura, Gyeto and Muher) range in total number of phonemes from 39 to 45 and each inventory consists of a slightly different set of segments.

Table 3.12: Comparison of Akan inventories

ID	Source	Phonemes	Consonants	Vowels	Tones	Citations
140	SPA	40	22	15	3	Welmers 1946 Ladefoged 1964 Stewart 1967 Schachter and Fromkin 1968 Crothers et al. 1979
N/A	UPSID_{317}	34	21	13	0	Welmers 1946 Stewart 1967 Schachter and Fromkin 1968 Maddieson 1984
208	UPSID_{451}	35	21	14	0	Welmers 1946 Stewart 1967 Schachter and Fromkin 1968 Dolphyne 1988a Ladefoged 1964 Maddieson 1984 Maddieson and Precoda 1990
N/A	Hartell	26 (34)	17 (25)	9	0	Bambose 1982 Dolphyne 1971 Dolphyne 1988a Fromkin 1977 Warren ND Hartell 1993
655	Chanard	31	18	9	4	Hartell 1993 Chanard 2006
1244	PHOIBLE	60	28	30	2	Dolphyne 1988a Moran 2012

The SPA inventory for Akan contains a total of 40 phonemes, including three tones (high, mid and low) and two lengthened vowels (/øː/ and /œː/).[34] These two vowels are (very) marginal phonemes.[35] So all in total, one would reference the original sources from which the SPA compiler's took the Akan inventory (Welmers, 1946; Ladefoged, 1964; Stewart, 1967; Schachter and Fromkin, 1968) and SPA as a resource whose authors interpreted an Akan inventory from those primary sources (Crothers et al., 1979).[36]

The next inventory of Akan is from UPSID$_{317}$. Although UPSID$_{317}$ inventories are not included in PHOIBLE, I mention it here to point out the slight differences in analysis between inventories in Maddieson 1984 and Maddieson and Precoda 1990.[37] The Akan inventory in UPSID$_{317}$ contains 34 total phonemes, including 21 consonants and 13 vowels. The inventory is based on three of the four same references as SPA, including Welmers 1946; Stewart 1967; Schachter and Fromkin 1968, and as noted by Maddieson (1984), UPSID$_{317}$ benefitted from the work of SPA. The UPSID$_{317}$ description of the Akan inventory differs slightly from SPA. UPSID$_{317}$ does not include the phonemes /ç, dʝ, k^{wh}, r̩/, but does include and notes the marginal phonemes /øː, œː/. Maddieson (1984) also does not syllabify /m, n/ and describes /d, n, r, s, t^{h}/ as underspecified for dental and alveolar place of articulation.

For the expanded and corrected UPSID$_{451}$, Maddieson's sources include Welmers 1946; Stewart 1967; Schachter and Fromkin 1968; Dolphyne 1988a; Ladefoged 1964. The UPSID$_{451}$ inventory is very close to the UPSID$_{317}$ inventory, but it adds /ɛ, ɲ, ɲ^{w}, ɥ/ and removes /w/, the two marginal phonemes /øː, œː/, and the underspecified dental/alveolar consonants, marking them as alveolar.[38]

[34] See Appendix E for SPA to IPA segment correspondences.

[35] Crothers et al. (1979, 50) state: "Welmers [1946] reports two words with the vowels /o-trema-long, o-open-trema-long/ occurring before /r-trill/, and analyzes them as /u.e/ and /upsilon.epsilon/ respectively (pg 20)."

[36] Should inventories, whether taken directly from the original resource or reinterpreted from the original resource(s) by someone else, be reference differently? This is an open question for tracking data provenance. Currently I use the same citation style for both originals and reinterpretations, as shown in Table 3.12.

[37] Maddieson and Precoda (1990, 104) expanded and corrected a second version of UPSID to "improve the accuracy of the data".

[38] A note in the UPSID$_{451}$ database under the Akan entry states, "Labialized palatals appear as labialized velars before back vowels. Velar stops and palatal affricates are largely complementary in distribution but

Next there is the Akan inventory compiled for *Alphabets of Africa.* Hartell (1993, 168) notes “Data taken from the bibliography and verified by Florence Dolphyne.” The data come from Bambose 1982; Dolphyne 1971, 1988a; Fromkin 1977; Warren ND. *Alphabets of Africa* was later digitized and put online by Chanard (2006). Chanard’s digitization does not include the noted palatalization and labial-palatalization segments /dw, hw, tw, gy, hy, ky/. Chanard also reinterprets the segments /ι, υ, nw, y/ as /ɪ, ʊ, n^{w}, j/, respectively. And four tones are added (high, mid, low and falling), although where the additional data come from is unclear because it is not stated.

Lastly, the Akan inventory added to PHOIBLE was extracted from Dolphyne 1988a, which provides details for each sound, including the use of labiopalatalized affricates and fricatives. Dolphyne (1988a) lists 60 phonemic segments, of which 28 are consonants, 30 are vowels and two are tones. This analysis contains many vowels because the description of Akan’s 10 vowel system is triplicated; each vowel has a lengthened and a nasalized phonemic counterpart.

Taken all together, the four inventories of Akan in PHOIBLE (and all six Akan inventories in general) are based on nine works with additional re-interpretations by Crothers et al. (1979); Maddieson (1984); Maddieson and Precoda (1990); Hartell (1993); Chanard (2006) and myself. The Akan inventories provide a nice example of why it is important to track data provenance and they illustrate the difficulty in doing so.[39] Currently I keep an entry for each segment inventory referenced in PHOIBLE. However, what is needed is a mechanism to track the history of changes of a reference to a particular inventory ID. One option that I am exploring is to use Slowly Changing Dimensions (SCDs) (Kimball and Ross, 2002). SCDs are data management methodologies used to preserve and track changes to a database over time. The current PHOIBLE database is simply a snapshot of its current content, but what would be very useful is for all reference data fields to be updated so that historical records and changes can be kept track of.[40]

show contrast before /a/.”

[39] It should also be noted that Akan is actually a macrolanguage term for two main subdivisions that have been designed as Fanti [fat] and Twi [twi] by ISO 639-3.

[40] Data provenance also has to do with tracking the reasoning of why certain decisions regarding changes

The reference table contains bibliographic citation information for each segment inventory in PHOIBLE. Different content of segment inventories may cause users to question whether the digitizations are true to their original sources of if errors were introduced in their digitization.[41] Therefore, the reference table is also in a one-to-many relationship with the file table, which is used to store the "phonological squibs" that I collected for inventories represented in PHOIBLE. A phonological squib is a PDF scan of the pertinent pages from which data from the phonological system was interpreted and extracted. Phonological squibs give users easy access to a fair use snippet of the original data source if they wish to consult it or if they think that they have found a mistake in my interpretation or processing of the data.

The reference table is also in a many-to-one relationship with the referenceType table. The referenceType table simply keep tracks of the BibTeX entry type for each reference record. This is a nice example of database normalization – instead of the reference table containing an additional column that records the BibTeX entry type for each record in the reference table (which would contain many duplicate BibTeX reference types, e.g. "book", "article", "phdthesis", etc.), there is a referenceType table that contains only the unique BibTeX entry types, each of which is mapped to one more referenceID records. The referenceType table provides information on how many of each publication type are represented in PHOIBLE, e.g. *n* number of references are from PhD theses.

There is one last point to consider about the relationship between an inventory and its reference. The inventory_reference table cannot exist with the inventory and reference tables, hence the use of solid lines to represent an identifying relation between the two. The consequence is that the foreign keys from the outside tables (inventory and reference) together form a composite primary key in the inventory_reference table. This means that the same combination of inventory.inventoryID and reference.-referenceID can only happen once. Thus the identifying relationship shows that a row

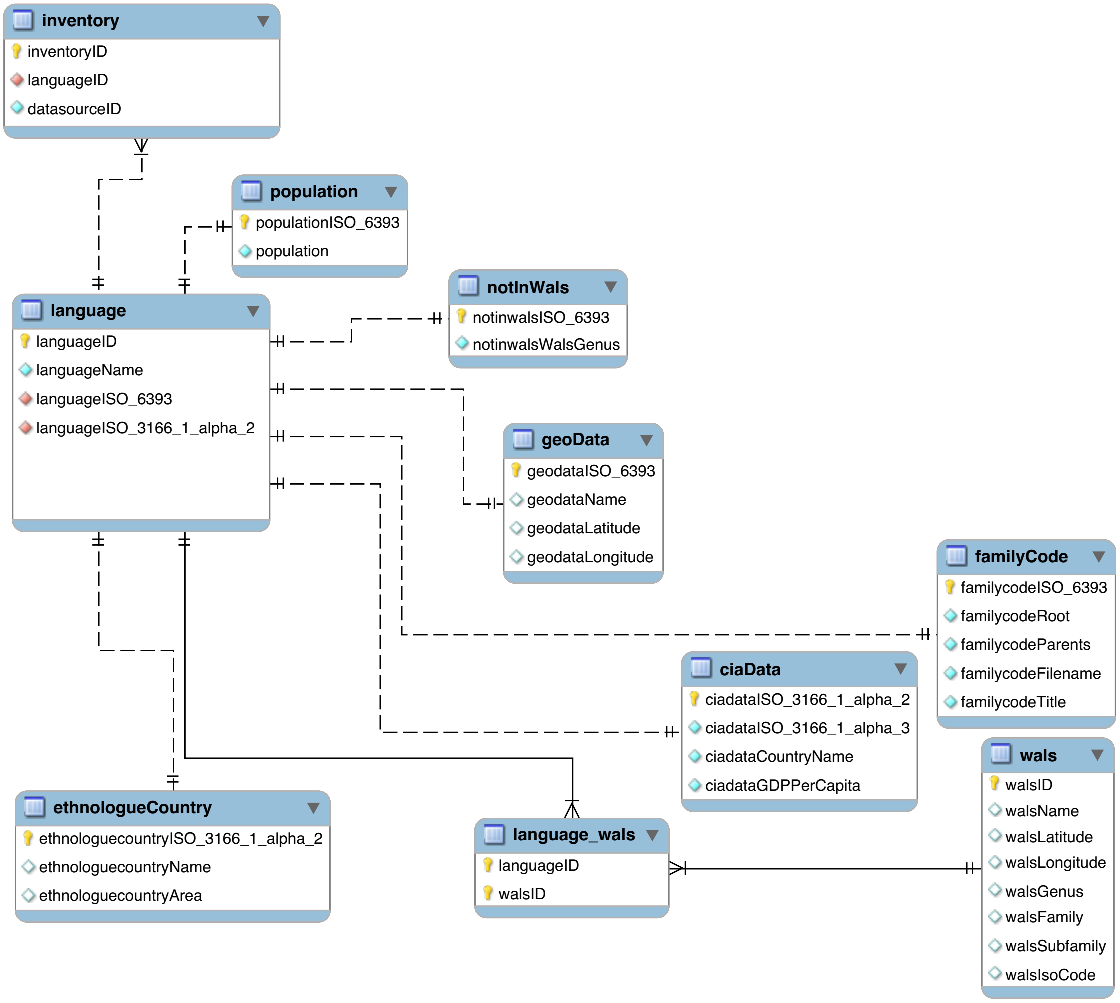
were made.

[41] See Section 2.3.

in inventory_reference cannot exist without either inventory or reference.

Moving on, the PHOIBLE database contains more than just segment inventory data and their bibliographic references. Each segment inventory has also been augmented with additional linguistic and non-linguistic data in tables that can be joined to segment inventory data via ISO 639-3 codes through the language table. From the lower half of the PHOIBLE database schema, these tables and their relations are shown in Figure 3.13.

Figure 3.13: Inventories and additional data



The first thing to note is that the inventory table is in a many-to-one relation with the language table. Each segment inventory in PHOIBLE is associated with an ID, a language name, its ISO 639-3 language name identifier and an ISO 3166-1 alpha-2 country code.[42] Using the language table and the ISO codes, I am able to link in language-specific data, including: population figures, geographic data, and genealogical information. I will now discuss each data source in turn.

Regarding population figures, these estimates are taken from the Ethnologue 16 (Lewis, 2009). I wrote a simple script to scrape the population estimates from each webpage. These data were then written to a tab-delimited file and imported into the population table, which is linked to the language table in a one-to-one relationship on ISO 639-3 codes. The extracted figures from the Ethnologue are numeric, ranging from 1 to 840,000,000, and there are several written descriptions, including: "No known speakers", "No estimate available", "Extinct" and "Ancient". These numbers and text descriptions are retained in the population table. I have also included the label "Missing E16 page" in 66 occurrences. Ethnologue 16 was published in 2009 and since then there has been several updates to ISO 639-3. These changes will be reflected in the next edition of the Ethnologue. For example, the code for [apf] for the language Pahanan Agta was added in change request 2009-086.[43] The request was adopted to split Paranan [agp] into Pahanan Agta [apf] and Paranan [prf]. A problem of course is that the current population table lists 16,700 speakers for Paranan and does not yet reflect its split into two distinct languages with two populations. When the next edition of the Ethnologue is released, its contents will have to be re-scraped for new population figures and the PHOIBLE database updated to include the new figures, while retaining the old ones.

Population figures present some other interesting challenges. In general obtaining "correct" figures is problematic because different sources may diverge by 100% or more in their reporting of population sizes (Bauer, 2007). In the Ethnologue, population figures can differ drastically depending on the countries where the language is spoken, e.g. Nzema [nzi] (Ghana) lists "262,000 in Ghana (2004 SIL). Population total all countries: 328,700". It is

[42] http://www.iso.org/iso/country_codes.htm

[43] http://www.sil.org/iso639-3/chg_detail.asp?id=2009-086

not always clear where the population figures come from. For example, for Kwambi [kwm] it simply lists “32,700 (2006)”. Perhaps that figure is from SIL personnel on the ground? In other places a citation is given, but the reference is not provided nor could I find it anywhere listed on the Ethnologue or SIL sites, e.g. the page on Opuuo [lgn] states “1,000 in Ethiopia (2007 A. Tsadik)” and the page on Narim [loh] (Sudan) lists “3,620 (1983 Fukui)”. There are also cases where it is unclear which population figure to use. For example, the page on Alaba-K’abeena [alw] (Ethiopia) states: “Population 162,000 (1994 census). 111,077 monolinguals (1994 census). 126,257 Alaba, 35,783 K’abeena. Ethnic population: 125,900 (1998 census).” And the page on Oshiwambo [kua] (Angola) lists: “421,000 in Angola (Johnstone 1993). Population total all countries: 668,000.” Some pages are simply difficult to parse: Otuho [lot] (Sudan) “135,000 (Voegelin and Voegelin 1977). Dongotono (1998), 2,500 Koriot, 1,000 Lomya.”; Saamia [lsm] (Uganda) “335,000 in Uganda (2002 census). 279,972 Basaamia and 75,257 Bagwe (2002 census).”; for Liberian English [lir] (Liberia) there is “No estimate available”, but under the “Language use” notes provided, Liberian English is noted as a trade language with 1,500,000 L2 speakers (1984 census). Population figures may be quite old, e.g. Kunyi [njx] has 52,000 speakers as of the 1984 census. Also problematic is the same language code used by different dialects, e.g. Kigiryama [nyf] (Kenya) lists “623,000 (1994 I. Larsen), increasing. 496,000 Giryama [nyf], 17,000 Kauma [nyf], 19,000 Jibana [nyf], 13,000 Kambe [nyf], 72,000 Rabai [nyf], 6,000 Ribe [nyf]”.

Moving past population and on to geographic data, several tables are involved and each links to the `language` table by either an ISO 639-3 or ISO 3166-1 alpha-2 code. The latter set of codes are used for the representation of names of countries. For example, the Ethnologue uses the two letter ISO 3166-1 alpha-2 codes to link between its `LanguageCodes`, `LanguageIndex` and `CountryCodes` tables.[44] Thus each `LangID` (ISO 639-3 code) and its corresponding canonical language name, alternative language names, language status, etc., is linked via a `CountryID` (ISO 3166-1 alpha-2 code) to the `CountryCodes` table, which contains both the country name and world area in which each language is

[44] `http://www.ethnologue.com/codes/default.asp`

spoken.[45] In the PHOIBLE relational database schema, the relation between the language and ethnologueCountry tables is one-to-one and made through the two-letter ISO 3166-1 alpha-2 codes, which act as a foreign key in language and the primary key in the ethnologueCountry. Thus for each segment inventory in PHOIBLE, the country and world region where it is spoken, as reported in Ethnologue 16, can be retrieved.

In the same way, I also link the language and ciaData tables in a one-to-one relation on ISO 3166 alpha-2 codes. In the ciaData table, I currently include only the CIA country name, the ISO 3166-1 alpha-3 code and the GDP per capita. Additional data is also available in the CIA World Factbook and could be linked to segment inventories, such as climate data, religion, median age and age structure, geographic coordinates, etc.[46]

Geographic coordinates for the majority of segment inventories are available in the geoData and wals tables.[47] Rows in the geoData data are in a one-to-one relation with the language, linked via ISO 639-3 codes. Data in the geoData table come from a database in the Department of Linguistics at the Max Planck Institute for Evolutionary Anthropology.[48] The data include Ethnologue language names, ISO 639-3 codes, and latitude and longitude figures for 6862 distinct ISO 639-3 codes. The latitude and longitude figures are separate, although comparable, to those in the wals table. The geo-coordinates in WALS were fine-tuned by Matthew Dryer and contain 2429 distinct ISO 639-3 codes. I have added the latitude and longitude figures from geoData to PHOIBLE because of their broad scope. The figures from WALS are also available because they are published with the WALS data set, discussed below.

Lastly, the genealogical information linked to each segment inventory comes from two sources: WALS (Haspelmath et al., 2008) and Ethnologue 15 (Gordon, 2005). WALS publishes its data online in downloadable delimiter-separated formats.[49] I imported the

[45] World areas include: Africa, America, Asia, Europe and the Pacific.

[46] https://www.cia.gov/library/publications/download/

[47] Currently 25 segment inventories in PHOIBLE have an ISO 639-3 code that does not have corresponding latitude and longitude figures.

[48] The aim of the database was to collect geographical coordinates for all ISO 639-3 codes. Roughly 10% still need verification and many locations were added by hand (Hans-Jörg Bibiko, p.c.).

[49] http://wals.info/export

WALS data into the wals table. Each row in this table contains a WALS language code, WALS language name, latitude and longitude coordinates, WALS genus, language family and subfamily, and an ISO 639-3 code. A pivot table is needed between the language and wals tables because there is a one-to-many relationship between WALS language codes and ISO 639-3 codes. For example, the language Angami (Sino-Tibetan; India) has the WALS code "agm", which corresponds to two distinct ISO 639-3 codes [nri] (Ethnologue language: Naga, Chokri) and [njm] (Naga, Angami). Thus the language_wals pivot table maps in a one-to-many-to-one relation a language.languageID and wals.languageID. This allows access from a segment inventory in PHOIBLE to its corresponding WALS data. Again, there are languages in PHOIBLE that are not in the WALS sample. For each of these additional 352 languages, I identified an existing WALS genus and mapped it to its ISO 639-3 code in the notInWals table. All languages represented in PHOIBLE are associated with a WALS genus, whether or not that particular language is included in the WALS sample.

The Ethnologue 15 language family data were taken from the Multitree project (LINGUIST List, 2009).[50] These data reside in the familyCode table, which is linked to the language table in a one-to-one relation on ISO 639-3 codes. Each ISO 639-3 code in the familyCode table is basically a leaf node in a particular language family. Thus each ISO 639-3 code is associated with a top-level language family stock (familycodeRoot), an immediate parent language family (familycodeParents), the filename of where the data was taken from (familycodeFilename) and a title representing that language family (familycodeTitle). For example, Standard German [deu] is an immediate child of the East Middle German [emge] branch of the language family (stock) Indo-European [ieur]. Note it may also be the case that the immediate parent of a language is also its language family stock, e.g. Quileute [qui] belongs to the Chimakuan family [chmn], which has no other known branches.

Assigning language families to ISO 639-3 codes posed a few problems. First, creoles and mixed languages are classified in their own "language families" in Multitree. The issue here

[50] Details are given in Section 4.4.

is that a creole may be assigned to more than one family, e.g. Jamaican Creole [jam] is assigned to both North American Pidgins and Creoles [napc] and Central American Pidgins and Creoles [capc] because it is spoken in both geographic regions. Another example are mixed languages, which form from bilingual situations, so in a sense the resulting language belongs to both (and neither) language family. I consider the cases of pidgins and creoles and mixed languages genealogically unclassifiable; each is assigned a language family stock code in the `familyCode` table, but the assignment is somewhat arbitrary (e.g. I assigned [jam] to [capc] because Jamaica seemed more Central America than North America to me, but technically Jamaica belongs to North America and so does Central America for that matter). This solution, or perhaps better put, this lack of a solution, is not entirely to my liking. However, the data warehouse procedure described in Section 3.2.2 requires that there be no ambiguity in language family assignment, i.e. either a family code is assigned in the `familyCode` for each ISO 639-3 code, which is in a one-to-one relation with the `language` table, or assignment is built into the procedure. I chose to use the former because it is more transparent.

To summarize, in this section I have discussed in detail the structure of the PHOIBLE relational database model and how the different data sources in PHOIBLE are connected. My design is practical for adding new segment inventories, for checking to see if their contents adhere to Unicode IPA, and for updating independent data sources that provide PHOIBLE's segment inventories with additional linguistic and non-linguistic information. However, the database schema as it is requires non-trivial prerequisite knowledge of relational design models and structured query languages, so that a user can query its contents. In the next section I give some examples of how one would query the PHOIBLE relational database model as its described in this section. Then I describe a data warehouse procedure I created that denormalizes the relational database data and outputs it into two flat files, which can be easily queried and the flat files' format is practical for end users that wish to do quantitative analysis with the data.

3.2.2 Data warehouse flat file tables

To illustrate the utility of denormalizing relational data into flat file tables, I will begin by showing some examples of how to query the PHOIBLE relational database. A few simple but interesting queries are:

- How many phonemes are there in a particular language?
- What are the set of phonemes in a particular language?
- How many languages contain a particular phoneme?
- Which languages contain a particular phoneme?
- What are the number of consonants, vowels and tones in a particular language?

As I discussed in the previous section and illustrated in Figure 3.10 on page 102, the relations between an inventory, its segments and how those segments are encoded, capture both segment types and segment tokens. This distinction encapsulates both the notion of language-specific sounds and languages containing the "same" sound across languages. The query in Example 3.7 returns the count of phonemes for the segment inventory associated with `inventoryID` 1. In the current database, this returns 40 phonemes for Korean, as reported for the segment inventory given in SPA (Cho, 1967; Martin, 1951, 1954; Martin and Lee, 1969; Kim, 1968, 1972; Crothers et al., 1979).

(3.7)
```
SELECT language.languageName, count(phonemeID)
FROM language, inventory, phoneme
WHERE inventory.inventoryID = phoneme.inventoryID
AND inventory.inventoryID = 1
```

Replacing "count(phonemeID)" with just "phonemeID" would return a list of rows containing the language ID and phoneme ID. Alas, this query only returns phoneme IDs and not the graphical representations of phonemes that linguists are used to working with. According to my relational model, a phoneme is made up of one or more component glyphs, which are themselves Unicode code points that can be rendered as glyphs, i.e. a particular representation of a grapheme via the font in which it is rendered.[51] The segment type-token distinction and the relationship between a segment and its component Unicode characters are encoded in the relations between `phoneme.phonemeID`, `glyph.glyphID`, `glyph_unicode.glyphID` and `glyph_unicode.unicodeID` and `unicodeIPA.unicodeIPAID`. Thus to return the graphical representations of segments, the query must incorporate aspects of the relational model design. This can be considered a disadvantage of using the relational database model because queries can quickly become quite complex, requiring clauses that combine fields from different tables. Example 3.8 shows a query that returns concatenated glyphs that represent phonemes for a segment inventory indicated by its inventory ID (removing the WHERE clause returns all inventories).

(3.8)
```
SELECT inventory.inventoryID,
GROUP_CONCAT(unicodeIPA.unicodeIPAGlyph
ORDER BY glyph_unicode.order ASC SEPARATOR '')
FROM phoneme
  INNER JOIN glyph_unicode ON
    phoneme.glyphID = glyph_unicode.glyphID
  INNER JOIN unicodeIPA ON
    glyph_unicode.unicodeID = unicodeIPA.unicodeIPAID
  INNER JOIN inventory ON
    phoneme.inventoryID = inventory.inventoryID
WHERE phoneme.inventoryID = 1
GROUP BY phoneme.phonemeID
```

The JOIN clauses are necessary to combine records in the relevant tables. Through database normalization, the redundancy of data in these tables has been minimized. The Unicode IPA description table (named `unicodeIPA` in the PHOIBLE relational database schema) is an example of applied normalization. My working format of that table is given in Appendix D. The table contains 177 unique rows. A snippet with database column names is given in Table 3.13.

[51] For terminology definitions, see Section 2.1.4.

Table 3.13: Snippet of Unicode IPA table

unicodeIPAID	unicodeIPAGlyph	unicodeIPAHex	unicodeIPAClass
116	t	0074	consonant
688	ʰ	02B0	consonant
690	ʲ	02B2	diacritic
810	_	032A	diacritic
643	ʃ	0283	consonant

First there are few things to note in the table. Since each character in Unicode is given a unique code point, we can use a representation of that code point for the primary key of the `unicodeIPA` table. I use the decimal representation of Unicode code points, shown in the `unicodeIPAID` field.[52] In the `unicodeIPAGlyph` cells, a graphical representation of each Unicode character is given. And in the `unicodeIPAClass` cells, there is a segment class label, denoting to what class a character belongs (consonant, vowel, tone or diacritic). By ordering complex segments, i.e. segments that are made up of one or more characters and/or diacritics, I can use the `unicodeIPAClass` label of the first Unicode character to determine the class of each segment. This compositional approach, of course, is not perfect. Note that the aspiration diacritic <ʰ> in row two is labeled "consonant". This is a bit of a hack due to the fact that the aspiration diacritic precedes its base consonant in pre-aspirated stops, e.g. pre-aspirated stops in Hopi [hop].

The relationships between the `phoneme` and `unicodeIPA` tables have been normalized to reduce redundancy. An example is that the rows in the `unicodeIPA` table are unique. Thus each segment in each segment inventory is actually modeled as a (possible) combination of segments from the `unicodeIPA` table. For example, the complex segment /t̪ʃʲʰ/, a

[52] The `unicodeIPAHex` field also contains unique code points, represented in hexadecimal, which could be used as unique identifiers.

palatalized voiceless aspirated palato-alveolar sibilant affricate found in the inventories of Kashmiri [kas] and Amuesha [ame] in SPA[53] and UPSID$_{451}$,[54] is made up of five characters: < t > + <_> + < ʃ > + < ʲ > + < ʰ >. Instead of storing the graphical representation of <t̠ʃʲʰ> twice in the database (or more if it is encountered in another language description), a phoneme ID is assigned to this segment type and it consists of a list of glyphs that are each associated with a unique Unicode ID.

One might ask, why not just store each segment separately as its graphical representation and disregard duplication? Often a trade-off for simplicity in one area will cause another area to become more complex. Thus there are inevitably conflicts of design that occur in relational database modeling. One reason that I chose to break segments down into their component glyphs and Unicode points is because creating a relatively short list of the unique Unicode IPA characters is far more efficient than going through the list of 1780 distinct segment types that currently exist in PHOIBLE and assigning each of them additional information such as a segment type label or a vector of distinctive feature values. This information can be generated compositionally from the `unicodeIPA` table and from additional information about which features belong to which segments. For example, an interesting query is to get the consonant, vowel and tone counts for each segment inventory. This allows a user to examine phenomena like consonant and vowel ratios across languages. After assembling a complex segment, I can identify its segment class by looking up the segment class for its first character in the `unicodeIPA` table. Once each segment in an inventory has been assigned a segment class label, those consonants, vowels and tones can be summed up. Another example has to do with feature vector assignment. By including features for each composite character in `unicodeIPA`, features can be assigned to contour and complex segments iteratively.[55] The Unicode IPA table also provides the additional benefit of acting as an error checker for segments' characters that are inserted into the database. If a non-standard Unicode IPA character was mistakenly entered into the data

[53] Kelkar and Trisal 1964; Fast 1953; Crothers et al. 1979

[54] Fast 1953; Wise 1958; Kelkar and Trisal 1964; Zakhar'in and Edelman 1971; Zakhar'in 1974; Bhat 1987; Maddieson and Precoda 1990

[55] See Chapter 6.

somewhere in the pipeline, it can be easily caught and corrected. Lastly, the segment type labels in the `unicodeIPA` table can be used to generate the composition of a segment, which provides an additional method for searching on segments and segment types. For example, if a user wants to query languages for triphthongs, they can search on segments that match “vowel-vowel-vowel”. This functionality is discussed below.

In addition to extracting information about segments in inventories, PHOIBLE provides users with additional data like genealogical group, geographic location, etc. Again, accessing the data via the relational database is an involved task because the model, although with its advantages for combining and keeping data updated, puts the burden of extracting the desired information into the query. The verbose query in Example 3.9 shows how one might extract segment inventories’ contents along with their inventory ID, ISO 639-3 language name identifier, language name, population, and geographic and genealogical information about the language. A snippet of this query’s result is given in Table 3.14.

(3.9)
```
SELECT inventory.inventoryID as ID,
  language.languageISO_6393 as ISO_6393,
  language.languageName as name,
  population.population as population,
  ethnologueCountry.ethnologuecountryName as country,
  ethnologueCountry.ethnologuecountryArea as area,
  geoData.geodataLatitude as latitude,
  geoData.geodataLongitude as longitude,
  familyCode.familycodeRoot as stock,
  wals.walsGenus as genus,
  GROUP_CONCAT(unicodeIPA.unicodeIPAGlyph
  ORDER BY glyph_unicode.order ASC SEPARATOR '') as glyph
FROM phoneme
  INNER JOIN glyph_unicode ON
    phoneme.glyphID = glyph_unicode.glyphID
  INNER JOIN unicodeIPA ON
    glyph_unicode.unicodeID = unicodeIPA.unicodeIPAID
  INNER JOIN inventory ON
```

```
    phoneme.inventoryID = inventory.inventoryID
  INNER JOIN language ON
    phoneme.inventoryID = language.languageID
  INNER JOIN population ON
    language.languageISO_6393 = populationISO_6393
  INNER JOIN geoData ON
    language.languageISO_6393 = geodataISO_6393
  INNER JOIN ethnologueCountry ON
    languageISO_3166_1_alpha_2 =
    ethnologuecountryISO_3166_1_alpha_2
  INNER JOIN familyCode ON
    language.languageISO_6393 = familycodeISO_6393
  INNER JOIN language_wals ON
    language.languageID = language_wals.languageID
  INNER JOIN wals ON
    language_wals.walsID = wals.walsID
GROUP BY phoneme.phonemeID
```

Table 3.14: Results for query in Example 3.9

ID	ISO6393	name	population	country	area	latitude	longitude	stock	genus	glyph
1	kor	Korean	42,000,000	Korea, South	Asia	37:30	128:0	asis	Korean	t̠ʃʰ
1	kor	Korean	42,000,000	Korea, South	Asia	37:30	128:0	asis	Korean	ʔ
1	kor	Korean	42,000,000	Korea, South	Asia	37:30	128:0	asis	Korean	n
1	kor	Korean	42,000,000	Korea, South	Asia	37:30	128:0	asis	Korean	o
2	ket	Ket	190	Russia	Europe	64:0	87:0	yeos	Yeniseian	ʔ
2	ket	Ket	190	Russia	Europe	64:0	87:0	yeos	Yeniseian	n
2	ket	Ket	190	Russia	Europe	64:0	87:0	yeos	Yeniseian	nʲ

On the one hand, a disadvantage of my relational database model is the complexity involved in querying it for data. On the other hand, an advantage of a relational database is the ability to extract data in structured formats that can be consumed by users and used as input and read into other programs. A flat file table like that given in Table 3.14 may be redundant in certain respects, but it is easily loaded into a program like R for statistical analysis. To create flat file tables from a relational database, its tables must be joined in various ways and the data extracted into the desired formats. This process can be undertaken with a SQL script, essentially a large SQL query, that denormalizes and extracts relational data into a single flat file database table. I call the output of this process a data warehouse.

In standard business practice, there is typically a division between a live “operational” database and a data warehouse. The operational database is built to handle transactions and is designed with rules of database normalization to optimize performance and data integrity. The notion of a data warehouse emerged in the late 1980s through work at IBM to meet the growing demand from businesses to undertake data mining and analysis of transactional database data. The term data warehouse was made popular by Inmon (1992), who’s emphasis was on integrating data into a collection that would aid business management in decision making. Thus data warehousing became the process that organizations use to integrate data from different sources to facilitate data mining, analysis, reporting and decision making.

There are several definitions for *data warehouse* because the integration process and forms in which data are stored differ from project to project. The data warehouse is designed for query and analysis rather than for transactional processing, so the models in which the data are stored and the types of formats from which data are extracted differ significantly. For example, the focus of a data warehouse is often to mine consumer activities to identify consumer trends. A data warehouse can be a flat file or another type of database, such as a relational database, object database, etc. A data warehouse can be normalized or denormalized.

My approach to data warehousing in this work follows from Kimball 1996, in which a data warehouse is defined as a copy of transaction data that is structured for query and analysis. Currently, I create two data warehouses that are generated from a SQL script and result in two flat file tables containing data from PHOIBLE's relational database with some additional information generated by the script, such as a trump ordering and the compositional make-up of segments, which I discuss below. Tables 3.15 and 3.16 illustrate the phoneme level and aggregated data warehouses.[56]

[56] The column headers are easily changed and are listed here as-is for convenience sake.

Table 3.15: PHOIBLE data set at the phoneme level

Source	ID	ISO6393	trump	root	wals_genus	population	latitude	longitude	phoneme_id	glyph_id	glyph	class	comb	n
SPA	1	kor	1	asis	Korean	42,000,000	37:30	128:0	1	1	t̺ʃʰ	cons	c-d-c-c	4
SPA	3	lbe	1	ncau	Lak-Dargwa	157,000	42:0	47:0	124	1	t̺ʃʰ	cons	c-d-c-c	4
SPA	5	kat	1	kart	Kartvelian	3,900,000	42:0	44:0	203	1	t̺ʃʰ	cons	c-d-c-c	4
SPA	6	bsk	1	asis	Burushaski	87,000	36:30	74:30	240	1	t̺ʃʰ	cons	c-d-c-c	4
SPA	14	khm	1	ausa	Khmer	12,300,000	12:30	105:0	632	19	uː	vowel	v-d	2
SPA	27	tha	1	taik	Kam-Tai	20,200,000	15:00	100:40	1150	19	uː	vowel	v-d	2

Table 3.16: PHOIBLE data set aggregated

Source	ID	ISO6393	trump	root	wals_genus	population	latitude	longitude	phonemes	cons	tones	vowels
Chanard	649	abi	1	ncon	Kwa	50,500	5:40	-04:35	39	21	0	18
PHOIBLE	922	lbj	1	sitb	Bodic	150,000	34:00	78:00	38	33	0	5
SPA	91	zun	1	nais	Zuni	9,650	34:55	109:0	54	44	0	10
UPSID	648	zun	2	nais	Zuni	9,650	34:55	109:0	25	20	0	5
PHOIBLE	1013	skv	1	skoo	Western Skou	700	-02:35	140:55	23	13	3	7

The data warehouse tables are not just a simple database export. I wrote a SQL stored procedure that combines the different relational tables and reverse engineers the normalization forms to denormalize the database's contents into flat files. Denormalized data sets are easy to manipulate and query, as I will show below. The data are extracted from the relational database and I apply some analysis to add a source trump ordering, information about the segment class composition of each segment and counts for total phonemes, consonants, vowels and tones. The data warehouse tables contain a copy of the relational data in PHOIBLE at a particular moment. Thus as more segment inventories are added or data in the relational database is updated (e.g. language codes), up-to-date data warehouses tables can be created with the new data by simply recalling the SQL stored procedure.

A SQL stored procedure is basically a series of SQL queries saved in a database so that it can be called at any time like a command or function. To generate the data warehouse tables, my SQL procedure first creates a source trump ordering table that assigns an order to be applied to duplicate segment inventories, i.e. inventories that contain the same ISO 639-3 language code. The current source trump ordering is set to select inventories first from source PHOIBLE, then SPA, then UPSID$_{451}$ and finally Chanard (AA). When there are duplicate inventories within the same source, the trump hierarchy is applied by order of ascending PHOIBLE ID. Some examples are given in the Table in 3.17.

The data warehouse flat file tables are built up stage by stage by the SQL procedure. Similar to the SQL query shown in Example 3.9 on page 121, data from tables `population`, `ethnologueCountry`, `geoData`, etc., are joined on primary keys with `language` and `inventory` and the relevant data for each segment inventory is extracted and added to the data warehouses. Again, the SQL function GROUP_CONCAT grabs the glyph combinations and concatenates them into cells in the `glyph` column in the phoneme level table. While I do the concatenation, I create another table that keeps track of the unique glyphs and sums the number of combined characters (shown in the column `NumOfCombined`) and determines each segment type's class (in `CombinedClass`). So for example, in Korean the segment <t̠ʃʰ> is a consonant that has a length of four characters, which combine as: c-d-c-c. I also use this information to dynamically generate the figures in the aggregated data warehouse table. The number of phonemes, consonants, vowels and tones are determined

Table 3.17: Example of the source trump hierarchy

Source	ID	ISO6393	Language Name	SourceTrumpOrder
SPA	124	hau	Hausa	1
UPSID	351	hau	HAUSA	2
Chanard	729	hau	hausa (Niger)	3
Chanard	730	hau	hausa (Nigeria)	4
PHOIBLE	1244	aka	Akan	1
SPA	140	aka	Akan	2
UPSID	208	aka	AKAN	3
Chanard	655	aka	akan	4

during the SQL procedure and then summed up and added to the aggregated table.

The phoneme level and aggregated tables are flat file databases, i.e. a database that consists of a single table (and file) that stores data in a flat structure consisting of a set of columns and rows, and contains one record per row. Each record is separated by some type of delimiter when exported, e.g. I export the output of the data warehouse SQL stored procedure, the phoneme level and aggregated tables, into a tab-delimited format and then load those files into R or Python.

On the one hand, a disadvantage of the flat file database tables is that they would be very cumbersome to maintain. The denormalization of the data causes much duplication, so for example if a language code is changed, then the maintainer of the data set would have to replace all occurrences of that language code in these tables. A “find and replace all” command may be invoked to speed along such a change, but by having collapsed all the data into one table, the maintainer loses the flexibility of updating say only the `geoData` or `population` tables.

On the other hand, flat file databases are very practical to query with SQL because there are no relational tables to join. I will now give some examples of SQL queries on the data warehouse tables that I find useful. My SQL procedure stores both tables into a database that I called `ReportingWarehouse`. My two tables are called `Master_ResultSet_PhonemeLevel` and `Master_ResultSet_Aggregated`. To query all rows from the aggregated data warehouse table, which is illustrated in Table 3.16 on page 128, the user can use the query given in Example 3.10 with the appropriate table name.

(3.10)

```
SELECT *
  FROM ReportingWarehouse.Master_ResultSet_Aggregated
```

This query returns 1336 rows from the aggregated data warehouse table, which equates to one for each segment inventory in PHOIBLE. The query shows how many phonemes there are in each language description in PHOIBLE, including counts for consonants, vowels and tones. If the `SourceTrumpOrdering` field is restricted to “1”, as in Example 3.11, the query will return the set of 1089 distinct segment inventories in PHOIBLE as per the current trump hierarchy.

(3.11)

```
SELECT *
  FROM Master_ResultSet_Aggregated
  WHERE SourceTrumpOrdering = 1
```

If a user only wants to get at information from a particular source, say the UPSID$_{451}$ inventories and their segment counts including the number of consonants and vowels, he or she could use the query given in Example 3.12 to retrieve 451 segment inventories.

(3.12)

```
SELECT *
  FROM Master_ResultSet_Aggregated
  WHERE source = "UPSID"
```

The WHERE clause in SQL statements is used to restrict the results to specified criteria. For example, a user might only be interested in segment inventories from languages spoken in Africa, as shown in the query in Example 3.13.

(3.13)
```
SELECT *
  FROM Master_ResultSet_Aggregated
  WHERE area = "Africa"
```

More specific still, perhaps the user only wants access to only Afro-Asiatic languages spoken in Africa, shown in Example 3.14.

(3.14)
```
SELECT *
  FROM Master_ResultSet_Aggregated
  WHERE area = "Africa" and root = "afas"
```

The user can continue to further specify his or her criteria. For example, a query to return Afro-Asiatic languages spoken in Africa, the segment inventories of which include a description of tone, given in Example 3.15.

(3.15)
```
SELECT *
  FROM Master_ResultSet_Aggregated
  WHERE area = "Africa" and root = "afas" and TopLevel_tone > 0
```

Moving on to the phoneme level data warehouse table illustrated in Table 3.15 on page 127, perhaps the user wants to know what exactly those tone segments are in the descriptions of Afro-Asiatic languages spoken in Africa, as described in the query in Example 3.16.

(3.16)
```
SELECT *
  FROM Master_ResultSet_PhonemeLevel
  WHERE root = "afas" and area = "Africa" and class = "tone"
```

Users can also query on a particular segment. For example, someone investigating tone might want to know which languages described in PHOIBLE have a high tone, as shown in Example 3.17. This query shows how many languages contain a high tone by displaying those languages.

(3.17)
```
SELECT *
  FROM Master_ResultSet_PhonemeLevel
  WHERE glyph = "˦"
```

Taking advantage of the `CombinedClass` column in the aggregated data warehouse table, users can also search for languages with contour tones that contain two tones, as shown in Example 3.18.

(3.18)
```
SELECT *
FROM Master_ResultSet_PhonemeLevel
WHERE CombinedClass = "t-t"
```

So far, my examples have restricted criteria to specific occurrences, but SQL also offers the LIKE operator to search for a specified pattern within a column. For example, if a user wants to search for descriptions of languages that contain contour tones with two or more tones, they could use the query given in Example 3.19, which would also return records that contain segments such as /˥˩˥/.

(3.19)
```
SELECT *
FROM Master_ResultSet_PhonemeLevel
WHERE CombinedClass like "t-t%"
```

When querying for language descriptions that contain ranges of segments, the LIKE operator is particularly useful. For example, someone might wish to test claims about diphthongs made in Miret 1998. One might start with a query similar to 3.18, but with vowels specified, as in Example 3.20.

(3.20)
```
SELECT *
FROM Master_ResultSet_PhonemeLevel
WHERE CombinedClass = "v-v"
```

This query would capture specifically those records that contain some combination of vowel and vowel. However, it would not catch diphthongs containing a diacritic, such as nasalized or lengthened diphthongs. To capture those diphthongs as well, one would want to again use the LIKE operator with the “%” wildcard, as in Example 3.21. This query would also capture triphthongs.

(3.21)
```
SELECT *
FROM Master_ResultSet_PhonemeLevel
WHERE CombinedClass LIKE "%v%-%v%"
```

Lastly, it is quite simple to query the phoneme level data warehouse table to find out what the set of phonemes in a particular language is. This search can be undertaken with either the language name or more precisely with the ISO 639-3 language name identifier. In Section 2.3.5, I discussed issues regarding alternative language names. Personally I find it easier and quicker to identify a particular language's ISO 639-3 code via the Ethnologue's website and then I use the code to query PHOIBLE, as shown in Example 3.22. This query returns two inventories for Nama [naq], one given in SPA and the other in UPSID$_{451}$.

(3.22)
```
SELECT *
FROM Master_ResultSet_PhonemeLevel
WHERE language_code_id = "naq"
```

As I mentioned before, exporting the phoneme level and aggregated data warehouse tables in a delimited format is also useful as an input format to other tools like statistical packages and programming scripts. I will briefly show some queries that can be undertaken using these tab-delimited data warehouse flat files and R,[57] a free software environment for statistical analysis and for creating plots and graphics.[58]

The first step is to read the table into R, as shown in Example 3.23. The data are read from the tab delimited "1089_Master_ResultSet_Aggregated.tab" file into the variable "data.all". The file contains a header ("header=T(rue)"). The data should be split on tab as a separator (sep="\t"). Quotation marks should be escaped (quote="\""). And decimal points are marked with a period (dec=".").

(3.23)
```
data.all <- read.delim("1089_Master_ResultSet_Aggregated.tab",
   header=T, sep="\t", quote="\"", dec=".")
```

[57] http://www.r-project.org/

[58] The analyses and graphics presented in Chapters 5 and 7 are made with R.

Once the data has been read in, the user can take advantage of the simplicities of the R programming language to query the data. I will show just a few examples. First, an important feature of R is the ability to easily subset the data given a column and its value. Some examples are given in Example 3.24.

(3.24)
```
1. data.trump <- subset(data.all, SourceTrumpOrdering == 1)
2. upsid <- subset(data.all, Source == "UPSID")
3. phoible <- subset(data.all, Source == "PHOIBLE")
4. chanard <- subset(data.all, Source == "Chanard")
5. spa <- subset(data.all, Source == "SPA")
6. vowels <- data.all$TopLevel_vowel
```

Line 1 would subset the rows in the aggregated data table into those that have a source trump order of "1". This would gather the set of unique segment inventories into the "data.trump" variable. Lines 2, 3, 4 and 5 simply subset the data from "data.all" into subsets based on the source type, e.g. if one wants to access just the UPSID$_{451}$ inventory data, then line 2 subsets the 451 rows containing information about UPSID$_{451}$-specific inventories into the "upsid" variable. Line 6 subsets all vowel counts from segment inventories in the PHOIBLE data set into a "vowels" variable. If in line 6 the "data.all" variable is changed to "upsid" (having already fired line 2), then the vowels variable would contain 451 rows and each would contain the total vowel count of a particular segment inventory in UPSID$_{451}$. This can of course also be applied to the total number of phonemes, consonants and tones by subsetting a particular column by its header label, e.g. "variable$header_label", so "data.all$phonemes", "data.all$TopLevel_consonant", etc. The subsetted data can then easily be probed for basic statistics, as shown in Example 3.25.

(3.25)
```
1. range(vowels)
2. min(vowels)
3. max(vowels)
4. mean(vowels)
```

Line 1 in Example 3.25 will show the range of vowels in the data, e.g. in UPSID$_{451}$ the range of vowels in segment inventories is 3-46; in the combined PHOIBLE data set this range is between 2-50. Lines 2, 3 and 4 show how to access the min, max and mean of "vowels" or other variable sets.

The examples provided here are quite elementary, but R is a very powerful tool that offers many research possibilities in combination with the PHOIBLE data warehouse flat file tables. For example, with just a few lines of code we can calculate the mean segment inventory size of unique languages in the PHOIBLE data that are spoken in a particular geographic area, which I do in Chapter 5 in Table 5.11 on page 250.

(3.26)

```
1. data.all <- read.delim("1089_Master_ResultSet_Aggregated.tab",
header=T, sep="\t", quote="\"", dec=".")
2. data.all <- subset(data.all, SourceTrumpOrdering == 1)
3. area.counts <- data.all[, c("area", "phonemes")]
4. africa <- subset(area.counts, area == "Africa")
5. mean(africa$phonemes)
```

In Example 3.26, line 1 again reads in the data. Line 2 gets a unique set of inventories based on the trump hierarchy. Line 3 creates a data frame of geographical regions and total phoneme counts. Line 4 subsets that data frame into just inventories in Africa. And line 5 calculates the mean. In coordination with R's *maps* and *fields* libraries, the geo-coordinates for each language in the PHOIBLE data set also allows users to plot languages as data points on a map. I provide an example in Figure 3.16 on page 149. In Chapter 5 I use both the aggregated and phoneme level data warehouse flat files and R to investigate statistical patterns in segment inventories and I implement in R a genealogical sampling method to take into account language descent using the genealogical data in the PHOIBLE data set. In Chapter 7, I also use the PHOIBLE flat files and R with some more advanced statistical approaches to investigate the purported correlation between population size and phoneme inventory size. Thus the flat files discussed in this section prove very useful for investigating properties of segment inventories in the world's languages.

To summarize, relational databases are typically designed with rules of database normalization to handle transactions by optimizing for performance and data integrity. In the previous section, I described PHOIBLE's relational database model and how different types of data are connected. The relational database is a strong tool for collecting and aggregated data, but it is not ideal for querying. In this section I discussed how I denormalized and extracted data from the relational database into two flat file tables that I call data warehouses. The data in these tables is generated from a SQL procedure that extracts the data using a data warehousing approach and then adds some additional analysis, such as the trump hierarchy, compositional make-up of segments, and total phoneme, consonant, vowel and tone counts. I then showed how these flat file tables can be easily queried and analyzed. In the next section I describe transforming the flat file tables into a graph data model and I discuss the advantages of this data model.

3.2.3 RDF graph

In the previous sections, I described PHOIBLE's relational database schema and how I use a data warehouse procedure to denormalize its data to extract two flat file tables. I use these flat files as input for programming scripts that transform the data into RDF. The benefits of having an RDF/OWL knowledge base implementation and three examples of how to use and query the knowledge base are presented in this section. I then show how the PHOIBLE segment inventory RDF graph can be merged with an RDF graph of distinctive features, so that segment inventories can be queried at an even deeper level than the segment. Furthermore, with OWL a user can add logically-defined statements to the graph, which a reasoner can use to inferred triples to the merged graphs. Examples below will make these processes clearer.

Hebeler et al. (2009) argue that the relational database forces users into a schema-centric perspective. Interaction with the data deals with low-level details of tables, columns, rows and keys. The main challenge is how to join data tables in ways that create the interrelated sets that allow queries to be answered, as we have seen in the previous section. Adding additional data requires the user to join new data based on existing tables' IDs or new relations must be established.

However, the RDF/OWL knowledge base forces users into an open data perspective. Data is the central driving factor and meaning is applied directly to relationships among the data, rather than being centered on programming instructions that extract meaning from the data. The knowledge base decouples the data from the programming instructions and provides a dynamic resource of distributed data. It is dynamic because it allows for inclusion of new types of data at any time and allows anyone to state anything about any topic, i.e. there is an open-world assumption and a non-unique naming assumption explicit in the data model.[59] This requires a different perspective than data-centric programming because of the ability of the model to encode logical inference. Adding a new statement to the knowledge base can ripple through it and transform the data in intended or unintended ways. Therefore, knowledge engineering takes center stage in knowledge base development.

Let us first consider a simple example of querying PHOIBLE for Crothers's (1978) observed near universal that most languages have the vowels /i, a, u/.[60] To simplify things, for the relational database query I will use the simple flat file structure from Section 3.2.2 and abstract away from my project-specific database schema. A few of the current 50k+ rows were given in Table 3.15 on page 127. Each row in the table corresponds to a segment in a particular inventory. Additional data that are specific to a language, like the language family code or population, are repeated.

Using a simple Python script, the data from the phoneme level table can be read in and written out as a simple RDF graph. Example 3.27 provides a snippet of PHOIBLE segment inventories in RDF, serialized in RDF/XML.

(3.27)
```
<?xml version="1.0"?>
<rdf:RDF
  xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:phoible="http://phoible.org/"
  xmlns:owl="http://www.w3.org/2002/07/owl#"
  xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#">
<rdf:Description rdf:about="http://phoible.org/id/iso639-3/ant">
  <phoible:hasSegment rdf:resource="http://phoible.org/segment/ɾ" />
</rdf:Description>
```

[59] See: Section 3.1.

[60] For overview and discussion, see Sections 4.3.1 and 5.6.

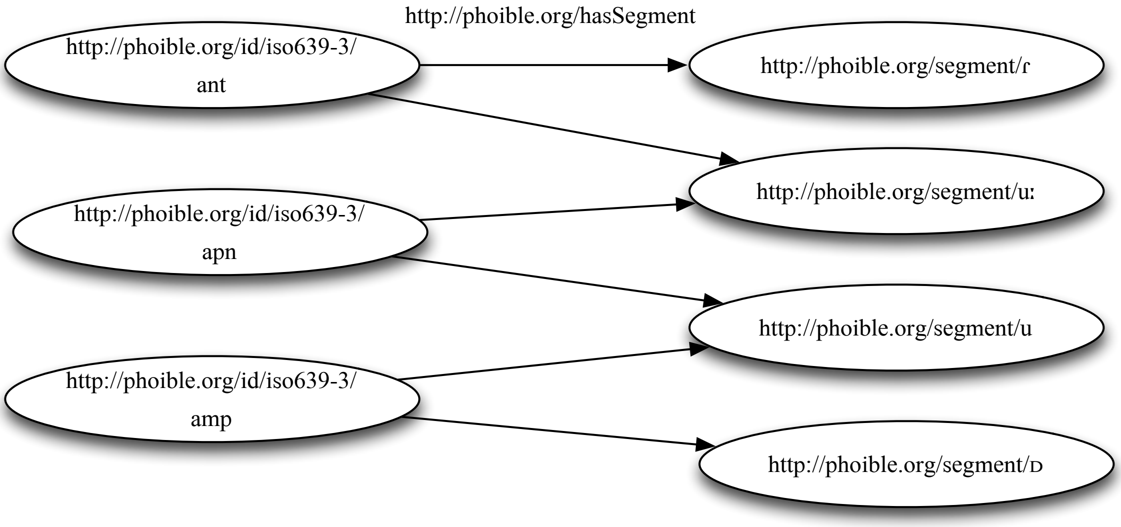
```
<rdf:Description rdf:about="http://phoible.org/id/iso639-3/ant">
  <phoible:hasSegment rdf:resource="http://phoible.org/segment/uː" />
</rdf:Description>
<rdf:Description rdf:about="http://phoible.org/id/iso639-3/apn">
  <phoible:hasSegment rdf:resource="http://phoible.org/segment/uː" />
</rdf:Description>
<rdf:Description rdf:about="http://phoible.org/id/iso639-3/apn">
  <phoible:hasSegment rdf:resource="http://phoible.org/segment/u" />
</rdf:Description>
<rdf:Description rdf:about="http://phoible.org/id/iso639-3/amp">
  <phoible:hasSegment rdf:resource="http://phoible.org/segment/u" />
</rdf:Description>
<rdf:Description rdf:about="http://phoible.org/id/iso639-3/amp">
  <phoible:hasSegment rdf:resource="http://phoible.org/segment/ᴅ" />
</rdf:Description>
</rdf:RDF>
```

RDF/XML is not the prettiest format for human consumption, so in Figure 3.14 I provide a graph illustration of this snippet. The RDF file is a model that consists of language codes and segments that are associated with those codes by the PHOIBLE-defined *hasSegment* predicate.

Now, to query the aggregated table data with MySQL for inventories that have the vowels /i, a, u/, a user could use JOIN statements to get the intersection of the results from queries that return languages that have an /a/, an /i/ and an /u/. The query is given in 3.28.

(3.28)
```
SELECT a.language_code_id
FROM ( SELECT DISTINCT language_code_id, glyph
FROM Master_ResultSet_PhonemeLevel
WHERE glyph = 'a') a
INNER JOIN (
SELECT DISTINCT language_code_id, glyph
FROM Master_ResultSet_PhonemeLevel
```

Figure 3.14: Snippet of PHOIBLE RDF segments graph


http://phoible.org/hasSegment
http://phoible.org/id/iso639-3/ ant
http://phoible.org/id/iso639-3/ apn
http://phoible.org/id/iso639-3/ amp
http://phoible.org/segment/ɾ
http://phoible.org/segment/uː
http://phoible.org/segment/u
http://phoible.org/segment/ᴅ

```
WHERE glyph = 'i') i
ON a.language_code_id = i.language_code_id
INNER JOIN (
SELECT DISTINCT language_code_id, glyph
FROM Master_ResultSet_PhonemeLevel
WHERE glyph = 'u') u
ON a.language_code_id = u.language_code_id
AND i.language_code_id = u.language_code_id
ORDER BY a.language_code_id
```

On the other hand, RDF's graph structure does not require joining various pieces since the query seeks to match triples within the graph. A SPAQRL query to retrieve the sample results is shown in Example 3.29.

(3.29)
```
SELECT ?languages
WHERE {
  ?languages phoible:hasSegment i .
  ?languages phoible:hasSegment a .
```

```
    ?languages phoible:hasSegment u
  }
```

Both queries return 835 records from 1089 distinct inventories (about 79%). This might seem a bit low considering Crothers's observation that 98.5% of languages in the SPA sample contained /i, a, u/. Upon closer inspection though, both queries miss the descriptions of languages that contain similar vowels outside of this tight range of precisely defined characters <i>, <u> and <a>.

Some language descriptions contain the same vowel qualities with long vowels but not their short counterparts. For example, four inventories in PHOIBLE are described as containing <i>, <uː> and <a>, but not <u>. One of those languages from UPSID$_{451}$ is Noni [nhu], which has the vowels /i, iː, e̞ː, ɛ, ɛː, uː, ʊ, o, oː, ɔ, ə, ɔː, a, aː/ (Hyman, 1981; Maddieson and Precoda, 1990).[61] Differences in vowel quality also play a role. Fifteen segment inventories contain /ʊ/ but no /u/, including Wik-munkan [wim] (/i, ɛ, ʊ, ɔ a/) (McConnel, 1945; Sayers and Godfrey, 1964; Maddieson and Precoda, 1990).[62] Wik-munkan is also a nice example of the difficulty in interpreting language descriptions.[63]

Returning to our query of which languages /i, u, a/ occur in, at this point some choices need to be made if we want to expand our search space for criteria like vowel length or quality. On the one hand, we can expand the SQL query by adding logic operators like OR to it, which would arguably make the query even more complicated. On the other hand, in the RDF

[61] The /e̞ː/ is my IPA rendition of UPSID$_{451}$'s *mid front unrounded vowel* <”e:>; compare UPSID$_{451}$'s *higher mid front unrounded vowel* /e/. See also Appendix F for our UPSID-to-IPA mappings.

[62] Crothers (1978, 103) collapses all three vowel systems into /i, u, a/, although the phonetic variations vary considerably in cases like [u].

[63] Wik-munkan is listed in SPA with the contrastive vowels /i, iː, ɛ, ɛː, ʊ, ʊː, ɔ, ɔː, a, aː/ (Sayers and Godfrey, 1964; Crothers et al., 1979). There is a footnote on the lengthened vowels in its inventory that states: “The vowel qualities for the long vowels are not separately specified, since the analysis is in terms of five vowels plus a ‘length’ phoneme” (Crothers et al., 1979, 630). It seems that Maddieson chose not to include Wik-munkan's length series, which is not clearly reported in SPA, when normalizing comparative segments for UPSID. In general vowel length does appear in inventories in UPSID$_{451}$. However, in an inventory such as Bambara's [bam], long vowels do not appear in the inventory, even though there is a comment in UPSID$_{451}$ that states “All vowels also appear long”. See: `http://web.phonetik.uni-frankfurt.de/L/L4105.html`. Putting confusing interpretations of such inventories aside, length is also considered a “series-generating component” (Maddieson, 2007), as is nasalization, length, voice quality and tone. There is strong disagreement on whether or not these features should be included or excluded in summaries or statistical analyses of segment inventories.

model we can simply add an additional layer of knowledge *to the model*, which allows us to leverage logical inference to query the knowledge base without changing the underlying data in it and without changing our query. By using logically-defined properties in OWL, and then merging the OWL and RDF graphs, we can establish relationships between resources in our graph that are inferred by a semantic reasoner, i.e. a piece of software that infers the logical consequences in the graph and that adds any logically inferred triples to that graph before the query is fired.

This process of adding additional logically-defined statements to the graph and running the reasoner is rather straightforward. One method to accomplish this is to use the OWL property *owl:sameAs*, which links an individual to an individual by stating that two URIs refer to the same individual. Example 3.30 uses the *owl:sameAs* property to indicate that the segment /uː/ is the same individual as the segment /u/.

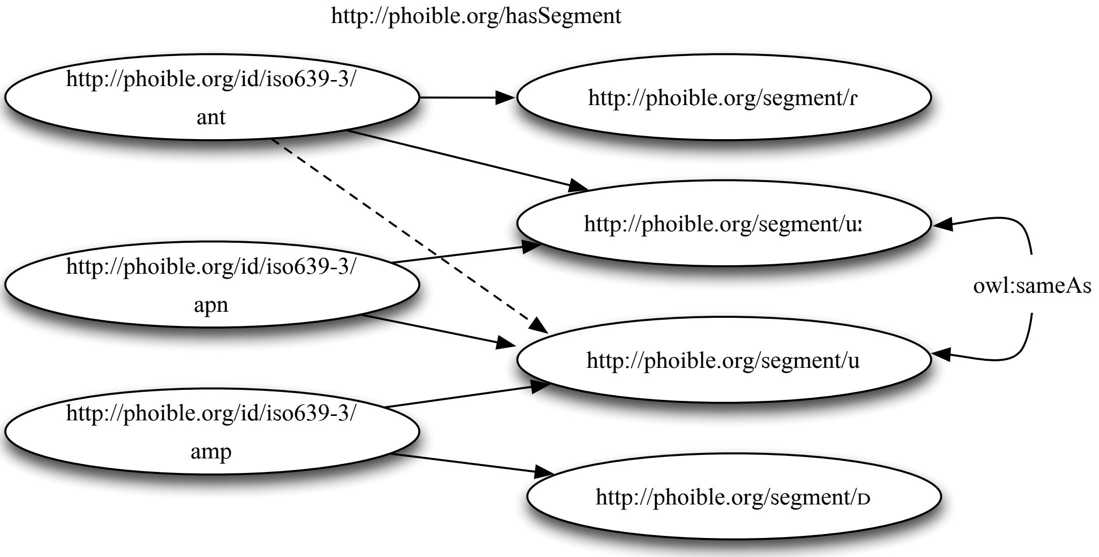
(3.30)
```
<rdf:Description rdf:about="http://phoible.org/segment/uː">
  <owl:sameAs rdf:resource="http://phoible.org/segment/u"/>
</rdf:Description>
```

When this additional knowledge is added to the knowledge base and loaded with the RDF file of segment inventories, an OWL DL[64] reasoner infers the additional triples and adds them to the graph before querying. Figure 3.15 illustrates the *owl:sameAs* relation and the subsequent inferred knowledge denoted by the dotted line. This information is not permanently added to the model, so it can be used for some queries both not others. In other words, the *owl:sameAs* predicate is not persistent. Instead of returning 836 records, querying for <i>, <u> and <a> now returns the 840 languages, which includes the four additional languages that are described as having /i/, /a/ and /uː/, but not /u/.

Alternatively, we could specify /u/ *owl:sameAs* /uː/ since the relation is symmetric and then just query for <i>, <u> and <a>. If we don't want our queries to differentiate vowels because of their length, we can remap all lengthened vowel individuals to their short counterparts. For Crothers's query, if we include /a/ *owl:sameAs* /aː/ our query returns 846 languages, and additionally with /i/ and /iː/, it returns 873 (of 1088 languages, just

[64] See Section 3.3 for a discussion of the different types of OWL.

Figure 3.15: Snippet of PHOIBLE RDF segments graph with inferred triples


http://phoible.org/hasSegment
http://phoible.org/id/iso639-3/ant
http://phoible.org/id/iso639-3/apn
http://phoible.org/id/iso639-3/amp
http://phoible.org/segment/ɾ
http://phoible.org/segment/uː
http://phoible.org/segment/u
http://phoible.org/segment/ᴅ
owl:sameAs

over 80%). This still isn't near Crothers's claim of 98.5%. However, equipped with OWL and the ability to add properties and restrictions to individuals, and SPARQL to query the knowledge base, we have tools to investigate the matter. For example, querying inventories that have /i, ɑ, u/, i.e. a back [ɑ] but no front /a/, returns an additional 42 inventories, which brings the result count up to 84% of languages in PHOIBLE.

Crothers's query is one example of how to use and interact with the PHOIBLE knowledge base. It illustrates an important property of working with RDF/OWL – the ability to manipulate the knowledge base through an ontology and to specify how to derive logical consequences and to create new entailments. Readers may have noticed the non-IPA symbol <ᴅ> in Example 3.27 and Figures 3.14 & 3.15. This symbol was included in PHOIBLE because UPSID_{451} distinguishes between a voiced alveolar tap (denoted by the symbol /ᴅ/) and a voiced alveolar flap (/ɾ/). The distinction does not exist in the IPA, where tap and flap are collapsed into one manner of articulation. In fact, if we look closely at the seven languages in UPSID_{451} that have a voiced alveolar tap and the 91 languages that have a voiced alveolar flap, there is no overlap between the two sets, i.e. there is no language in UPSID$_{451}$ that contrasts voiced alveolar tap and voiced alveolar flap.[65] In the feature set used in UPSID$_{451}$, the two phonemes contrast in precisely two features "tap" and "flap". Thus the distinction may be an effect of transcription symmetry and is important for some contrastive aspect of language-specific inventories. Nevertheless, when querying for contrastive segments across languages in the database/knowledge base, one might wish to treat these two segments as the same segment.[66] Moreover, the tap/flap distinction is not the only distinction in UPSID$_{451}$ that one might wish to collapse for various reasons. Another is the voiceless retroflex sibilant fricative (23 languages)[67] and the voiceless retroflex fricative (1).[68] Or perhaps one would like to remap or collapse the underspecification of some or all of the 99 dental/alveolar sounds found in UPSID$_{451}$.

Instead of investigating a (near) universal of segment inventories, now let's look at investigating a property of a specific language. Querying segment inventories in the knowledge base for segments or series of segments is straightforward. This example comes from Scott Sadowsky, who works on Mapudungu [arn], a language spoken in Chile. Sadowsky wanted to know how many languages have the following phoneme co-occurrences and if any languages have all four phonemic oppositions:

1. Both (i) a voiced dental/interdental nasal, and (ii) a voiced alveolar nasal (e.g. dental/interdental /n̪/ and alveolar /n/).

2. Both (i) a voiceless dental/interdental plosive, and (ii) a voiceless alveolar plosive (e.g. dental/interdental /t̪/ and alveolar /t/).

3. Both (i) a voiceless dental/interdental fricative, and (ii) a voiceless alveolar fricative (e.g. dental/interdental /θ/ and alveolar /s/).

[65] Henning Reetz presents a nice HTML interface for browsing UPSID$_{451}$ inventories online. For a list of inventories in UPSID$_{451}$ with a voiced alveolar tap, see http://web.phonetik.uni-frankfurt.de/S/S0773.html. For inventories with a voiced alveolar flap, see http://web.phonetik.uni-frankfurt.de/S/S0774.html. A PHOIBLE web interface is forthcoming and will be available with static URLs at: http://phoible.org/.

[66] For example, Hyman (2008, 89) collapses Maddieson's [ᴅ] with [ɾ].

[67] http://web.phonetik.uni-frankfurt.de/S/S0787.html

[68] http://web.phonetik.uni-frankfurt.de/S/S0152.html

4. Both (i) a voiced dental/interdental lateral approximant, and (ii) a voiced alveolar lateral approximant (e.g. dental/interdental /l̪/ and alveolar /l/).

5. How many have all four of these oppositions phonemically?

The query given in Example 3.31 retrieves any inventories where the triple pattern matches: *?languages* (the variable), *phoible:hasSegment* (predicate), and */n/* and */n̪/* (objects).

(3.31)
```
SELECT ?languages
WHERE {
  ?languages phoible:hasSegment n̪ .
  ?languages phoible:hasSegment n
}
```

The same query can then be used for the other phoneme pairs, given in (2), (3) and (4), by simply replacing the object segments. The query in (5) is just a conglomeration of the previous four queries, shown in 3.32.

(3.32)
```
SELECT ?languages
WHERE {
  ?languages phoible:hasSegment n̪ .
  ?languages phoible:hasSegment n .
  ?languages phoible:hasSegment t̪ .
  ?languages phoible:hasSegment t .
  ?languages phoible:hasSegment θ .
  ?languages phoible:hasSegment s .
  ?languages phoible:hasSegment l̪ .
  ?languages phoible:hasSegment l
}
```

The results of the five queries are given in Table 3.18. Using R's *maps* and *fields* libraries and geo-coordinates data available in PHOIBLE's data warehouse flat files, I've plotted the results on a geographical map, shown in Figure 3.16. Sadowsky's intuition to investigate the possibly peculiar phoneme combinations in Mapudungu shows that out of 1089 inventories, 48 languages contrast dental /t̪/ and alveolar /t/, 40 contrast dental /θ/ and alveolar /s/, 21 dental /n̪/ + alveolar /n/ and 11 dental /l̪/ and alveolar /l/. However, only Mapudungu has all of these contrasts in the PHOIBLE data set.

Table 3.18: Distribution of segments /n̪, n, t̪, θ, s, l̪, l/ across PHOIBLE

n̪+n	t̪+t	θ+s	l̪+l	n̪+n+t̪+t+θ+s+l̪+l
Alyawarra	Alyawarra	Aja	Alyawarra	
Anywa	Anywa	Albanian	Arrarnte	
Arrarnte	Arrarnte	Amahuaca	Arrernte	
Arrernte	Arrernte	Aneityum	Digueno	
Boiken	Betta Kurumba	Aragonese	Diyari	
Digueno	Brahui	Asmat	Kalakatungu	
Dinka	Brokskat	Baka	Macedonian	
...	...	...	...	
Mapudungu	Mapudungu	Mapudungu	Mapudungu	Mapudungu
...	...	...	...	
Total = 21	48	40	11	1

Of course this isn't the whole picture. Querying contrastive segment types can only get us so far. Mapping vectors of distinctive features to segments in the knowledge base provides users with a deeper level of granularity for investigating patterns in and across segment inventories. However, there are several computational issues to overcome in assigning features to segment types.[69] The first major hurdle is that distinctive feature sets have poor typological coverage when compared with segment types that appear in broad cross-linguistic segment inventories like PHOIBLE. We cannot simply use a feature matrix as a look-up table for assigning feature vectors to segment types (without first defining a feature vector for every segment type in the data set). Moreover, segment types belong to one of three segment classes, i.e. simple, complex and contour, and each requires a different

[69] See discussion in Chapter 6.

method for collapsing features. A simple segment type is a single segment or a segment with one or more diacritics. Diacritics overwrite features in the base segment and can occur before or after the base. Complex segments consist of dually articulated segments like /kp/ in which certain features overwrite other features. Contour segments, e.g. pre- and post-nasalized consonants, affricates and contour tones, present the main challenge in automatically assigning features to segment types from a feature set because they encode temporal movement of phonetic features. Thus the single-tiered distinctive feature vector that is used in typical distinctive feature matrices does not straightforwardly merge with another feature vector. This is an issue that feature geometry set out to address (Clements, 1985; Sagey, 1986; McCarthy, 1988; Clements and Hume, 1995).

The distinctive feature set developed in this work to address these issues is an expanded feature set based on Hayes 2009, which I will call Hayes′ (Hayes prime). Hayes′ has been expanded with several feature types (e.g. fortis, ATR, click, tone) to achieve coverage of all segment types in PHOIBLE.[70]

Returning to Sadowsky's question of phoneme co-occurrences, at the segment level the query for alveolar and dental phoneme pairs was too specific, i.e. it only asks about these specific groups of phonemes. Other languages could also have four alveolar/dental phoneme pairs, but the segments may differ along different feature planes, say in manner of articulation or voicing, e.g. an affricate rather than a plosive pair, or voiced fricatives instead of voiceless ones.

Table 3.19 illustrates the (partial, but relevant) feature vectors and how they contrast for alveolar and dental phonemes. These alveolar and dental segments belong to the simple segment class, i.e. the alveolar sounds [t, s, n, l] can be assigned the feature vectors assigned to them in Hayes 2009. The dental sounds [t̪, n̪, l̪, s] are also assigned the alveolar sounds' feature vectors, but the features of the dental diacritic [+anterior, +distributed] overwrite the relevant cells of the alveolar sounds' feature vectors.[71] This results in the set of alveolar

[70] See Section 6.3 for an evaluation of the typological coverage of features in Hayes 2009 and the UPSID_{451} feature set (Maddieson and Precoda, 1990), as applied to the contents of PHOIBLE.

[71] In Section 6.4, I present the problems involved in mapping features to segment types and discuss my computational approach that allows users to query segments and segment inventories in the knowledge base at the feature and feature geometry levels using RDF/OWL.

and dental pairs shown in Table 3.19 that contrast minimally in [±distributed].

The SPARQL queries in Examples 3.33 & 3.34 illustrate how to query for languages containing the segments [t, n, l, s] and for languages containing the segments [t̪, n̪, l̪, θ].

(3.33)
```
SELECT DISTINCT ?languages
WHERE {
  ?languages phoible:hasSegment ?segments .
  ?segments phoible:hasFeature feature:ANTERIOR .
  ?segments phoible:notHasFeature feature:DISTRIBUTED
}
```

(3.34)
```
SELECT DISTINCT ?languages
WHERE {
?languages phoible:hasSegment ?segments .
?segments phoible:hasFeature feature:ANTERIOR .
?segments phoible:hasFeature feature:DISTRIBUTED
}
```

Modeling segments, distinctive features and their relationships in an RDF/OWL knowledge base allows us to investigate segment inventories at the feature level. This model could also be implemented in relational database tables, as Maddieson and Precoda (1990) did for the UPSID$_{451}$ data. Again, the additional of yet another relational database table, or multiple ones in the case of different feature sets, would increase the complexity of querying the database's contents.

However, now that we have distinctive feature vectors mapped to segment types, we can harness the power of OWL to develop an ontology (or minimally a taxonomy) of features by defining the hierarchal and logical relationships between feature classes. Figure 3.17 is a visualization of an OWL file that encodes the Hayes′ features in a feature geometry modeled on Clements and Hume 1995. The "is-a" relationships in the hierarchy represent OWL *subClassOf* relations. Daughter classes inherit the features of their parent node. Elements of feature geometry can be used to query segments in inventories or query on class types as a shorthand for feature bundles by merging RDF graphs, e.g. return all roots (segments) that are [+nasal] but underspecified for place of articulation (i.e. the archiphoneme /N/ discussed in Section 2.3.4).

Figure 3.16: Dental/interdental and alveolar phonemic contrasts in several languages

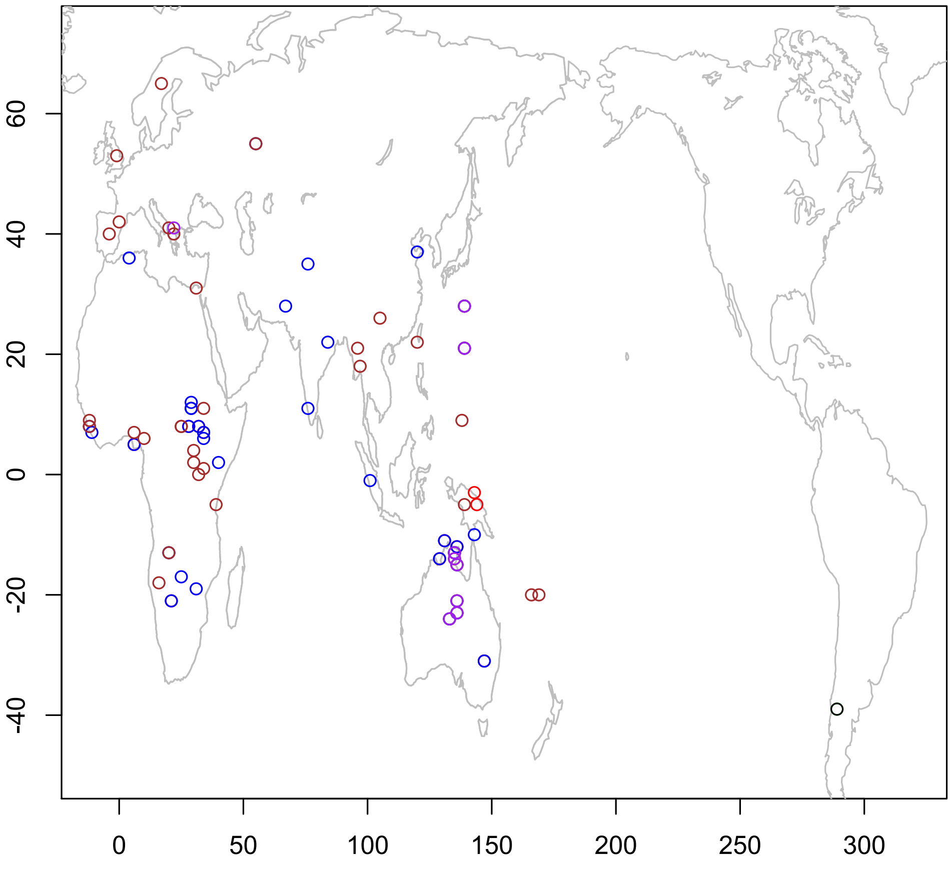
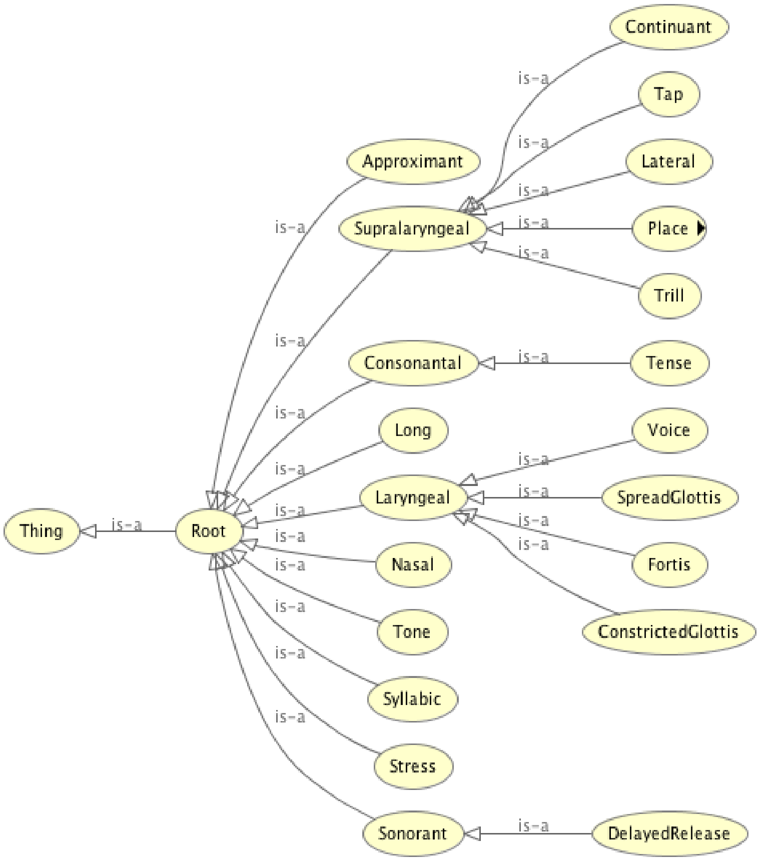


Table 3.19: Dental and alveolar feature pairs

segment	sonorant	continuant	delayed release	approximant	nasal	voice	coronal	anterior	distributed	strident	lateral
t	-	-	-	-	-	-	+	+	-	-	-
̪								+	+		
t̪	-	-	-	-	-	-	+	+	+	-	-
n	+	-	0	-	+	+	+	+	-	-	-
n̪	+	-	0	-	+	+	+	+	+	-	-
l	+	+	0	+	-	+	+	+	-	-	+
l̪	+	+	0	+	-	+	+	+	+	-	+
s	-	+	+	-	-	-	+	+	-	+	-
θ	-	+	+	-	-	-	+	+	+	-	-

Figure 3.17: Hayes′ feature geometry



In this section I presented a simple RDF model of the PHOIBLE data and gave an example of how additional knowledge can be added to knowledge bases via OWL properties to define relationships in the data. I then used this knowledge to query patterns of segments and features found in a sample of the world's languages. I also showed how segments can be modeled as a set of distinctive features in an RDF graph and merged with an RDF graph of segments to produce a resource to query segment inventories at the level of features.

3.2.4 Summary

There are many different ways to store data. Therefore, it is important to decide on an approach that meets the particular needs of the users of the data. I've shown in this section that there is no one way to model data that addresses all query types, while making it easy for users to work with a typological data set. Each data model has its pros and cons and I have discussed them in this section.

3.3 Knowledge representation

In the previous section, tabular data, a relational database and knowledge representation in a graph data structure were compared and shown to encode data in different ways and for different purposes. For the PHOIBLE segment inventory database, I used data modeling techniques to define the requirements for querying the database. This type of data modeling is commonly called database modeling, because the intention is to implement a database schema to support the functions of the proposed application. On the other hand, the data modeling of and RDF/OWL graph can be considered a knowledge engineering task. The task of knowledge engineering is to represent knowledge of a particular domain in a machine readable format. What does the domain being modeled look like? The task is to identify similarities and relationships between things. The knowledge engineer constructs an ontological theory that begins with concepts, relations and desired inferences (Sowa, 2000; Farrar, 2003; Farrar and Langendoen, 2010). Instead of modeling data according to database normalization techniques for general purpose querying, the ontological theory explicitly defines objects, properties of objects, and relations among objects in the data. The idea behind this approach is to analyze expert knowledge of a particular domain and then encode it in a knowledge representation language. As such, there may be different solutions for modeling a particular domain of affairs and different knowledge engineers' approaches may lead to differently structured knowledge bases.

Knowledge representation is the intersection of theories and techniques from the fields of logic, ontology and computation. It involves the application of logic and ontology in creating a computable model of a particular domain (Sowa, 2000). Knowledge representation is concerned with the design of formalisms for implementing a computationally and epistemologically adequate conceptualization of a particular domain (Baader et al., 2003). The product of knowledge representation modeling, the knowledge base, is a machine readable description of the domain. The central assumptions in the knowledge base are captured in the ontological theory, i.e. the set of logical statements that describe knowledge of the domain. These sets of statements are often referred to simply as *ontologies*. The ontology can be used by automated reasoning tools to produce new knowledge, enhance search, and prove the consistency of logical propositions in the knowledge base.

Knowledge representation languages are formalisms used to represent knowledge. Popular knowledge representation languages include logic, frames, production rules and semantic networks. Each knowledge representation language has its own advantages and disadvantages. These formalisms each have both a syntactic and an inferential feature. The syntactic feature provides a mechanism for explicitly encoding information in the knowledge representation language. The inferential feature provides mechanisms for deriving implicit information from that knowledge store. This section explores knowledge representation, Description Logics, ontology and the knowledge base.

3.3.1 Representing knowledge

Knowledge representation is the study of representing knowledge in formal structures and identifying what kinds of reasoning can be computationally modeled with that knowledge. Knowledge-based systems have been implemented in different formalisms including frames, rules and semantic networks. These techniques have in common the ability to denote objects, object properties and relations among objects. Knowledge-based systems have at their core a knowledge base and mechanisms for deriving inferences from the logical propositions encoded in that knowledge base. Propositions in the knowledge base are explicitly encoded objects, properties and relations in the domain of discourse – a model for a particular domain of expert knowledge. The basic idea is that the model's formal representation makes the connection between some state of real world knowledge and a computable model of them that can be used for tasks like scientific investigation.

What is it about a knowledge representation language that provides the ability to perform the tasks that the knowledge engineer desires? Nonmonotonic reasoning aside, Hayes (1985, 4) remarks, "virtually all known representational schemes are equivalent to first-order logic".[72] Using formal logics for knowledge representation languages provides a precise model theory. There are several computational requirements for representing knowledge, as pointed out in Jurafsky and Martin 2009, chap. 14. Here I address those required for the task at hand: modeling languages' segment inventories and combining segments with different distinctive feature sets in a knowledge representation language.[73]

The first requirement is verifiability. The knowledge represented must be able to be verified, i.e. the truth of propositions in the knowledge base must be determinable. The state of affairs described in the knowledge base can then be compared to the state of affairs that is being modeled. The second requirement is unambiguous representations; the system should have the ability to reach a final representation that is unambiguous. Third, inference and variables are required for representing knowledge. Inference is the ability of a system to reach a conclusion based on evidence and the ability to reason over truth propositions that are logically derivable, but not explicitly encoded in the knowledge base. Variables are needed for matching propositions in the knowledge base against queries. Lastly, expressiveness is the measure of the level of expressivity of a knowledge or meaning representation language. This requirement rests on the interpretability of the formalism used to describe the model. The expressivity is defined by the logic that provides a formal semantics for the knowledge

[72] See also Hayes 1977; 1979.

[73] Jurafsky and Martin (2009) note the canonical form as a computational desideratum for representing meaning to linguistic input. However, for the task at hand this is irrelevant and therefore is not mentioned as a computational requirement for representing knowledge.

representation. For example, first-order logic is more expressive than Description Logics, a family of highly structured languages that are a fragment of first-order logic (Baader et al., 2003). These logic foundations provide different levels of expressivity. The expressivity also determines to what degree the data in a knowledge representation can be reasoned over. Using Description Logics provides improved computational tractability, but they are less expressive than first order logic. Farrar and Langendoen (2010) show that linguistic data can be modeled in OWL-DL, the web ontology language that is most closely expressed by the Description Logic $\mathcal{SHOIN}(\mathbf{D})$. Pellet, a complete OWL-DL reasoner, can be used to perform computationally tractable inference on OWL-DL knowledge bases (Sirin et al., 2007).[74]

These computational requirements are considerations that must be addressed to guarantee that the knowledge representation achieves its purpose. To represent meaning, a representation formalism is needed. The formal representation should tell the user something about the domain of the model and it should accurately describe facts concerning the state of affairs of the intended model. The model, or in ontological terms, the knowledge base, is the formal representation of a state of affairs modeled from the real world. If this model accurately reproduces that state of affairs, then the user is able to leverage the knowledge representation language to access explicit and implicit information about the modeled state of affairs.

To summarize, this section has described the computational requirements for representing knowledge in machine-readable formats. The language of knowledge representation, its underlying logical foundation, is discussed in the next section on Description Logics, a family of knowledge representation languages that are proven computable fragments of first-order logic.

3.3.2 Description Logics

To represent the meaning of linguistic expressions in formal structures, some type of logic formalism is needed. This section introduces the class of logical formalisms known as De-

[74] http://clarkparsia.com/pellet/

scription Logics (DL) (Baader et al., 2003; Baader and Sattler, 2001; Calvanese et al., 2001).[75] DL is a mathematical theory and a formalism for representing knowledge. It is equipped with a logic-based semantics that provides the logical formalism for ontologies. An ontology, discussed in detail in the next section, is a set of statements that denote a particular conceptualization of a domain. Because real-world domains are incredibly complex, a conceptualization of a domain is an abstract and simplified view of the world (Gruber, 1993). As a means of axiomatization for conceptualization, logic is used in ontology development.

Formal logic languages provide a mechanism for evaluating the verifiability of a statement. They may also facilitate certain types of inference. An important computational desideratum for modeling the semantics of language is expressiveness (Jurafsky and Martin, 2009, chap. 14). Different logic formalisms have different expressive power, i.e. the degree to which ideas are expressible in a formalism. For example, first-order logic (FOL) is a well understood language and is expressive enough to handle many aspects of natural language semantics (e.g. quantifiers, conjunction, disjunction, etc.). However, FOL is generally undecidable, therefore its deductive system cannot provide the truth value of certain types of statements in a finite time. DLs are an alternative to FOL that provide improved computational tractability at the cost of expressivity. They are more expressive than propositional logic, more computationally tractable than FOL, and more efficient than FOL at determining decision problems (a question of a formal system with a yes or no answer). The formal language used to represent knowledge restricts what kinds of domain knowledge can be encoded. For a particular DL, its expressiveness is determined by the concept and role constructions it supports (Horrocks et al., 2003). DLs are advantageous because they always yield a correct answer in finite time and many DL systems come with reasoning services that use explicitly represented knowledge to automatically deduce implicit knowledge (Baader et al., 2008).

Like all formal logics, DLs have a proof theory. The proof theory determines entailments from a set of statements in the logic formalism. DLs are decidable structured fragments of FOL and their expressivity is encoded with labels (represented with letters) for describing

[75] See Baader et al. 2003 for a full account of the semantics of DL. For basic notions of DLs see Baader and Nutt 2003, and for linguistic examples see Farrar and Langendoen 2010.

the logic operators allowed. OWL-DL is derived from the $\mathcal{SHOIN}(\mathbf{D})$ family of description logics. “S” stands for the modal logic S4 (Horrocks et al., 2003); “H” indicates role hierarchies; “O” indicates individuals (nominals) are included; “I” indicates that inverse roles are allowed; “N” indicates number restrictions are allowed (cardinality restrictions); and “D” indicates the use of datatype properties, data values or data types.

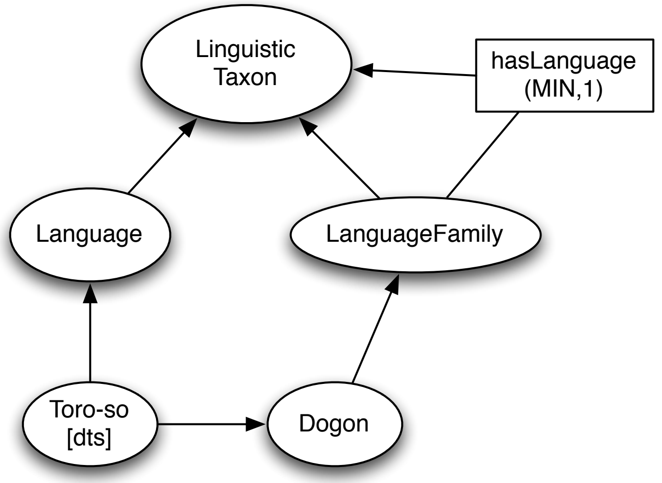
To provide an example of representing knowledge in DL, consider the basic graph in Figure 3.18. The nodes represent concepts (sets or classes of individual objects) and the links between nodes represent relationships among the concepts. The relationship between `Dogon` and `LanguageFamily` represents an “is-a” relationship. “is-a” is a subsumption relationship where one concept (or class) is a subclass of another. The more specific concept inherits the properties of the more general and these relations define a hierarchy over the concepts. In Figure 3.18 `Dogon` is a language family, and subsumes (or inherits) all the properties of a `LanguageFamily`, just as `Toro-so` is a language and subsumes the characteristics of `Language` (e.g. has `MorphoSyntacticProperty`), and `Language` and `LanguageFamily` subsume properties of `LinguisticTaxon`.[76] A feature of DLs is their ability to represent relationships beyond “is-a” (Nardi and Brachman, 2003). In this simple network, `LanguageFamily` has a value restriction expressing a limitation that it must have one or more Languages.

Table 3.20 provides a comparison of constructors in $\mathcal{SHOIN}(\mathbf{D})$ and OWL-DL and illustrates relations beyond “is-a”.[77] These basic DL constructors can be used to create logical statements, resulting in the axioms (assertions of knowledge) that define restrictions on concepts and roles (the links between concepts). In DL concepts represent unary predicates and roles represent binary predicates. A concept is instantiated by individuals and represents a class in the domain being modeled, making an individual an instance of that concept (Nardi and Brachman, 2003). A role (or relation or link) is a binary relation

[76] From GOLD, version 2010, a LinguisticTaxon is: “the class of Taxons whose instances are used in the scientific classification of language varieties. That is, instances of LinguisticTaxon have instances that are human language varieties.”, see `http://linguistics-ontology.org/gold/2010/LinguisticTaxon`.

[77] Farrar and Langendoen (2010, 9) note that in their table: “D is assumed to be a built-in data type and not a declared concept.”

Figure 3.18: Language family subsumption graph



between individuals. By definition, DL has only binary relations, so higher arity relations are disallowed (Farrar and Langendoen, 2010). In DL terminology, concept, individual and role are used instead of class, object and property (or instance), shown in Table 3.21.

In a DL knowledge base, concept descriptions are used to build statements about the domain being modeled. In FOL predicates have equal ontological status, but in DLs their semantics are typically split into concepts and roles (Farrar and Langendoen, 2010). The concept split is known as TBox and ABox and separates concepts and roles from individuals (Baader et al., 2008). By definition, a DL knowledge base (KB) is an ordered pair of the TBox (T) and ABox (A), i.e. KB = <T,A>. T is the union of the set of concepts and roles in the domain. It relates axioms between concepts and roles. A is the set of individuals in the domain. It relates axioms to individuals. The TBox (terminological knowledge) consists of axioms about the properties of concepts and roles, and relationships between them (like the schema in a database setting). Concepts correspond to unary predicates that represent an object (category or kind) in the domain. Concepts are instantiated by individuals, or in other words, an individual is instantiated as an instance of a concept. The ABox (assertion box) consists of facts about instances and individuals. In regard to class

Table 3.20: A comparison of $\mathcal{SHOIN}(\mathbf{D})$ and OWL-DL constructors (Farrar and Langendoen, 2010, 9)

Constructor	$\mathcal{SHOIN}(\mathbf{D})$	OWL-DL
conjunction	$C_1 \sqcap C_2$	unionOf(C_1, C_2)
disjunction	$C_1 \sqcup C_2$	intersectionOf(C_1, C_2)
negation	$\neg C$	complementOf(C)
oneOf	$\{o_1, ..., o_n\}$	oneOf $\{o_1, ..., o_n\}$
exists restriction	$\exists R.C$	someValuesFrom(C); onProperty(R)
value restriction	$\forall R.C$	allValuesFrom(C); onProperty(R)
atleast restriction	$\geq nR$	minCardinality(n); onProperty(R)
atmost restriction	$\leq nR$	maxCardinality(n); onProperty(R)
datatype exists	$\exists R.D$	someValuesFrom(D); onProperty(R)
datatype value	$\forall R.D$	allValuesFrom(D); onProperty(R)
datatype atleast	$\geq nR$	minCardinality(n); onProperty(R)
datatype atmost	$\leq nR$	maxCardinality(n); onProperty(R)
datatype oneOf	$\{v_1, ..., v_n\}$	oneOf $\{v_1, ..., v_n\}$

membership within the TBox's concepts, the ABox describes the roles between instances and other assertions about instances in the knowledge base generated through inference. This split between the TBox and ABox can be useful for reasoning. For example, the ABox can be used for instance checking and the TBox used for classification, since it encodes properties and relations between concepts. The separation may also affect performance in decision procedures for reasoning.

In summary, Description Logics are a family of computationally tractable logic formalisms used in knowledge representation. They are a mathematical theory that have attracted much attention in their role in formally specifying semantics (or metadata) of Web contents as part of the development of a Semantic Web of data (Jurafsky and Martin,

Table 3.21: Terminology of DL vs OWL

DL	OWL
concept	class
individual	object
role	property

2009). OWL-DL most closely resembles the DL $\mathcal{SHOIN}(\mathbf{D})$ and has been successfully used to implement an ontology for describing linguistic morphosyntactic terminology (Farrar, 2003; Farrar and Langendoen, 2003, 2010). Axioms in the knowledge base form a conceptualization of a particular domain and are captured in an ontological theory of that domain. These statements are often simply referred to as *ontology.*

3.3.3 Ontology

The word ontology is derived from Greek *ōn, ont-* "being" + *-logy* and means the study of the nature of being, or the study of "existence". This original sense prevails today. In philosophy, ontology belongs to the branch of metaphysics and its object of study is reality. Ontology concerns itself with a description of concepts (or individuals) and how they relate. These sets can be grouped, subdivided or hierarchically organized.

Modern advances in mathematics and computer science caused ontology to acquire an additional meaning. In 1992, Gruber defined ontology in terms of computer science.[78] What is an ontology? Gruber's (1993) short answer: "An ontology is a specification of a conceptualization." This sense was meant in the context of sharing knowledge, particularly among Artificial Intelligence (AI) software, i.e. "semantics independent of reader or context" (Gruber, 1993). This co-option was troublesome and the term *ontology* may be AI literature's most misused (Bateman, 1995). The use of ontology in both philosophy and

[78] Accessed on July 1, 2011: `http://www-ksl.stanford.edu/kst/what-is-an-ontology.html`

computer science, however, share a common trait: the study or description of entities and their relationships that exist or may exist in some domain. An ontology is the product of such a study.

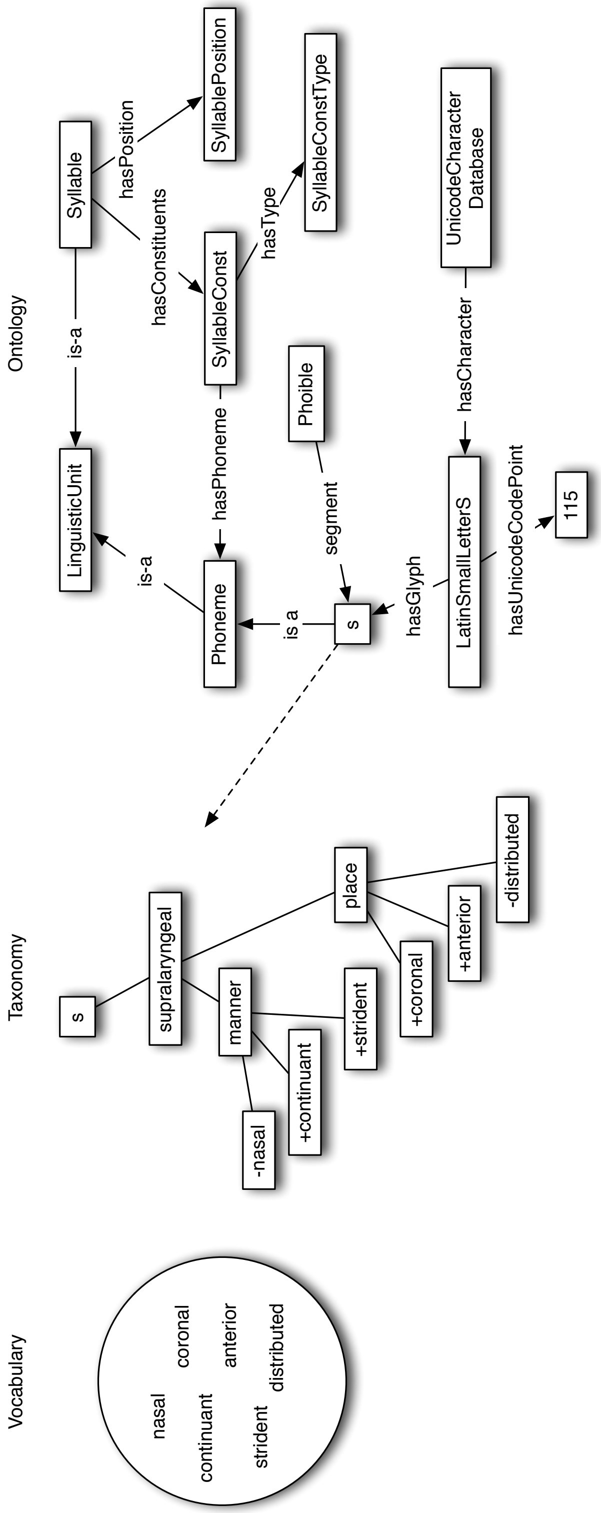
An ontology is used to model domain knowledge. It contains information regarding classes and their relationships, whether abstract or concrete. An ontology uses a well-defined vocabulary of terms to describe concepts and their relationships within a particular domain. Therefore, ontology can actually refer to a vocabulary, a taxonomy or a description of a domain. A vocabulary is a collection of defined terminology. When those terms are given hierarchical relationships, they become a taxonomy.

Compare the images in Figure 3.19, which juxtaposes a collection of terms of phonetic features, a taxonomic phonetic feature representation from Clements 1985, and an example ontology of various concepts and their relations. The simple collection of phonetic feature terms becomes a taxonomy when the terms are extended through hierarchical relationships.[79] The features coronal, anterior and distributed characterize the place node that dominates them and the supralaryngeal node dominates the manner and place nodes (Clements, 1985, 248). In the example ontology, there are many relations, including non-hierarchical ones, that model the relationships between different linguistic units, and segments and their technological encoding.[80]

[79] Some additional information is added by specifying the binary value of each feature, because this is a (partial) hierarchical representation of the [s] segment in Clements (1985). The hierarchical structure for this feature geometry representation remains constant; it changes to the binary specification of the features that allows the representation to describe different segments.

[80] The dotted line denotes that the segment taxonomy can be connected to the ontology.

Figure 3.19: Vocabulary vs taxonomy vs ontology



Whereas a taxonomy is a hierarchical classification of terminology within a domain (typically in a tree format that represents parent-child relations), an ontology is a model of a domain and it specifies the characteristics of the domain by precisely defining the relationships between categories (aka “concepts”, “terms” or “things”). An ontology captures knowledge that is not necessarily hierarchal; it can define any relation between categories. With an ontological description, the semantics behind vocabulary terms and their relationships can be described in a formal logic-based model. In Figure 3.19, for example, the supralaryngeal tier node dominates the manner and place of articulation tier nodes. In one possible ontology, the manner and place of articulation nodes can be modeled as subclasses of the supralaryngeal node. The phonetic features nasal, continuant, coronal, and so forth, are then defined as properties of these classes. In Figure 3.19, an instantiation of these classes and their properties results in an instance of the [s] segment. In the example ontology in the same figure, the taxonomy plays one part in the larger model of the domain that models linguistic units that have to do with phonetics and phonology (segments, phonemes, syllables, etc.) and the technological factors for encoding segments via a Unicode Character Database and PHOIBLE.

The design of an ontology depends on the application in mind. For any domain, there is not a single correct ontology. Instead, knowledge (or ontology) engineering is driven by competency questions, i.e. questions that the ontology should be able to answer (Grüninger and Fox, 1995). These competency questions are used to define the ontology’s requirements. The development of an ontology includes defining a set of data and its structure so that applications can use that knowledge to investigate the data. Therefore, competency questions also provide a framework for evaluating different ontological approaches for the same requirements.

An ontology is not an application. It is tool for specifying semantics and defining formal logic-based knowledge models. In Section 3.1, I showed how RDF can be used to model information in a graph data structure. OWL, an ontology language and another component of the Semantic Web, provides the features for utilizing and interpreting OWL semantics (McGuinness and van Harmelen, 2004). OWL adds restrictions to the content and structure of RDF graphs, thus allowing processing for computationally decidable reasoning. To use these computational capabilities, some type of framework for storage and retrieval of information, and for the logical interpretation of the ontology is needed. The Semantic Web framework is a collection of integrated tools and technologies that provide the ability to create and work with a knowledge base.

3.3.4 Knowledge base

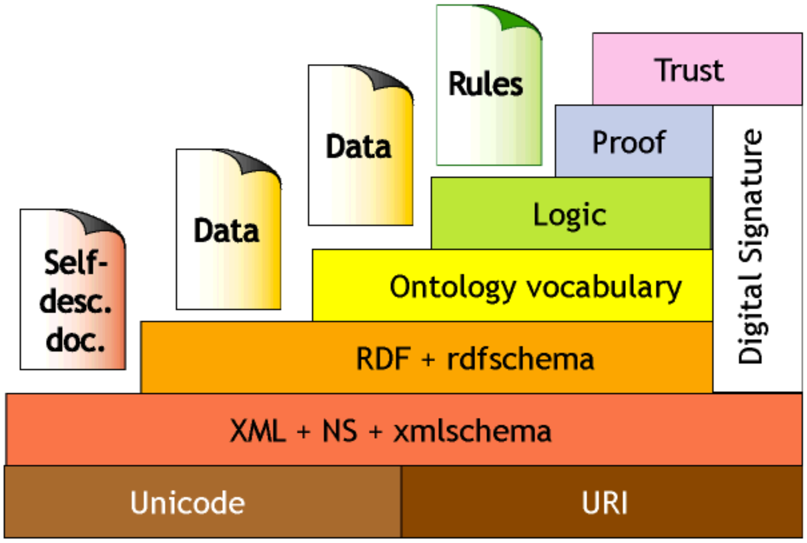
At the core of knowledge-based systems is the knowledge base. The knowledge base system is typically a set of software components that provides the ability to create a collection of information and to describe that information ontologically. Such an application framework provides the functionality to create, describe, process and make inference over information in the knowledge base. In this sense, the knowledge base is the capability of what several integrated technologies allow a user to achieve. For example, to use RDF for data modeling and OWL for defining knowledge models, some real-world application must implement these technological specifications to provide users with tools to utilize these capabilities. One such framework, and the one used in this work, is the Semantic Web framework.

Tim Berners-Lee and colleagues coined the term Semantic Web and gave a vision to a "web of data" (Berners-Lee et al., 2001). This vision is motivated by the fact that the Web has evolved mainly as HTML webpages that publish information for human consumption: HTML markup displays content that is interpretable by a Web browser, yet the inherent meaning of content in webpages is not interpretable by computers because they lack rich machine-readable metadata that machines can exploit. Thus the goal of the Semantic Web vision is to make possible the processing of information published on the Web by computers (Cardoso and Sheth, 2006). Figure 3.20 shows Berners-Lee's illustration of the Semantic Web stack – the hierarchy of languages and technologies utilized in the formation of the Semantic Web.[81]

In Semantic Web architecture, an application framework stores data in the knowledge base, performs inference, and provides query endpoints and an application programming

[81] This illustration is taken from a talk by Berners-Lee available at: `http://www.w3.org/2000/Talks/1206-xml2k-tbl/slide10-0.html`. Note that the architecture of the Semantic Web stack is evolving as layers are formalized (Horrocks et al., 2005; Kifer et al., 2005).

Figure 3.20: Semantic Web stack



interface for data retrieval. The knowledge base interacts with aggregated data sources and performs the logic inference of the domain-model ontological reasoning. Ontologies are the digital architecture that provide interoperable semantics of metadata. These semantics are machine interpretable because by following certain standards in creating the data, tools that “understand” (access and inference) the data can link data from disparate sets if they share any common node. A node in the Semantic Web, i.e. a concept, individual or class, is a Uniform Resource Identifier (URI).

The Semantic Web is built in layers. Triples are built with URIs that define the subject, predicate and object of a statement. Each triple/statement describes a fact. The subject and predicate are defined with a URI. The object of the statement can be either a URI or some other definable data type, such as a string literal or an integer. The URI is a key feature in the overall architecture because each provides a unique identifier within a global namespace. Since triples are built with URIs, they can be easily merged from many different sources via common URIs or defining of relationships between URIs via additional triples. In Example 3.35 the ISO 639-3 code [nob] for Bokmål Norwegian is a subclass of the (macrolanguage) ISO 639-3 code [nor] for Standard Norwegian.[82] In Example 3.36, the segment used to represent Maddieson's voiced alveolar flap is treated as the same segment used for the voiced alveolar flap. The URI is a key design feature because it provides the mechanism for global naming and connects each resource in the statement to a Web resource (through a process called content negotiation, the URI might resolve to a human-readable webpage or machine-readable data).

(3.35)
```
<rdf:Description
  rdf:about="http://phoible.org/id/iso639-3/nob">
<rdfs:subClassOf
  rdf:resource="http://phoible.org/id/iso639-3/nor"/>
```

(3.36)
```
<rdf:Description rdf:about="http://phoible.org/segment/ᴅ">
<owl:sameAs rdf:resource="http://phoible.org/segment/ɾ"/>
```

RDF expresses information in triples, i.e. in the form of subject-predicate-object statements. RDFS (RDF Schema) is an additional ontology language built on top of RDF that can be used to define simple class types and relations in an RDF graph. For example, with RDFS, the *subClassOf* relation can be used to define inheritance between subjects and/or objects. OWL is another, more powerful ontology language that builds on RDF's structure and it adds more logic relations to RDF graphs by defining restrictions that include equivalency, transitivity, cardinality, etc. Together, RDF, RDFS and OWL can be used to create knowledge bases that can be logically evaluated; the truth conditions of statements in the RDF graph can be verified within a finite amount of time. The knowledge base's assumptions are encoded in an ontological theory and their statements can be processed by automated reasoning tools to infer new knowledge; thus generating new information that can then be added back to the RDF/OWL knowledge base. Automated reasoning tools can be used to prove the consistency of information encoded in the knowledge base or to enhance search via the addition of implicit knowledge derived from explicit ontological statements through logical inference.

[82] Currently there are no standard ISO 639-3 URIs.

3.3.5 Summary

Knowledge representation is “the application of logic and ontology to the task of constructing computable models for some domain” (Sowa, 2000, xii). It is a multidisciplinary field that leverages theory and techniques from logic, ontology and computation. Mathematical formalization of these areas, as well as probability, helped AI to make the leap from ideas originally explored by philosophers in antiquity to modern day information science. Logic provides the formal structure for knowledge representation and the rules for inference. The logic assumptions in a knowledge base are captured in an ontological theory. In this section I have given an overview of knowledge representation and given a description of how the representation of knowledge can be implemented in Semantic Web technologies like RDF and OWL.

3.4 Conclusion

In this chapter I provided a brief overview of modeling data in different formats and I discussed aspects of knowledge representation. In Section 3.1 I provided some data modeling basics. In Section 3.2 I described the PHOIBLE data models in detail and showed how users can query the different data model instantiations. In Section 3.3 I discussed the details of knowledge representation. I focused on knowledge representation within the Semantic Web framework, which uses RDF graph data structures and Description Logics formalized in OWL to create knowledge bases. In Section 8.4.6 in Chapter 8 I will describe future work with RDF/OWL and the PHOIBLE data to create *Linked Data*. Linked Data is a recommended best practice for describing and marking up resources for sharing and connecting information and knowledge on the Web.

Chapter 4

PHOIBLE

4.1 Introduction

PHOIBLE is an online repository of cross-linguistic phonological segment inventory data that contains additional linguistic and non-linguistic information about languages.[1] It is a convenience sample that includes phonological segment inventories for 1089 of the world's 6909 known living languages, so roughly 16%.[2] Additional linguistic information linked to each segment inventory includes language family information (language stock via Ethnologue and genus via WALS) and each segment is linked to a set of distinctive features. Non-linguistic data linked to the segment inventories includes population figures, geographic location (world region, predominate country where the language is spoken and geo-coordinates) and per-capita GDP by country.[3]

The amount of detail for each segment inventory ranges from phonemic descriptions to descriptions of phonemes, their allophones and their phonological environments. This is because PHOIBLE subsumes the segment inventory databases from the Stanford Phonology Archive (SPA; Crothers et al. 1979), the UCLA Segment Inventory Database (UPSID; Maddieson 1984; Maddieson and Precoda 1990), *Alphabets des langues africaines* (AA; Hartell 1993; Chanard 2006), and an additional 485 "PHOIBLE inventories" that were gathered because they were not previously included in these databases. All segment data in PHOIBLE were standardized and compiled into a single data repository through a process commonly called Extract, Transform and Load (ETL) (Inmon, 1992; Kimball, 1996). The SPA, UP-

[1] Figure 1.1 on page 3 provides an illustration of PHOIBLE's contents. PHOIBLE is available online at `http://phoible.org`.

[2] This figure is based on the Ethnologue 16th edition (Lewis, 2009).

[3] Population figures, geographic areas and countries where languages are predominately spoken are from the Ethnologue (Lewis, 2009). Geo-coordinates are from WALS (Haspelmath et al., 2008) and GDP figures are from the Central Intelligence Agency's World Factbook (Central Intelligence Agency, 2010).

SID, AA and PHOIBLE inventories each underwent an individualized ETL process because each data source provided its own set of challenges in forming a unified, all-Unicode IPA data repository. Additional linguistic and non-linguistic data were added to the PHOIBLE database via tables and associated with segment inventories via their ISO 639-3 codes, so that these data can be easily updated in future releases.

This chapter is set up as follows. In Section 4.2, I discuss the motivation behind creating PHOIBLE. I explain the ETL processes and challenges faced in merging its disparate data sets in Section 4.3. And in Section 4.4 I describe PHOIBLE's genealogical coverage.

4.2 Motivation

Since the 1970s, investigations into phonological universals have been undertaken using cross-linguistic segment inventory data sets. This work began with SPA (Crothers et al., 1979). The compilers of SPA gathered detailed segment inventories to test and make claims of phonological universals and to provide statistics on the distribution of phonological segments in the world's languages. SPA was the predecessor to Maddieson's UPSID databases.

More than twenty years after UPSID$_{451}$ was made publicly available, it remains the standard reference sample for research on segment inventories and phonological universals. A small but representative sample of publications that use UPSID's segment inventory data include: *Segmental Complexity and the Structure of Inventories* (Rice and Avery, 1993), *Differentiating 451 Languages in Terms of their Segment Inventories* (Pericliev and Valdés-Pérez, 2002), *On the back of the tongue: Dorsal Sounds in Australian Languages* (Butcher and Tabain, 2004), *Modeling the Co-occurrence Principles of the Consonant Inventories: A Complex Network Approach* (Mukherjee et al., 2008), *Areal-typological Constraints on Consonant Place Harmony Systems* (Kochetov et al., 2008), *Universals in Phonology* (Hyman, 2008), and *The Role of Features in Phonological Inventories* (Clements, 2009).[4]

Hyman (2008, 94) provides an excellent example of why access to a broader set of segment inventories is desirable. Based on the UPSID$_{451}$ sample, he postulates "Consonantal

[4] There are well over 1000 published articles that cite or use data from Maddieson 1984 and Maddieson and Precoda 1990.

Universal #4: Every phonological system has coronal phonemes".[5] In little time, Blevins (2009) refuted this phonological universal and argued that Northwest Meeko [mek-nws][6] lacks coronal phonemes; they are described as predictable allophones of velars.

In this case, it seemed to me that the solution to the problem of making claims about phonological universals (or generalizations about segment inventories) lies in broad access to current research on the phonologies of the world's languages. Surely, collecting the analyses of phonological inventories for all documented and described languages and making them available on the Web is within today's technological grasp (though there remain many challenges as discussed in Section 2.3). This is one motivation that has driven the development of PHOIBLE.

Another motivation that has driven development is to create an extensible, transparent and interoperable repository of data that is openly available to research communities. This goal has been influenced by my work on the National Science Foundation funded E-MELD project.[7] One aim of the E-MELD project was to develop technological infrastructure to preserve and share data from the digital documentation of (endangered) languages. There is a growing research community with goals shared by the E-MELD vision working towards a *cyberinfrastructure* (called *e-Science* in European initiatives) designed to support multidisciplinary scientific research. Cyberinfrastructure is the convergence of computing, digital standards, information management and a cultural shift that supports the sharing of data. In linguistics, this is increasingly important as both languages and language documentation are at risk of endangerment and extinction. A well-established cyberinfrastructure in linguistics requires digital architecture and adherence to standards to ensure that data from a variety of resources is accessible and interoperable (Bender and Langendoen, 2010). In the next section, I describe the challenges of merging legacy databases and new data sets into a single interoperable data repository.

[5] This universal echoes the finding made in Maddieson 1991 that all languages in the UPSID sample have at least one coronal consonant.

[6] This code comes from an extended version of the ISO 639-3 language identification codes that contains dialect information provided by Multitree: `http://multitree.linguistlist.org`. Mekeo [mek] has four variants: North [mek-nor], Northwest [mek-nws], East [mek-eas] and West Mekeo [mek-wes].

[7] `http://emeld.org`

4.3 *Extract, transform, load*

Combining disparate data sets is problematic and challenging. Integrating segment inventories from many different resources into one interoperable data set posed two main challenges. The first and simpler challenge involved adding metadata to each record. How can each segment inventory be identified with information about its origin (language, bibliographic reference, database of origin) so that inventories can be indexed and compared? My starting point has been to identify each resource from which a segment inventory was extracted with an ISO 639-3 unique language identifier. ISO 639-3 codes, however, are not enough for a database that contains different analyses and dialect descriptions for the same language. For example, there are many different segment inventories of varieties of English that all fall under one language code [eng].[8] Therefore, each segment inventory in PHOIBLE is given a unique identifier and is associated with bibliographic metadata.[9]

The second and more complex challenge is both linguistic and technological. How can the segments in segment inventories, which are typically idiosyncratic in their transcription, be brought to a level where they can be compared linguistically and computationally? The first step is to interpret the segments in transcription systems into IPA, PHOIBLE's interlingual pivot. This involves reading linguists' phonological descriptions and interpreting their analyses.[10] Re-encoding segments and phonetic descriptions into IPA brings up the issue of diacritic ordering. To my knowledge, the IPA does not define an explicit ordering scheme for diacritics. Instead, linguists tend to use an order implicitly based on speech production and expressed through timing units that encode the acoustic sequence of sounds as they occur in a temporal and linear order, e.g. /p^{h}/ and /k^{w}'/. Moreover, when diacritics appear above and below the segment, the order is no longer visually distinguishable, e.g. a nasalized creaky vowel, /ã̰/.[11] To the linguist this probably poses little problem and I

[8] See Section 2.3.4 for discussion.

[9] These details are discussed in Section 3.2.

[10] For an overview of the challenges involved, including interpreting segments and phonetic descriptions into IPA, see Sections 2.3.3-2.3.5. For segment-specific issues that I encountered in each data source in PHOIBLE, see Sections 4.3.1-4.3.4 below.

[11] In Section 4.3.4, I briefly discuss some of the choices that I have made regarding segment and diacritic

imagine most linguists give little thought to whether they should first click on the nasalization diacritic and then the creaky voice diacritic, or vice versa, when creating segments through an IPA picker or another input method. To the computer there are two series of character combinations that can be rendered for <ã̰>, as shown in Table 4.1. For segments with more diacritics, the combinations are essentially n-factorial, where n is the number of diacritics (although as mentioned, many diacritics that following the base segment have a linguistically-implicit ordering). Diacritic ordering is a critical issue. If <ã̰> and <ã̰> aren't ordered the same computationally, they are literally two different (sequences of) characters, even though visually they are homoglyphs. Of course the example in Table 4.1 is just one of many different possible homoglyphs that can be created with Unicode IPA characters and diacritics.

Table 4.1: Rendering sequences of Unicode characters as segments

ã̰	ã̰
U+0061 + U+0330 + U+0303	U+0061 + U+0303 + U+0330
LATIN SMALL LETTER A + COMBINING TILDE BELOW + COMBINING TILDE	LATIN SMALL LETTER A + COMBINING TILDE + COMBINING TILDE BELOW

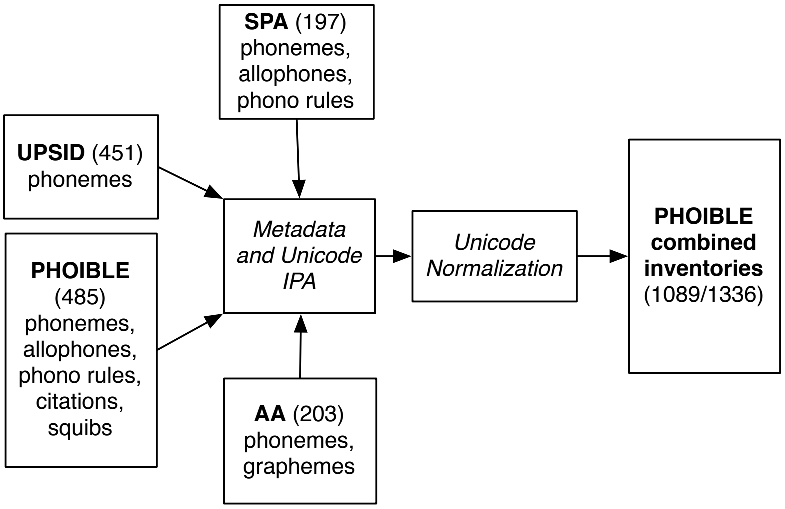
After mapping all segment types to IPA so that segment inventories can be compared linguistically, the second challenge is making the segments interoperable computationally. This issue is addressed by using Unicode normalization to decompose each segment type into an algorithmically determined sequence of characters.[12] Unicode defines the order of normalization forms, thus assuring that equivalent strings will have the same binary representation.

ordering. A full account of my decisions is given in Appendix C.

[12] The details of Unicode normalization forms are quite complex. Refer to Unicode Standard Annex #15 for details: `http://www.unicode.org/reports/tr15/`.

Figure 4.1 gives a high-level illustration of the implementation of PHOIBLE. In the next four sections, I explain the individual ETL approaches for SPA, UPSID$_{451}$, AA and PHOIBLE inventories.

Figure 4.1: Implementation of PHOIBLE


SPA (197) phonemes, allophones, phono rules
UPSID (451) phonemes
PHOIBLE (485) phonemes, allophones, phono rules, citations, squibs
Metadata and Unicode IPA
Unicode Normalization
PHOIBLE combined inventories (1089/1336)
AA (203) phonemes, graphemes

4.3.1 SPA

The first computerized database of phonological segment inventories is the Stanford Phonology Archive (SPA; Crothers et al. 1979). SPA was inspired by Joseph Greenberg's research on universals and his personal archive of data from notebooks and his memory (Crothers et al., 1979, i-ii). The utility of a computerized archive was clear: a device for scholars wishing to ask questions about universals, but who did not have access to data like Greenberg's paper records and his knowledge of languages. The aim of the archive was to develop machine-searchable files so that researchers could look for patterns, examples and evidence of phonological universals (Sherman and Vihman, 1972). SPA was produced by the Stanford Language Universals Project (1967-1971) and its segment inventories include descriptions of phonemes, allophones and comments on phonological contexts for 197 different languages.[13]

[13] A sample of an inventory printed in the *Handbook of Phonological Data From a Sample of the World's Languages: A Report of the Stanford Phonology Archive* is provided in Figure 4.4 on page 178.

SPA brought together in one place detailed segment inventory data from a "carefully selected sample of the world's languages" (Crothers et al., 1979, i). These data were collected independently of the questions that SPA intended to answer, with the intention of providing a valid data sample. SPA aimed to provide a balanced representation of diverse language families and geographical areas (Sherman and Vihman, 1972). The SPA sample included the eleven most commonly spoken languages within its 200 language sample. The project's intent was to provide a resource to support or refute cross-linguistic hypotheses in phonetics, phonemic systems, phonotactic constraints and phonological processes (Vihman, 1974). However, creating an unbiased sample of languages to test phonetic and phonological hypotheses is difficult.[14] Using SPA, Sherman (1975, 3) may have been the first to raise the issue of how to create a representative language sample that "that adequately and proportionately represent[s] areal, genetic and typological diversity of the languages of the world". How to create a statistically unbiased sample of cross-linguistic data is still an area of intense debate.[15] The intent today remains the same as then: to make statistically valid generalizations over incomplete data sets.

SPA's developers raised two important questions that remain relevant to typological database projects today. The first asked, "what constitute[s] adequate descriptive categories for linguistic phenomena?". And the second, "what are appropriate media and formats for storing, controlling, and accessing descriptive linguistic data?" (Sherman and Vihman, 1972, 163).

The first question addresses the issue of creating a comparable data set. Several problems ensue from this question. For example, how does a cross-linguistic resource provide an unbiased set of data when each resource is an idiosyncratic description of a field linguist's observations? Language descriptions by different researchers do not include the exact same observations because they are impressionistic accounts. The shortcomings of extracting phonological descriptions from published sources was apparent early on. Different terminologies and different theoretical approaches posed problems of interpretation for maximal

[14] See discussion in Section 2.3.

[15] See Section 2.3.2.

interoperability of comparable data sets (Crothers et al., 1979).

The second question asked by SPA's developers is technological and remains relevant today (arguably even more so with the increasing variety of digital formats and recording media). What is the appropriate format for creating an accessible data repository for long-term archiving? SPA provides us with a historical example. In the SPA Handbook's forward, Charles Ferguson writes, "Also, we had hoped that the Archive would become widely accessible both through a continuing Archive unit at Stanford and through the use of tapes at other universities and centers of language research. As of this writing (May 1979), it seems that Stanford archive retrieval services will be severely curtailed and that the University of California at Berkeley is the only other place where a copy of the Archive tapes is available and in regular use for phonological research." (Crothers et al., 1979, vi).

Although SPA was novel in its technological approach, the archive was never really usable on a computer and the grant was cut before the project could finish all it intended to accomplish (Scott Drellishak via Marilyn Vihman, p.c.). Unfortunately, the immense work that went into creating the computerized version of SPA became largely obsolete and later inaccessible to researchers.[16,17]

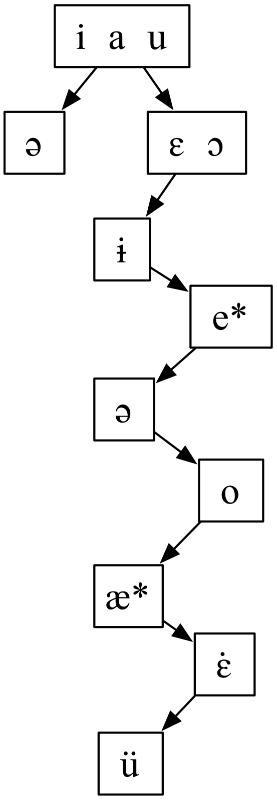
In a review of *Universals of Language II* (Greenberg et al., 1978), a volume devoted to topics in phonological universals and based on work with SPA and the Stanford Universals Project, Javkin (1980, 830) states, "[SPA] can be expected to change substantially the course of research in phonological universals." Crothers (1978) took full advantage of utilizing SPA by describing typological universals of vowel systems. Crothers's claims included the observation that 98.5% of languages in the SPA sample have the vowels /i a u/. He also included a dispersion model (an implicational hierarchy) of proposed vowel universals, reproduced here in Figure 4.2.[18]

[16] However, much of SPA's content was used in UPSID, which was later made publicly available.

[17] Several years ago we attempted to retrieve the SPA data by contacting the Linguistics Department at Stanford University. However, they reported that the data were no longer available from the Phonology Archiving Project. Fortunately, Marilyn Vihman, an author of the handbook and publications editor for SPA, had a printout of the massive 900 page resource, which she kindly mailed to the University of Washington so we could digitize it.

[18] Vowels marked with * can be interchanged. The segments <ü> and <ɛ̇> represent a high front rounded vowel and a lower-mid central unrounded vowel, respectively (Crothers, 1978, 137).

Figure 4.2: Vowel hierarchy based on inventories in SPA (Crothers, 1978, 133)



The ability to query a segment inventory database for evidence and counter-evidence indeed provided a new avenue for research in investigating phonological universals and the cross-linguistic frequency of linguistic phenomena like segments. The utility of SPA for cross-linguistic research on language universals was clear and inspired much future work in the field, including Maddieson's UPSID database, which has become the reference standard for investigating the nature of speech sound inventories (discussed in the next section).

The SPA sample contains phonemes, allophones and a description of their phonological environments for 197 distinct languages. To extract the inventory data from SPA, the paper copy was scanned into PDF and its contents digitized by hand into an Excel spreadsheet.[19] After digitizing SPA and mapping its segment descriptions into IPA, the segment inventory data were transformed via a Python script into an intermediate CSV format that contains segments in Unicode IPA and metadata for each inventory.[20] These data were then written to an XML file and imported into PHOIBLE's MySQL relational database. The ETL process that transformed the SPA Handbook into interoperable segment inventories is illustrated in Figure 4.3.

Figure 4.3: Stanford Phonology Archive conversion process

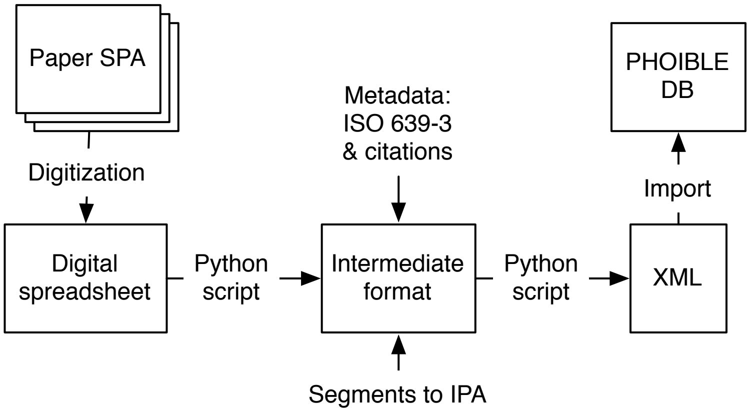


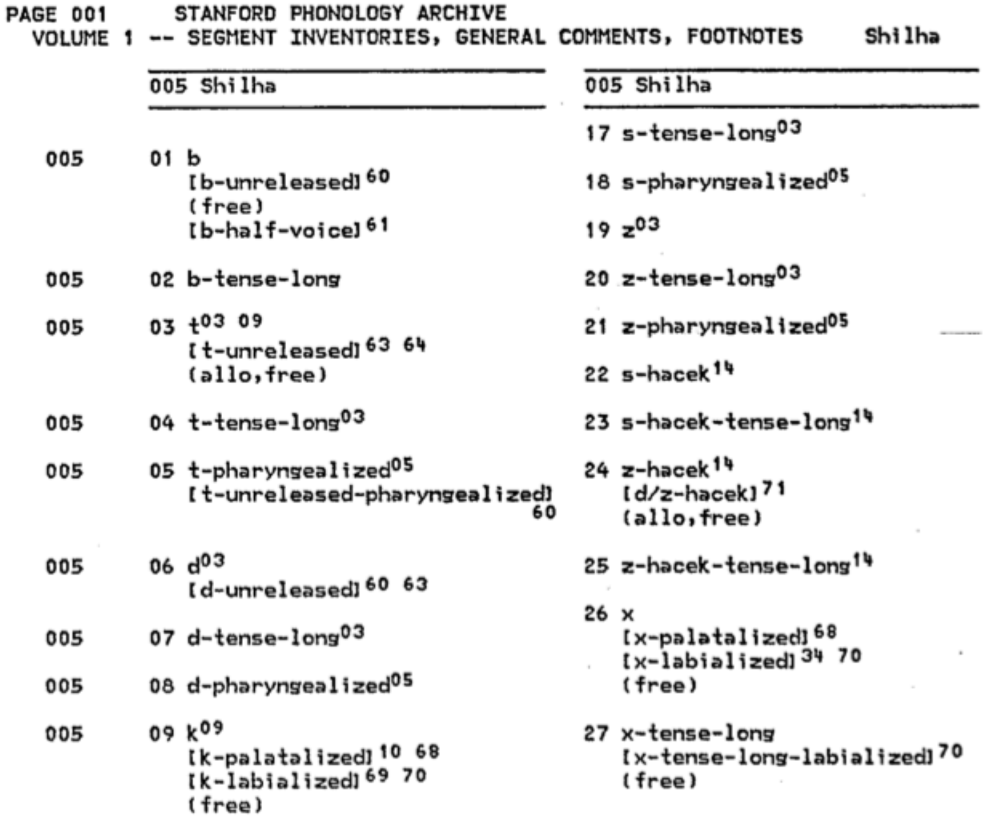
Figure 4.4 shows a portion of the segment inventory for Shilha [rif]. In SPA there are 1545 segment types encoded in written descriptions like "d-pharyngealized". Each phoneme is numbered (to its left) and its allophones are provided below it in square brackets. Phonemes and allophones may be followed with a numeric code in superscript; they are associated with notes provided after the inventory. A full description of segments and codes used in SPA is given in Sherman and Vihman 1972 and Crothers et al. 1979.[21]

[19] When I started the project, my attempts with various Optical Character Recognition (OCR) software programs were fruitless. The digitization process was started by Scott Drellishak and continued by Michael McAuliffe. To avoid typos in the digitization, Drellishak set Excel's to autocorrect input, e.g. when "/g" was keyed in, it was automatically replaced with the correct Unicode IPA <g> LATIN SMALL LETTER SCRIPT G at U+0261.

[20] This process allowed me to keep separate the original data that were digitized into a spreadsheet, the SPA-to-Unicode IPA mappings and the transformed version of SPA that includes an ISO 639-3 code for each inventory. This modular process allows me to update, say, a particular SPA segment description's IPA rendering, and then the conversion pipeline can be easily rerun to update the PHOIBLE database.

[21] In at least one case, there appears to be a typo in the original SPA data set. The typo appears in

Figure 4.4: Segment inventory for Shilha (Tarifit [rif]) from SPA



```
PAGE 001      STANFORD PHONOLOGY ARCHIVE
  VOLUME 1 -- SEGMENT INVENTORIES, GENERAL COMMENTS, FOOTNOTES     Shilha

            005 Shilha

005         01 b
               [b-unreleased] [60]
               (free)
               [b-half-voice] [61]

005         02 b-tense-long

005         03 t[03] [09]
               [t-unreleased] [63] [64]
               (allo,free)

005         04 t-tense-long[03]

005         05 t-pharyngealized[05]
               [t-unreleased-pharyngealized]
                                          [60]

005         06 d[03]
               [d-unreleased] [60] [63]

005         07 d-tense-long[03]

005         08 d-pharyngealized[05]

005         09 k[09]
               [k-palatalized] [10] [68]
               [k-labialized] [69] [70]
               (free)

            005 Shilha

            17 s-tense-long[03]

            18 s-pharyngealized[05]

            19 z[03]

            20 z-tense-long[03]

            21 z-pharyngealized[05]

            22 s-hacek[14]

            23 s-hacek-tense-long[14]

            24 z-hacek[14]
               [d/z-hacek] [71]
               (allo,free)

            25 z-hacek-tense-long[14]

            26 x
               [x-palatalized] [68]
               [x-labialized] [34] [70]
               (free)

            27 x-tense-long
               [x-tense-long-labialized] [70]
               (free)
```

To make the segment types interoperable with segment inventory data from other sources, each SPA segment description was interpreted into a Unicode IPA representation (see Appendix E).[22] For the most part, these mappings were straightforward. A few examples are provided in Table 4.2.

In some cases, however, mapping a SPA written description to an IPA representation was problematic. For example, there is no IPA diacritic to represent "half voice" in "ash-half-

the segment inventory for the language Ga [gaa], record number 095. Reportedly in Ga, "All vowels are somewhat nasalized in the environment of nasal consonants" (SPA citation: Berry, J. n.d. The Pronunciation of Ga. Cambridge, Eng.: Heffer informants). Whereas the other vowels have "nasalized-weak" allophones (e.g. "u" and "u-nasalized-weak") the "a-front-nasalized-weak" is listed as phoneme, although it should be listed as an allophone of "a-front-nasalized" in Crothers et al. 1979, 52.

[22] Michael McAuliffe undertook the initial pass through the segments and then changes and corrections were made by Richard Wright, Dan McCloy and myself.

Table 4.2: Examples of SPA and Unicode IPA correspondences

SPA	IPA
x-uvular-tense-labialized	χʷ
t/s-hacek-preglottalized	ʔtʃ
t/c-fricative-aspirated-labialized	tçʷʰ

voice-long" and there is no weak nasalization diacritic for segments like "a-nasalized-weak".[23] These cases are typified by segment types used for allophonic distinctions. More problematic are sets of phonemically contrastive features in SPA that have no IPA representation, e.g. "tense" (fortis) and "lax" (lenis) consonants like "x-uvular-tense" and "x-uvular-lax". Collapsing these features (or simply ignoring them) because they do not exist in IPA is not ideal. In some languages descriptions in SPA, like Oneida [one], the lack of the tense feature would collapse an allophonic distinction. In other language descriptions like Lak [lbe] or Sa'ban [snv], however, a lot of phonemic contrasts would be lost (7 and 6, respectively).[24] This would reverberate in the number of phonemes used for statistical calculations and other possible analyses. Instead I had to violate pure IPA and chose to use diacritics to mark tenseness or laxness of consonants, regardless of the consistency in which they are used by researchers across language descriptions.[25] For example, for the tense consonants, I chose to use the "strong articulation" diacritic from the "extensions of the IPA" Unicode block at U+0348, combining double vertical line below. This symbol has been used in the literature and at this time seems to be a decent choice. These decisions are noted

[23] One approach to represent partial devoicing is to use the combination of a voiceless diacritic and a tie bar /o̥͡o/ (Hayes, 2009).

[24] See for example Table 2.8 on page 67.

[25] Ladefoged and Maddieson (1996, 95) describe the diverse meanings in which the terms fortis (tense) and lenis (lax) have been used as phonological labels in the linguistic literature.

with the segment correspondences in Appendix E.[26]

After the segment inventory data were digitized and the segments assigned IPA representations, each inventory was identified with an ISO 639-3 language code to make SPA's segment inventories compatible with other segment inventory databases' inventories. In several cases the language name provided in SPA is now a group of related languages. Two examples from SPA are provided in Table 4.3, which shows macrolanguage codes for Haida and Objibwa and their ISO 639-3 language codes. The term macrolanguage was introduced in the Ethnologue 16th edition to cover a set of closely related languages, or significantly different dialects.[27] For certain inventories in SPA, without expert knowledge it is difficult to identify the now more specific language variant that was originally documented in its broader sense. When the publication was identified with a specific code by WALS or the Ethnologue, or both, I used that code.[28] When neither resource referenced the publication, the original documentation was consulted. In some cases I could identify the language from information within the original documentation, e.g. indication of a particular dialect now considered a distinct language or by the geographic description of where the language is spoken. In other cases I am still seeking more verification by consulting other sources and by contacting experts in these languages. In some cases I have simply used the ISO 639-3 marco language code for the time being (Akan [aka] is one example). A list of language names, ISO 639-3 language name identifiers and bibliographic citations for each inventory in PHOIBLE is provided in Appendix B.

A final note about SPA is in regard to its contents. I have not gone through each inventory and verified from the original sources if the contents in the SPA Handbook match

[26] These correspondences can be easily updated and I welcome suggestions and community consensus on how segments like "half-voice-long" should be represented in IPA.

[27] There are two other situations for using macrolanguage codes (see: `http://www.sil.org/iso639-3/scope.asp`). One uses a standard variety as a cover-term for two or more languages. For example, a "Standard Arabic" is generally used by speakers from many distinct Arabic languages. The macrolanguage code [ara] is therefore used as a cover code for 30 or so distinct Arabic languages, e.g. Omani Arabic [acx], Saidi Arabic [aec], Moroccan Arabic [ary], etc. The other uses a macrolanguage code when subcommunities of a single language are diverging. For example, Serbo-Croatian [hbs] is a macrolanguage code for Bosnian [bos], Croatian [hrv] and Serbian [srp]. In this case, both communities and linguistic varieties are diverging; communities are trying to make their variety different from neighboring ones (Jelena Prokić, p.c.).

[28] The Ethnologue and WALS sometimes disagree on which code is assigned to which language. An example and discussion is given in Section 4.3.2.

Table 4.3: Example macrolanguages in SPA

Macrolanguage	Languages
Haida [hai]	Northern Haida [hdn] (Canada)
	Southern Haida [hax] (Canada)
Objibwa [ojg]	Chippewa [ciw] (United States)
	Ojibwa, Central [ojc] (Canada)
	Ojibwa, Eastern [ojg] (Canada)
	Ojibwa, Northwestern [ojb] (Canada)
	Ojibwa, Severn [ojs] (Canada)
	Ojibwa, Western [ojw] (Canada)
	Ottawa [otw] (Canada)

precisely to the original descriptions.[29] One example that I encountered is SPA's description of Ticuna [tca], an isolate spoken in Brazil. The inventory description, taken from Anderson 1959, lists nine tonemes: high, higher-mid, mid, lower-mid, low, high-falling, higher-mid-falling-mid, higher-mid-falling-low, and mid-falling. However, a review of the segment inventory by John Crothers (JHC) contains these remarks (Crothers et al., 1979, 949):

1. lower-mid – "Although the Andersons regard /high/ and /lower-mid/ as distinct tonemes, they probably are not phonemically different from the /higher-mid/ and /low/ tonemes respectively. [JHC]"

2. high-falling – "/high-falling/ occurs infrequently, mostly on bound pronominal morphemes. Undoubtedly not a distinct phoneme. [JHC]"

[29] Notes from the compilers appear in the handbook with each inventory, but these notes have not yet been entirely digitized.

3. higher-mid-falling-mid and mid-falling – "/higher-mid-falling-mid/ tone and the /mid-falling/ tone are the only falling tones which occur with any frequency. Some bound pronominal morphemes seem to alternate between the two tones. It cannot be considered certain that these two falling tones contrast. [JHC]"

4. higher-mid-falling-low – "/higher-mid-falling-low/ occurs infrequently, mostly on bound pronominal morphemes. Undoubtedly not a distinct phoneme. [JHC]"

I have not changed SPA's published segment inventory contents to reflect Crothers's observations because this would go against my methodology of keeping the data faithful to its original source, and then letting users manipulate the data set's contents for their own purposes. The way forward is to add new inventories by addressing Crothers's concerns by going through the original materials, as well as more recent publications on these languages (if they exist). These inventories are then added to PHOIBLE and users can specify which of the alternate inventories they wish to sample. Cases like Ticuna, like other cases in which errors on my part were introduced via the ETL process (and later corrected), have revealed themselves through simple statistical analysis of the data sets, e.g. Ticuna is an outlier for the number of tones represented in its inventory.[30]

In general, tone is a problematic area for phonemic analysis because tones exist on a suprasegmental level and can interact with the morphosyntax in complex ways.[31] Furthermore, many questions are raised in an analysis of tones, especially when entering a segment inventory into a database. For example, one must consider whether the language is strictly a register tone language with some underlying number of tones, or if the language uses tonal melodies whose ordering is contrastive. In my experience, some authors list a downstepped H as a contrastive phoneme, while analytically it may be an allophone of high tone. Yet in another description, the author lists high and low tones as "grammatical function only", thus implying that they are not lexically contrastive. Some researchers, including Nettle

[30] See Chapter 5.

[31] Current formal approaches lack machinery in describing complex tone systems like Dogon tonosyntax (Heath and McPherson, submitted).

(1995, 361), count each permitted combination of vowel and tone separately by multiplying the number of vowel phonemes by the number of tones or contrastive lengths. These examples show that tone is a problematic area for phonemic analysis and when creating typological data sets.

To summarize, in this section I gave a brief overview of SPA, its contents and some of the research questions that its compilers raised as the first "computerized" segment inventory database for phonological typology. Then I described the conversion process that I developed to transform the SPA data from a printed paper resource into a digital format. I discussed the challenges in making the SPA data interoperable with other segment inventory databases, which involved identifying ISO 639-3 language name identifiers for each inventory and transforming SPA's segment descriptions into Unicode IPA. Once these data were made interoperable, I imported the SPA data into the PHOIBLE database.

4.3.2 UPSID

In the early 1980's, Maddieson developed the UCLA Phonology Segment Inventory Database (UPSID), a computer-accessible database of contrastive segment inventories. The initial sample of 317 languages drew from 192 of SPA's inventories. However, changes were made. As noted in Maddieson 1984, 6: "Our decisions on the phonemic status and phonetic description do not always coincide with the decisions reached by the compilers of the SPA and we have sometimes examined additional or alternative sources, but a great deal of effort was saved by the availability of this source of standardized analyses." More than just increasing the number of inventories in SPA, Maddieson implemented a quota sample that aimed to include only one language from each small language family to create a typologically diverse and genealogically balanced sample of languages. The intent of the quota sample was to provide statistically valid generalizations of the world's languages for surveying segment type frequencies and patterns of their occurrence and co-occurrence. The results were published in Maddieson's (1984) influential book, *Patterns of Sounds*.

In 1990, UPSID$_{317}$ was expanded to include 451 segment inventories, roughly 6.5% of the world's languages (Maddieson and Precoda, 1990). The entire set of sources used in

UPSID$_{317}$ were re-examined, additional resources consulted, and some errors in the language inventories were corrected (Maddieson and Precoda, 1990, 104). UPSID$_{451}$ became the first widely used computer database of segment inventories. Indeed much of what is currently known about segment inventories and segment frequencies is based on UPSID$_{451}$.

UPSID$_{451}$ was designed to make possible statistically valid generalizations about segmental occurrences and co-occurrences about living languages (Maddieson and Precoda, 1990). The data were made available through a DOS software package that allowed users to "count or select and output to a file the particular subset of data that is crucial to the questions they want to address" (Maddieson and Precoda, 1990, 109).[32] It was noted that the computer program was relatively simple and for advanced analyses the data should be output and used as input in a statistical software package. There are also at least two other ways to access UPSID. Reetz (2005) developed and put online an HTML interface to the UPSID$_{451}$ data.[33] The website provides access to each segment inventory (including its contents and bibliographic citations), basic descriptive statistics regarding the frequency of segments and a search interface. There is also a Prolog interface to the UPSID$_{317}$ data developed by Ron Brasington.[34]

UPSID$_{451}$ contains phonemes and their featural descriptions for 451 distinct languages. PHOIBLE uses the publicly available DOS files.[35] Although the segment inventories did not need to be digitized by hand like SPA, the segment data and corresponding metadata had to be extracted from now old DOS files. The data were initially converted into a Microsoft Access database for another research project at UW.[36] The Access version was exported into Microsoft Excel spreadsheets. The UPSID$_{451}$ data tables were originally designed in a relational database fashion. Thus it was easy to import them directly into relational

[32] http://www.linguistics.ucla.edu/faciliti/sales/software.htm

[33] http://web.phonetik.uni-frankfurt.de/upsid.html

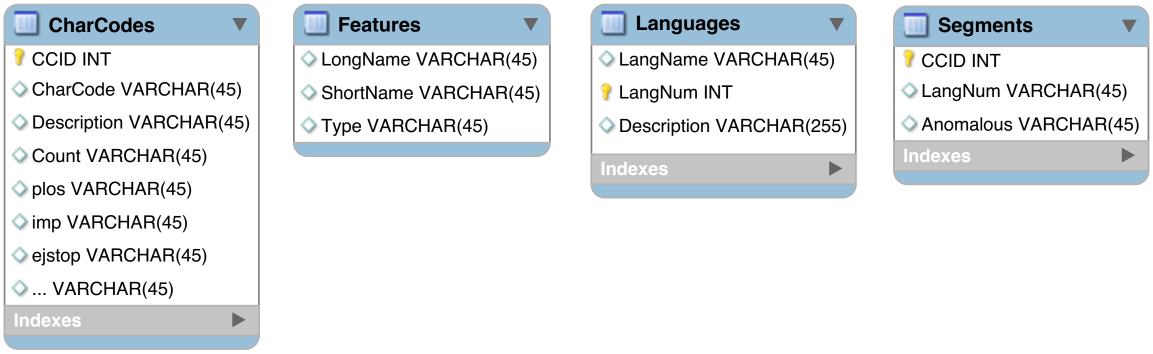
[34] http://www.personal.rdg.ac.uk/~llsling1/Upsid.interface.www/UPSID.interface.html

[35] http://www.linguistics.ucla.edu/faciliti/sales/software.htm

[36] Scott Drellishak wrote C code to extract and transform the DOS data into Microsoft Access for a seminar on cross-linguistic universals taught by Sharon Hargus at the University of Washington.

database tables. The original UPSID$_{451}$ tables are shown in Figure 4.5.[37]

Figure 4.5: UPSID$_{451}$ database schema



CharCodes
CCID INT
CharCode VARCHAR(45)
Description VARCHAR(45)
Count VARCHAR(45)
plos VARCHAR(45)
imp VARCHAR(45)
ejstop VARCHAR(45)
... VARCHAR(45)
Indexes

Features
LongName VARCHAR(45)
ShortName VARCHAR(45)
Type VARCHAR(45)

Languages
LangName VARCHAR(45)
LangNum INT
Description VARCHAR(255)
Indexes

Segments
CCID INT
LangNum VARCHAR(45)
Anomalous VARCHAR(45)
Indexes

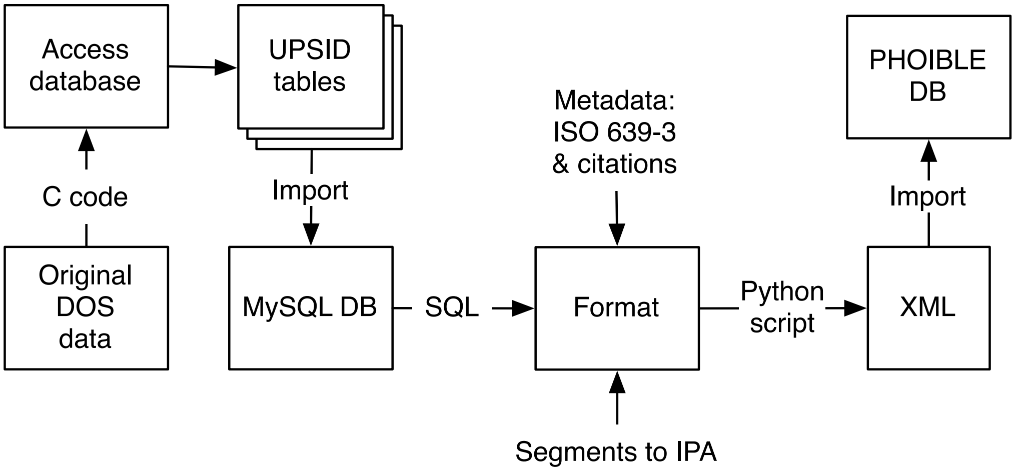
After each segment was converted into Unicode IPA and each inventory identified with a language code, these data were imported into MySQL database tables. A SQL query was used to join the UPSID$_{451}$ tables and then I output the transformed data into an intermediate format. A Python script then converted the output into an XML file, which was used to import the data into the PHOIBLE database. Figure 4.6 illustrates the ETL process that was undertaken with the UPSID$_{451}$ data. Although the ETL process has more steps than that of SPA's, the fact that the original data were already available in ASCII-encoded electronic format saved me time in the overall transformation process because the original data did not have to be retyped by hand.

Like SPA, to get the UPSID$_{451}$ data to interoperate with the segments from other segment inventories, each of its 921 segment types, represented with ASCII codes and written descriptions, were transformed into Unicode IPA characters. Table 4.4 provides a few examples of these correspondences.

There were a few problematic cases encountered in transforming segments into IPA. The first was what to do with segments underspecified for place of articulation, such as those labeled "dental/alveolar". These occur in 98 of the 921 segment types (of which 40

[37] Note, I have abbreviated the number of features in the `CharCodes` table for illustration's sake. There are 64 features in total. I also did not provide relationships between the tables because the original data did not explicitly state any.

Figure 4.6: UPSID$_{451}$ conversion process



occur in only one language in the UPSID$_{451}$ sample). Again, like the contents of SPA, I have chosen not to make changes to the original database's contents, except to interpret segment descriptions into IPA.[38] I have simply encoded these underspecified sounds with a vertical line <|> to indicate "or", e.g. /t̪|t/ indicates dental /t̪/ or alveolar /t/. In the case of underspecified segments then, a user interested in finding all t-sounds in the world's languages would have to query the database on /t/, /t̪/ and /t̪|t/. However, querying these segments is not problematic if one uses features instead of segments. I developed technological infrastructure in the form of an RDF/OWL knowledge base, so that users can underspecify their segment queries by either adding logic restrictions to the relationships between segments or by coarsening their queries at the level of distinctive features.[39]

Second, what do we do with sounds that are underspecified for manner of articulation? An example is "voiced alveolar r-sound". In fact, about 11% of r-sounds were dropped from

[38] This includes errors in UPSID$_{451}$ segment inventories as suggested by Vaux 2009 (discussed in Section 2.3.1). Following my methodology, to address Vaux's remarks one would simply add additional segment inventories with the changes that he has suggested and rank them higher than the UPSID inventories. This would leave intact the original UPSID resource for those who wish to compare their results with those of previous studies that used the original UPSID data.

[39] See Section 3.2.3 regarding the data model and Chapter 6 regarding features.

Table 4.4: Examples of UPSID_{451} segment and Unicode IPA correspondences

UPSID description	UPSID ASCII	IPA
long labialized pharyngealized voiceless uvular fricative	XW9:	χʷˤː
breathy voiced low central unrounded to high back rounded diphthong	auh	a̤ṳ
voiced alveolar lateral affricated click	g#	g‖

the UPSID_{317} analysis because authors failed to specify the manner of articulation of the r-sound in their language descriptions (Maddieson, 1984). Consequently an analysis of the most common r-sounds was not possible. The UPSID_{451} data set also contains cases where both place and manner of articulation are underspecified, e.g. "voiced dental/alveolar r-sound". This theoretically-driven issue of underspecified *archiphonemes* resonates in many inventory descriptions, not just those in UPSID_{451}. I have marked these cases with "*R", "*L" and "*N" for the time being and I have given them partial featural descriptions in PHOIBLE.[40]

The third problem was again the lack of an IPA representation for a particular segment. UPSID_{451} uses the feature description "fricated", which I currently represent with U+0353 COMBINING X BELOW. A full list of the UPSID_{451} and IPA segment correspondences, as I have interpreted them, along with notes regarding their conversions is given in Appendix F.[41]

After the segment transformations, each inventory was identified with an ISO 639-3 code, thus making the contents of UPSID_{451} compatible with the language descriptions in

[40] See Section 6.4.

[41] Two segment types in UPSID_{451} do not appear in any inventory. These are "G<", a voiced uvular implosive, and "h2", perhaps a typo. See also Reetz's list of typographical changes: `http://web.phonetik.uni-frankfurt.de/upsid_changes.html`.

other databases. Language descriptions in UPSID$_{451}$ again illustrate common problems in identifying language codes with language descriptions: language name resolution and identification of a language description's specific language variant, particularly in regard to macro-languages.

Language name resolution is exemplified by UPSID$_{451}$ language names "MIEN" and "TSESHAHT"; each is provided with an alternative language name, respectively "YAO" and "NOOTKA". In the case of Mien, the language name search space is large and ambiguous. The Ethnologue lists Mien as Lu Mien [ium], with alternative language names: Ban Yao, Highland Yao, Mian, Mien, Myen, Pan Yao, Yao, Yiu Mien and Youmian. In SPA, the language is listed as Yao, and the sources (Purnell, 1965; Mao et al., 1982) from which UPSID and SPA extracted the segment inventory also use the language name Yao. However, there is another language also called Yao [yao], spoken in Africa instead of Asia, with alternative language names: Achawa, Adsawa, Adsoa, Ajawa, Ayao, Ayawa, Ayo, Chiyao, Djao, Haiao, Hiao, Hyao, Jao, Veiao and Wajao. Language name disambiguation is a difficult task, exemplified by the fact that the Ethnologue lists 47,000 known alternative language names. On the other hand, in the case of the Tseshaht language name in UPSID, searching Ethnologue and its list of alternative language names returns no results. In these cases, WALS was helpful because it includes many of the UPSID$_{451}$ reference citations and each is associated with an ISO 639-3 language identifier. Thus I was able to use this information in tagging UPSID$_{451}$ inventories with language codes.

Discussed in detail in the previous section for languages in SPA, macrolanguages correspond to a one-to-many mapping between a macrolanguage and individual language identifiers. Two examples of language names used in UPSID$_{451}$ that now fall under the category of marcolanguages are given in Table 4.5.

In many cases a group of related languages does not have a macrolanguage code, so a particular language identifier needed to be assigned to a segment inventory. The identification of a language description's specific language variant, and therefore ISO 639-3 code, is exemplified by the segment inventory description of Andamanese used in UPSID$_{451}$. Today, Andamanese is not considered one language, but a language family. There are two genera consisting of 13 languages: Great Andamanese (Central (6) and Northern(4)) and South

Table 4.5: Some macrolanguages found in UPSID$_{451}$

UPSID language name	Macrolanguage code	Possible languages
Azerbaijani	[aze]	North Azerbaijani [azj] South Azerbaijani [azb]
Kanuri	[kau]	Central Kanuri [knc] Manga Kanuri [kby] Tumari Kanuri [krt]

Andamanese (3) (Lewis, 2009). In the Ethnologue, there are no citations that match the ones used in UPSID$_{451}$.[42,43] WALS offers a little more insight and references a separate publication by one of the authors[44] and associates that publication with Great Andamanese (Ethnologue: A-Pucikwar [apq]; Classification: Andamanese, Great Andamanese, Central). For the time being then, I have chosen [apq] as this entry's ISO 639-3 code, knowing it may be the incorrect identifier (A-Pucikwar is also listed as the last remaining Great Andamanese language; the other nine are now extinct). By assigning this inventory to a language identifier within a small language family, the potential lack of precision is unlikely to adversely impact statistical analyses that sample from genealogical groups or geographic regions. My decisions regarding which ISO 639-3 codes are associated with which language resources are documented in Appendix B.

During the development of PHOIBLE, I have relied mainly on the Ethnologue and WALS for assigning ISO 639-3 codes to particular language descriptions. A notable case that raises broader issues of attribution was when the Ethnologue and WALS assigned different codes to the same publication. Xiriâna (also spelled Shiriana) [xir] and Ninam [shb] (alternative

[42] Radcliffe-Brown, A. 1914. Notes on the languages of the Andaman Islands. Anthropos 9: 36-52.

[43] Voegelin, C.F. and Voegelin, F.M. 1966. [Andamanese]. Languages of the World: Indo-Pacific Fascicle 8 (Anthropological Linguistics 8/4): 10-13.

[44] Radcliffe-Brown, A. R. 1948. The Andaman Islands. Free Press.

names Xirianá and Shiriana), both spoken in Brazil, are each cited as the language described in *Shiriana Phonology* (Migliazza and Grimes, 1961), by Ethnologue and WALS, respectively. After some investigation and consultation with Amazonianists, Haspelmath (p.c.) reports that the WALS references (Migliazza and Grimes 1961, Borgman and Cue 1963 and Gómez 1990) are related to the Yanomam family, so they match Ninam [shb]. Apparently Ethnologue's bibliographic entry for Migliazza and Grimes 1961 is incorrectly labelled as the unclassified Arawakan language, Xiriâna [xir]. And one can see why, with such easily confusable language names.

The issues raised here are in regard to identifying a language described in a specific publication. Michael Cysouw and Jeff Good have coined the term *doculect* to describe the language variety described in a particular document. As Haspelmath (p.c.) points out, evidence about languages resides in descriptive documents, so to say that two doculects describe the same language variety is an additional claim above the level of the documents themselves. Language identification is a difficult task, especially when one is faced with a grammar of language X, but X is now known to be a group of distinct languages.

Finally, after the ETL process was applied to the UPSID$_{451}$ inventories that were extracted from the original DOS files, I was able to evaluate the accuracy of the output of the ETL process by comparing the segment and frequency counts from the transformed data against Reetz 2005. Errors from the conversion process were then identified and fixed.

To summarize, in this section I gave a brief overview of UPSID. I then discussed the problems in mapping UPSID's segments to IPA and the challenges in assigning an ISO 639-3 language name identifier to each segment inventory. I described the ETL process that was implemented to transform the contents of UPSID$_{451}$ from an old DOS program into an interoperable data format that includes segment inventory metadata and Unicode IPA segments. I then imported the interoperable data into the PHOIBLE database, which allows users to query the inventories at the segment level. The contents of the PHOIBLE database have been transformed into an RDF/OWL knowledge base. The knowledge base allows users to query segment inventories at the level of distinctive features, which addresses the problem of querying segments that are underspecified for place or manner of articulation.

4.3.3 Alphabets of Africa

Another segment inventory database is *Systèmes alphabétiques des langues africaines* (AA; Chanard 2006).[45] This online resource is a digitization of segment inventories from *Alphabets des langues africaines*, a compilation of the phoneme inventories and orthographies of 200 languages spoken in Africa (Hartell, 1993). In this compilation, each phoneme inventory and its associated orthography was provided by a language specialist or garnered from one or more language publications. Published by UNESCO, the aim of AA is to provide a description and make accessible the diversity of phonological systems in African languages and to illustrate the different solutions adopted by different countries in their development of alphabets for these languages. Chanard's website allows users to browse languages' phonemic systems through IPA-like charts that show the correspondences between each phoneme and its grapheme(s). The languages are listed by language name, ISO 639-3 code, country and language family (genus level). Chanard's website provides a rich resource for segment inventories of African languages.

AA contains phonemes and their orthographic representations for 203 languages. The ETL process I developed began when I scraped the contents of the AA webpages with a program that I wrote to download the pages and parse out the phonemes and graphemes that were embedded in HTML tables (Moran, 2009). This was accomplished with a Python script and a few regular expressions. The ETL process for AA is illustrated in Figure 4.7.

After the website was scraped, each segment inventory was parsed out and written to a simple tab-delimited flat file. These files contain the metadata for each language and each row in the file associates a phoneme to its corresponding grapheme. An example is given in Table 4.6. After the data were written to flat files, each file's segments were checked for Unicode IPA compliance and corrected if necessary. Then the data were transformed into an XML representation and imported into the PHOIBLE database.

The path to data interoperability of segment types was simpler than that of SPA and UPSID$_{451}$ because the segment inventory data were already digitized, and for the most part,

[45] By "AA" I mean the segment inventories and associated data in Hartell 1993, the digitized and updated version by Chanard (2006), or both depending on the context. Chanard's online version is available at: `http://sumale.vjf.cnrs.fr/phono/`.

Figure 4.7: AA ETL process

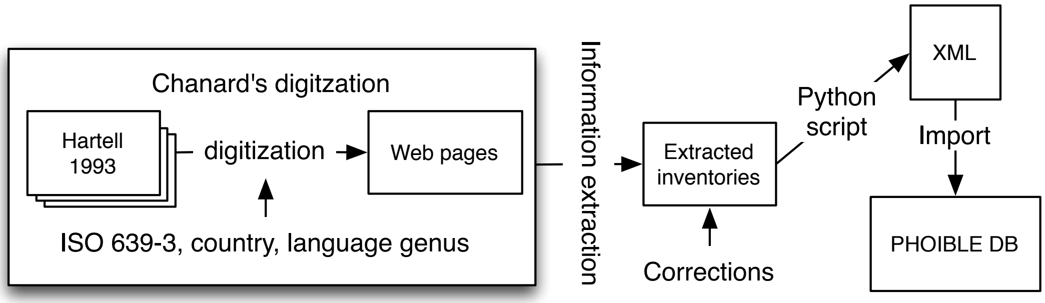


Table 4.6: Selected Sissala [sld] phoneme and grapheme correspondences (Hartell, 1993)

Phoneme	Grapheme
a	a
ʧ	c
ɲ	ny
j	y

represented in correct Unicode IPA. The path to make the segment inventories compatible was also less laborious because AA contains an ISO 639-3 language identifier for each segment inventory. These additions, however, also introduced errors in the data (in addition to errors found in the inventories presumably from the digitization). For codes, for example, the languages Daba and Kɔɔzime are marked [dab] and [nje] in AA, but they are now listed [dbq] and [ozm] in the ISO 639-3 standard. This is probably due to the nature of the ISO 639-3 code set. The codes are being updated annually, but that does not mean websites' contents are also being updated to reflect those changes. I found these errors by checking Chanard's codes against the latest version of the ISO 639-3 code set.

Errors in the digitization of segments for Chanard's online version were more difficult to catch. First, Christopher Green and I went through each extracted segment inventory and verified its contents against Hartell 1993 and we corrected any discrepancies that we found between Chanard 2006 and Hartell 1993. This included adding missing segments, removing additional segments and changing some segments into Unicode IPA (for example, Hartell (1993) uses some Africanist transcription conventions). I wrote a Unicode IPA validator to verify that all segments taken from AA adhered to their correct Unicode IPA codes points before the inventories were loaded into the PHOIBLE database.[46] This Unicode IPA validator takes as input a list of segments, splits them into characters (when they are not singletons), and then checks each character against a unique list of Unicode IPA code points that was curated by hand.[47] Several different types of errors appear in the data.

First, some incorrect symbols are simply erroneous Unicode IPA characters. For example, the Unicode Standard specifies LATIN SMALL LETTER SCRIPT G <ɡ> at U+0261 for the IPA voiced velar stop. AA uses the standard keyboard <g> LATIN SMALL LETTER G at U+0067. This is a common mistake found in online resources using IPA. Another example is AA's use of LATIN SMALL LETTER SHARP S <ß> at U+00DF instead of GREEK SMALL LETTER BETA <β> at U+03B2 for the bilabial fricative. Both mistakes are easy to make because these symbols are homoglyphs. Additionally, the Unicode Consortium decided to not to include additional code points in the IPA block for symbols already encoded in other character ranges, e.g. the bilabial fricative <β> resides in the Greek and Coptic block, the Latin letters in IPA reside in the Basic Latin block, etc. Only IPA-specific characters reside in the IPA Extensions block, i.e. the 96 characters in the range from U+0250 to U+02AF.

A second issue is the now decommissioned IPA segments used in AA, including /ɩ/ and /ɷ/, which are used in the Africanist transcription tradition. In PHOIBLE these were changed to their current IPA equivalents /ɪ/ and /ʊ/. A third issue is theoretical and was raised in Section 4.3.2, namely, what should be done with the use of archiphonemes in language descriptions? For the time being, I have simply marked archiphonemes with an asterisk and a capital letter.[48] A fourth issue relates to AA's use of a now depreciated

[46] Although I provide details of the validator in this section, it was also used on the contents of SPA and UPSID_{451}, which were collected chronologically after AA.

[47] See Appendix D for the complete list of Unicode IPA characters.

[48] Note that since capital letters are not legit IPA characters, they were added to the Unicode IPA descrip-

private use area (PUA) character at U+F25E (in earlier versions of SIL Doulos) to encode LATIN SMALL LETTER V WITH CURL at U+2C74. I have changed this to the sanctioned LATIN SMALL LETTER V WITH RIGHT HOOK <ⱱ> at U+2C71, introduced in Unicode version 5.1. Lastly, there are two undocumented graphemes: <ř> in inventories Banda (Sudan) [bfl] and Murle [mur]; and <r*> in Kabiyɛ [kbp], which appears to be a marker that <r> only appears in loanwords (Hartell, 1993, 288). For these inventories I consulted additional sources and vetted the segment inventories and made the appropriate changes.

Multiple reuse of linguistic data poses a problem for data accuracy (Thomason, 1994; Lewis et al., 2006). Can we trust that the data retains its integrity, i.e. can we assume that the data are unchanged from the original resource to the final one? The AA-to-PHOIBLE data path began when a researcher collected documentation from a native speaker of a particular language. This was most likely an impressionistic analysis of the sounds in the language (opposed to a rigorous acoustic analysis) as they were heard by the field linguist. He or she then undertook a phonemic analysis of the language, positing phonemes based on criteria such as which allophone occurs most frequently or is least affected by its environment. At some point this work was written up and published (and typographic errors may have crept into the manuscript). Impressionistic analyses of the same language may differ, so multiple resources on the same language were sought out for the AA compilation. The resources were consulted and the phonemic and graphemic inventories were selected. The AA compilation had to be typeset and published, further introducing opportunities for mistakes like typos. Thirteen years after AA was published, the data were digitized by Chanard (2006). The digitization is another point in the data's path that introduces the possibility of errors. Take for example, digitizing the data in a text editor. The data enterer would need to ensure that the character set is UTF-8 and not ASCII, since the AA data were digitized in Unicode IPA. If the document is uploaded or downloaded, the transfer session would have to be set to binary transfer (or some other lossless transfer) because an ASCII transfer (the default in some FTP programs) would corrupt the characters. The character set needs to remain intact to ensure accuracy and integrity. After the digitization,

tion table in Appendix D. Archiphoneme segments are assigned distinctive features using underspecification. This issue is discussed in Section 6.4.

the data were transformed into a database implementation, which introduces programmatic challenges and possibly more errors. Additionally, if say for example MySQL was used, the database, its tables, and even the fields that contain the data must be set to the correct character encoding, otherwise the imported data will be corrupted. Finally, my own work identifying and extracting the data and transforming it into a format for the PHOIBLE database introduces many points for introducing errors.

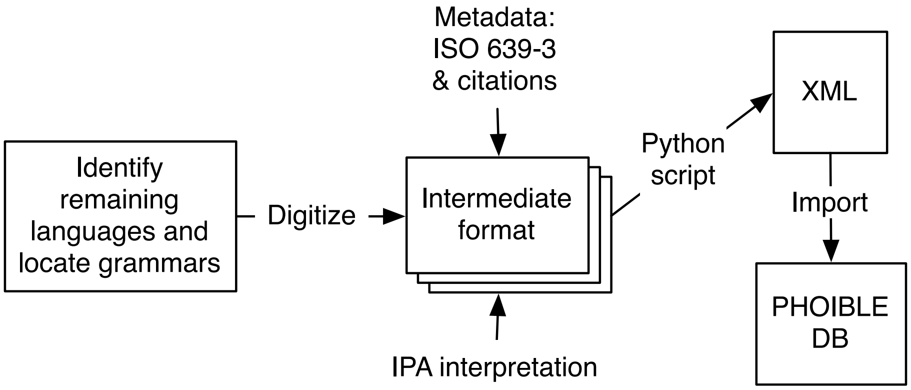
To summarize, in this section I have briefly discussed the two different versions of *Alphabets of Africa* and I described the ETL process that I created to extract the online AA's segment inventories and make them interoperable with the SPA and UPSID$_{451}$ inventories in the PHOIBLE data set. Although AA was "born digital" and it includes metadata for each inventory, its contents had to be nevertheless checked for Unicode IPA compliance and its inventories had to be matched to the correct ISO 639-3 language name identifiers.

4.3.4 PHOIBLE inventories

The PHOIBLE inventories increase the scope of SPA, UPSID$_{451}$ and AA by 485 languages. These additional inventories were extracted from roughly 150 PhD dissertations, 75 books, and numerous articles from peer-reviewed sources such as *Illustrations of the IPA* in the Journal of the International Phonetic Alphabet (JIPA). PHOIBLE inventories include minimally a description of each language's phonemes, but allophones and phonological conditioning environments are included when they were described in the resource. For this work I have not reinterpreted authors' phonological analyses. However, each original description was evaluated and I have thrown out any inventories that were deemed not rigorous in their scientific methodology (e.g. practical orthography descriptions without supporting linguistic evidence).

Little attempt was made to increase the PHOIBLE sample of inventories in a genealogically balanced way. By this I mean that although I sought out documented languages in families that had no or little representation in PHOIBLE, at the same time I did not discriminate against including inventories from families that were already well represented if good phonological descriptions were able to be located. My ultimate goal is to attain comprehensive coverage of documented languages. Any language not included in the databases of SPA, UPSID$_{451}$ and AA was targeted. In cases when an inventory from one of the databases seemed questionable, additional linguistic descriptions of that language were sought out and the additional segment inventory was added to PHOIBLE. Therefore, there may exist multiple resources for the same language.[49] The collection of PHOIBLE inventories provided challenges distinct from the ETL processes described in the previous sections. The ETL process is illustrated in Figure 4.8.

Figure 4.8: PHOIBLE inventories digitization and transformation process



Linguistic descriptions had to be identified, the quality of the description evaluated, and the data extracted through digitization. Each linguistic description was identified with an ISO 639-3 language name identifier and the segments in each inventory were interpreted into IPA. The bibliographic metadata for each resource was also collected. The initial digitization was undertaken in Excel spreadsheets that were then exported and transformed into an intermediate mediate CSV format. The data were checked for compliance to Unicode IPA (even we sometimes made mistakes with keyboarded characters) and then transformed into XML and loaded into the PHOIBLE database.

[49] In some cases an additional segment inventory was accidentally added when one already existed in another database. On the other hand, several AA inventories seemed questionable, so Christopher Green and I first vetted those inventories based on the original resources when they could be found. When they could not be located, we added additional segment inventories for those languages based on other sources. A trump hierarchy can be used when selecting a unique list of languages from the combined PHOIBLE inventories. See Section 3.2.2.

The main challenge with extracting segment inventories from language descriptions is whether or not phonetic and phonemic segments are easily interpretable or explained in the text. Again, our rule of thumb has been to trust the linguist's analysis and extract inventories as they are posited in his or hers description. In this manner we leave the "normalization" for typologists (cf. Hyman 2008). During the data collection I decided to explicitly mark marginal phonemes and archiphonemes so that they could be included or excluded from studies. Where there were issues in descriptions that were problematic to interpret, I have noted these issues in the original spreadsheets. If a description proved more than minimally problematic, we simply did not include its inventory in PHOIBLE.

As with the data sets described in the previous sections, the use of characters in phonological descriptions that are not explicitly recognized by IPA presented a challenge when extracting segment inventories from language descriptions. The extraction of segment inventories for PHOIBLE also required me to make decisions regarding segment and diacritic ordering. For example, when more than one diacritic appears below a segment, I chose to first use the place feature (dental, laminal, apical, fronted, backed, lowered, raised), followed by the laryngeal setting (voiced, voiceless, creaky voice, breathy voice), and finally by the syllabic or non-syllabic marker. So for example, a creaky voiced syllabic dental nasal appears as /n̪̰̩/. When there was more than one diacritic to the right of a segment, I chose the order: unreleased/lateral release/nasal release → palatalized → labialized → velarized → pharyngealized → aspirated/ejective → long. For example, a labialized aspirated long alveolar plosive is represented as /t^{wh}ː/. These conventions are provided in Appendix C.

To summarize, the ETL process for PHOIBLE inventories was rather straightforward: find a description of a language by an author that is rigorous in his or her scientific methodology (preferably a language not yet represented in PHOIBLE), read and interpret the author's description of the language's segment inventory, and then input the relevant information into a spreadsheet (while following the segment conventions outlined in this section). The inventory is then transformed into an intermediate CSV format and the relevant metadata is added (bibliographic citation and ISO 639-3 code). Lastly, the segment inventory is transformed into XML and then imported into the PHOIBLE database. The main challenge in this overall process is in interpreting an author's description and analysis of a given language's phonological system and encoding the segment data in (Unicode) IPA.

4.3.5 Summary

In this section I described the challenges and the ETL processes used to bring together SPA, UPSID_{451}, AA and the PHOIBLE inventories into one interoperable data set. The initial lack of interoperability between the resources described in this section highlights the need for technological infrastructure that supports research across disparate data sets. The use of standards such as ISO 639-3, IPA and Unicode will promote interoperability between PHOIBLE and other resources, such as those being added to the Linguistics Linked Open Data cloud,[50] an effort being spearheaded by the Working Group on Open Data in Linguistics.[51]

4.4 Genealogical coverage

Combining the SPA, UPSID_{451}, AA and PHOIBLE segment inventories together results in a sample that represents 16% of the world's languages. At the time of writing, there is no simple and straight forward means to evaluate the genealogical coverage of a large typological data sample on a family-per-family (or genus-per-genus) basis. Even though many genealogical language classifications are working hypotheses, it is nevertheless important to establish what the genealogical coverage of a typological data set is, thereby allowing the coverage of different data sets to be compared. In this section I describe a method I developed for evaluating the genealogical coverage of a data set by using a list of ISO 639-3 language name identifiers and simple XML representations that represent language family trees, extracted from the Linguist List's Multitree project (LINGUIST List, 2009).[52] I use this method to assess the genealogical coverage of PHOIBLE by comparing its contents with language families in the Ethnologue 15th edition, currently the most-up-to-date data available through Multitree.

[50] http://linguistics.okfn.org/resources/llod/

[51] http://linguistics.okfn.org

[52] http://multitree.linguistlist.org

To evaluate PHOIBLE's genealogical coverage, an index of its contents must be evaluated against an index of languages encoded by genealogical groups. Indices of languages date back at least as far as Hervas 1784. Since the 18th century, our knowledge about the diversity of languages and their relations has greatly increased. In the 20th century, several comprehensive language indices were compiled, including Ruhlen 1975, Ruhlen 1987, Voegelin and Voegelin 1977, and Moseley et al 1994. However, the most comprehensive list is the Ethnologue (Lewis, 2009). The first edition appeared in 1951 and cataloged 46 languages. By the 7th edition in 1969, it already listed 4493 living languages. In 1971 a computerized database was constructed for its contents and three-letter language identifiers were assigned to each language, "on the order of international airport codes".[53] These three-letter language identifiers evolved over the years and were recently reconciled with the ISO 639-2 and ANSI Z39.53[54] standards to become officially recognized by the International Organization for Standardization (ISO) as ISO 639-3. ISO 639-3 provides codes for the representation of names for nearly 7500 languages, including living, extinct, ancient, historical and constructed languages. SIL International is the registration authority for ISO 639-3 and oversees the annual change requests (additions, deletions or modifications) of the language codes.[55] Thus the standard evolves to reflect what is known about the world's languages and projects that adhere to ISO 639-3 are faced with the challenge of updating their metadata to reflect these annual codes changes.

There are no standardized computable representations of language families. To alleviate this problem, one option is to scrape the Ethnologue for their structure and contents. The Ethnologue presents language families through hyperlinks of connected webpages. This is a detailed process that requires analyzing the structure of connected webpages and recursively following links through sub-families until all languages are found. Furthermore, the relevant parts of the webpages have to be identified and the data correctly extracted. This proposed webpage scraper would also be brittle because changes to the structure of the Ethnologue webpages would break the script. Despite these challenges, the Multitree project has already

[53] See references in: http://www.ethnologue.com/ethno_docs/introduction.asp.

[54] MARC language codes: http://www.loc.gov/marc/languages/.

[55] http://www.sil.org/iso639-3/

crawled and scraped the Ethnologue 15 website's contents, put its language families data into CSV files that fit Multitree's internal working format, and devised and added their own unique four-letter language family codes and dialect information (Danielle St. Jean, p.c.).

Multitree's purpose is to generate visualizations of scholarly hypotheses about language families from a searchable database. For each hypothesis of a language family, Multitree also publishes an XML database dump of that data. Although the XML file adheres to a schema that is specific to the Multitree database, it nevertheless encodes the parent-child relationships of languages within each genealogical classification along with metadata about that language family. Multitree's XML data are represented in a tree data structure (recursively embedded hashes), so extracting the relevant information such as the ISO 639-3 language identifiers from within the <codes> tags is straightforward with an XML parser. The XML data encode the structure of the phylogentic tree that is displayed on the website, which is then easily preserved in a simpler XML file (minus the Mutltitree database-specific information).

To assess PHOIBLE's genealogical coverage for each language family, I downloaded the Multitree's XML representations of the Ethnologue's language family classifications. I wrote a script to extract the phylogenetic tree structure with language and language family codes. Then for each language family, I compared the PHOIBLE segment inventory index (in ISO 639-3 codes) and computed PHOIBLE's genealogical coverage. The distribution of languages in language families is very skewed.[56] The six language families in Table 4.7 represent over 60% of the world's languages. About half of PHOIBLE's segment inventories belong to these six language families. Appendix A provides the full list of 114 language families in the Ethnologue and shows PHOIBLE's coverage for each. Also on page 302, Figure 7.6 illustrates with a line plot the genealogical coverage of PHOIBLE in comparison to the number of languages in each of Ethnologue's language families.

[56] About a third of all language families, as listed in Ethnologue, have one language.

Table 4.7: Genealogical coverage of PHOIBLE for major language families

Language family	Ethnologue	PHOIBLE	Coverage
Niger-Congo	1516	270	17.8%
Austronesian	1271	81	6.4%
Trans-New Guinea	565	52	9.2%
Indo-European	450	51	11.3%
Sino-Tibetan	411	31	7.5%
Afro-Asiatic	375	51	13.6%

4.5 Summary

To summarize, in this chapter I have described in detail the contents of PHOIBLE, the ETL processes that were undertaken to merge the different segment inventory databases into one interoperable data set, the challenges involved in those processes, and the combined genealogical coverage of these resources. In the next chapter, I use PHOIBLE to investigate descriptive typological hypotheses about segment inventories in the literature.

Chapter 5

SEGMENTS AND INVENTORIES

5.1 Introduction

A segment inventory consists minimally of the set of consonants and vowels in a language. This set may be stated purely in terms of contrastive sounds, i.e. the set of phonemes employed by a language as postulated by a linguist, or it may also include the set of allophones that describe the non-contrastive sounds in the language, i.e. its phonetic inventory. However, as straightforward as these definitions appear, defining what goes into a segment inventory is an area of debate that impacts the conclusions reached in phonological typology studies. For example, authors of phonological descriptions do not necessarily agree on the phonemic status of segments that are breathy, creaky, nasalized, lengthened, pharyngealized, etc. These secondary phonation types can radically change the size of a segment inventory.[1] For example, compare the range of vowel inventory size in UPSID$_{451}$ (3-46) (Maddieson, 1984; Maddieson and Precoda, 1990), which includes secondary phonation types, with WALS (3-14) (Maddieson, 2008c; Haspelmath et al., 2008), which does not.

In this work I have taken a data-driven approach in collecting segment inventories from different tertiary databases and from secondary resources like grammars and phonological descriptions. These resources vary widely in their descriptions and analyses of languages' segment inventories. The technological architecture that I have developed allows users to decide whether they want to keep or remove certain segment types from their experiments, such as diphthongs, tone or vowels with secondary phonation types. In this chapter, I investigate whether descriptive typological facts about segment inventories still hold up when we probe a much larger database of languages.

In Section 5.2, I provide some background about the resources and work from which I examine properties of segment inventories. In Section 5.3, I examine the distribution of

[1] See discussion in Section 2.3.4.

segment types and investigate to what extent the genealogical skewing of inventories in PHOIBLE affects segment type frequency. The genealogical resampling method I use in this section is also applied in Section 5.4, in which I look in detail at aspects of segment inventories. In Section 5.5, I investigate the ratio between consonants and vowels across inventories, which is one area of typological interest because it has often been equated with complexity in phonological systems. Lastly, in Section 5.6, I ask whether Crothers's (1978) observation, based on the segment inventories in SPA, that the vowel systems in most languages contain /i, a, u/ still holds in the PHOIBLE data set. I use a statistical technique called multi-dimensional scaling to visualize how vowel systems expand after /i, a, u/.

5.2 Background

The Stanford Phonology Archive (SPA) was the first computerized segment inventory database used to test statistical claims about phonological universals (Crothers et al., 1979).[2] The ability to query a database of segment inventories for evidence and counter-evidence provided a new research tool for investigating phonological universals and the cross-linguistic frequency of segments. For example, Crothers (1978) utilized SPA to describe typological universals of vowel systems and observed that 98.5% of languages in the SPA sample have the vowels /i a u/. However, SPA did not provide a genealogically balanced sample of languages, which lead Sherman (1975) to raise the important issue of language sampling. How does one devise a cross-linguistic language sample that captures genealogical, areal and typological diversity?

The UCLA Phonological Segment Inventory Database (UPSID) compiled by Maddieson drew on the work of SPA, but it included substantially more languages (from SPA's 197 to UPSID$_{317}$ in Maddieson 1984 and later increased to UPSID$_{451}$ in Maddieson and Precoda 1990). Additionally, Maddieson aimed for a genealogically balanced sample, and inclusion of segment inventories was restricted by a quota sample, thereby limiting the sample to one language from each small language family (as determined at the time with the language

[2] For background, see Section 4.3.1.

family information available). Using UPSID, Maddieson's investigations led to explicit statements about the probable frequency of segments in the world's languages, the shape of phonological inventories and universal phonological tendencies. Indeed most of what is currently known about the distribution of sounds in the world's languages is based on UPSID.

The compilers of SPA and UPSID were faced with the decision of whether to stick with the original analysis of a phoneme inventory or to reanalyze the original phonemic analysis according to a consistent standard. I have taken the opposite approach in this work by accepting linguists' analysis at face value and by simply not including segment inventories from phonological descriptions that seem to lack scientific rigor. Whereas one author might consider diphthongs as phonemic and another considers them a sequence of two different phonemic vowels in succession, I simply add both analyses to PHOIBLE. Additionally, as long as a phonemic contrast has been purported in one language, I try to preserve it.[3] The infrastructure I built allows users to include or omit inventories given their linguistic preferences. I have, however, reinterpreted phonetic symbols and feature descriptions from all inventories into a consistent transcription standard for linguistic (and technological) interoperability. This means that I have interpreted and mapped the SPA and UPSID symbols and phonetic descriptions into Unicode IPA.[4]

During the development of PHOIBLE, my aim was to include as much detail as possible for each segment inventory. Thus, I included the allophonic information available in SPA, the graphemic data provided in AA, and when extracting inventories from published phonological descriptions for PHOIBLE inventories, I added phonemes, allophones, tone and phonological conditioning environments when this information was described by the author. Therefore, there are certain misrepresentations in the combined PHOIBLE database. For example, inventories include tone when they are treated as phonemic segments in SPA, AA or PHOIBLE inventories. Alternatively, if the inventory came from UPSID$_{451}$, it does not contain a description of tone. This reverberates in investigations of tone in languages in the

[3] For example, the voiceless/voiced contrast in implosives in Seereer-Siin [srr] (Mc Laughlin, 2005).

[4] See discussion in Chapter 4. Appendices E & F provide the SPA-IPA and UPSID-IPA correspondences.

PHOIBLE data set. Since there are sometimes duplicate inventories identified by the same ISO 639-3 code, I instantiated a mechanism to create a unique sample of languages via a definable trump hierarchy. For the purposes of investigating the typological distribution of segments in this chapter, I use the hierarchy: PHOIBLE inventories > SPA > UPSID > AA.[5] So when there are duplicate languages represented in the combined PHOIBLE data set, the pecking order is to first take a segment inventory from PHOIBLE, then SPA, UPSID_{451} and finally AA.[6] In the case of tone then, if an inventory is only represented in UPSID and happens to be a tonal language, tone is not included. Another user may wish to run queries against PHOIBLE minus SPA or UPSID or some combination thereof. This is possible because each inventory has been given a source identifier.[7]

In the following sections I present typological observations of segment inventories by comparing the PHOIBLE, UPSID_{451} and SPA databases. UPSID_{451} is intended to be a genealogically balanced sample of languages. SPA and AA are convenience samples. The PHOIBLE inventories are also a convenience sample, i.e. I collected inventories from the available literature and I did not adhere to any genealogically-balanced sampling procedure. Thus the entire PHOIBLE sample, which brings together these four databases, is one large convenience sample. Therefore, in this chapter I also devise and use a genealogical stratification sampling procedure to approximate the distribution of segments by correcting for genealogical bias. UPSID_{451}'s quota sample provides a point of comparison. I begin by looking at the distribution of segment types in the inventories in the combined PHOIBLE data set.

[5] In this experiment, SPA trumps UPSID in 157 languages because SPA contains descriptions of tone and UPSID does not. Additionally, there are 8 inventories in PHOIBLE that trump UPSID.

[6] If two or more inventories for the same language code are provided in one of the databases, e.g. there are multiple inventories for Fulfulde [fub] that I digitized for PHOIBLE, then the trump ordering is applied in ascending order of their inventory IDs. For example if there are four Fulfulde entries with inventory IDs 1, 2, 3 and 4. Then the trump order would be 1 for inventory ID 1, 2 for inventory ID 2, and so on. This is an arbitrary order, but one that can be reconstructed easily given the order of inventory IDs.

[7] See Section 3.2.

5.3 Distribution of segments

The International Phonetic Alphabet (IPA; International Phonetic Association 2005) is a system of phonetic notation that provides a standardized set of symbols for transcribing speech segments in the world's languages.[8] This set of symbols contains letters and diacritics that can be combined in various ways to denote the articulatory properties of a speech segment. The segments in a particular language are typically stated in terms of a set of contrastive sounds, referred to as a segment (or phonemic) inventory. In this section, I show that as the number of segment inventories in PHOIBLE increases, the number of segment types also increases.[9] I also show the frequency of segment types in the PHOIBLE data set before and after implementing a resampling method that estimates the genealogical bias of PHOIBLE's contents. In Section 5.4, I discuss the distribution of segment inventory sizes before and after applying the genealogical sampling method discussed in this section.

There is a large number and very wide range of segment types used in language descriptions and they show some interesting patterns. The first is the ratio of unique segment types with regard to the number of language descriptions in which they are found. In the UPSID$_{451}$ sample, 920 segment types appear in the descriptions of 451 languages.[10] In the PHOIBLE sample, 1780 segment types appear in the descriptions of 1089 (distinct) languages.[11] Figure 5.1 shows the increase in the cumulative number of segment types as languages are added to the PHOIBLE data set. So far, as I add new inventories, new segment types continue to appear at a rate as if the curve was bounded to infinity. I don't know of any obvious reason why the curve should be quadratic. I expected an asymptotic curve growing towards an upper boundary, but the current curve does not reach a maximum and for the current data there is no sign of any slowing towards an asymptote.

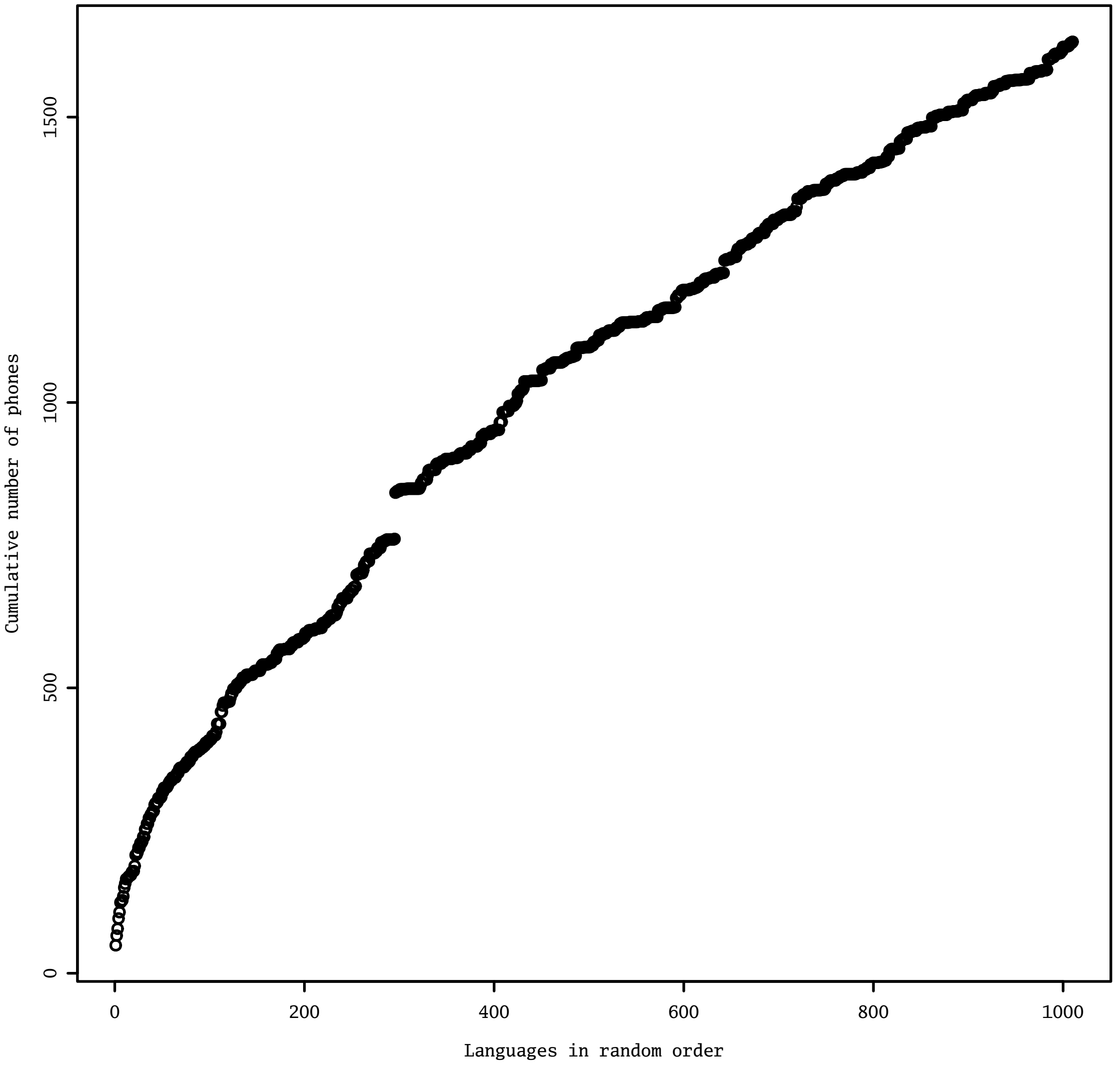
[8] See discussion in Section 2.3.5.

[9] See Section 2.1.2 for a description of the segment type-token distinction.

[10] For this analysis I have removed the <h2> segment in UPSID$_{451}$, which appears to be a typo (it does not appear in any inventory).

[11] In 1089 languages, there are 38244 segment tokens, of which 25922 are consonants, 11257 vowels and 1065 tones. These figures are only for phonemes and do not include allophones.

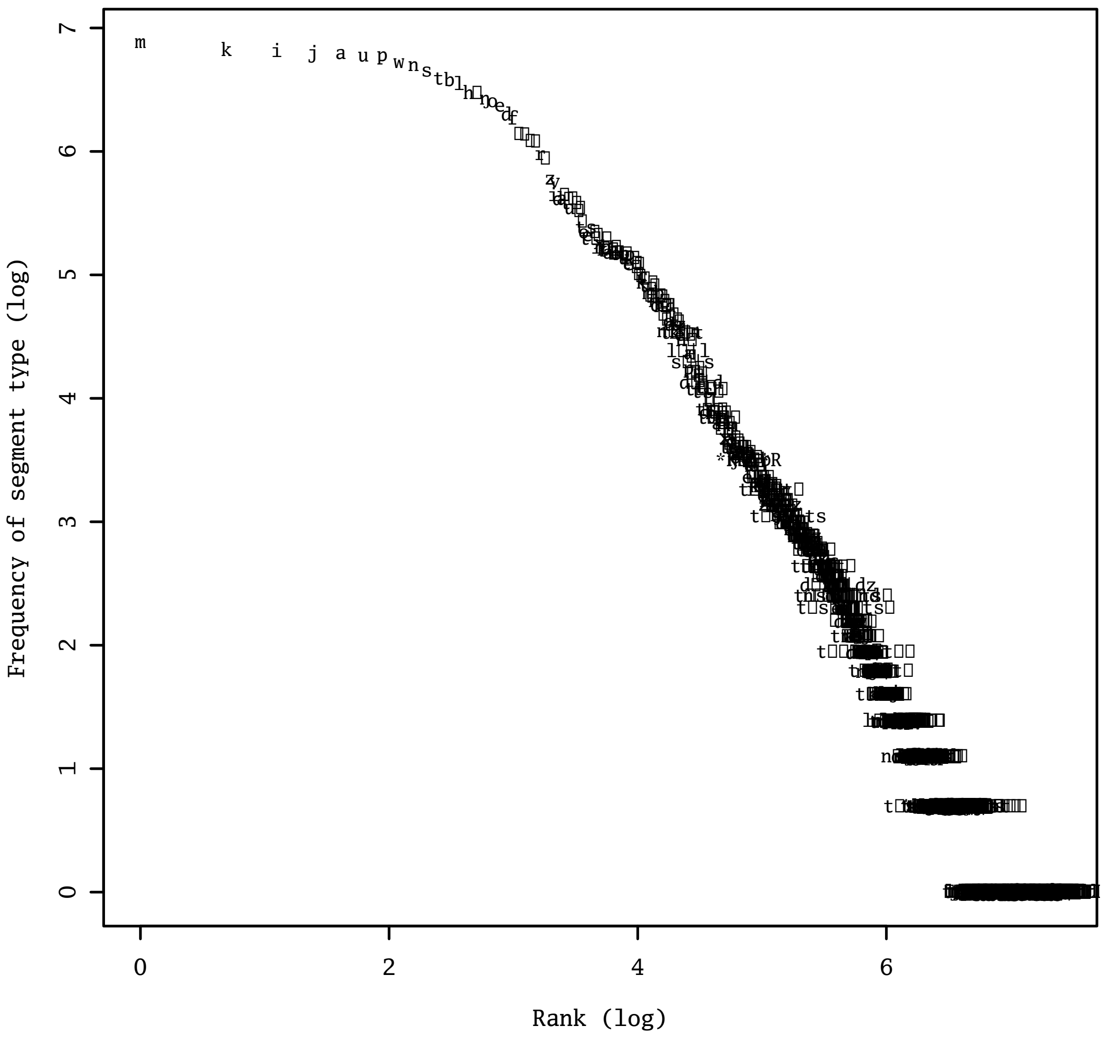
Figure 5.1: Cumulative number of segment types vs languages in random order



The languages in Figure 5.1 are plotted randomly and each iteration shows the same shaped curve. The coefficients of the interpolation of the log log plot suggest a quadratic relation. The intercept is 3.7076 and the log of the cumulative grouping is 0.5415. In the plot, the gap appears between !Xu [ktz] and the segment inventory that precedes it. !Xu is described as having 141 phonemes, 66 of which occur in no other language (Snyman, 1970, 1975; Maddieson and Precoda, 1990).

The second interesting pattern is the distribution of all segment types and their frequencies in inventories in PHOIBLE, shown in Figure 5.2. The log frequency of segment types (N=1780) is plotted against their log rank. The most frequent segments can be clearly seen at the upper left, e.g. /m, k, i, a, u, p/. These segments appear in most of the language descriptions in PHOIBLE. As the curve falls, the frequency of each segment type decreases and the number of unique segment types increases. At the bottom right of the plot, a mass of one-off segment types appears as one large blob. In fact, in both the UPSID_{451} and PHOIBLE data sets, around half of all segment types appear only once. In UPSID_{451}, 427 of 920 segment types are one-off occurrences, roughly 46.5%. In the PHOIBLE data, 909 of 1780 segment types are one-off occurrences, that is 51%. Thus half of all segment types found in languages descriptions in the data set are language-specific.

Figure 5.2: Frequency of segment types (log) vs rank (log)



Looking at these patterns regarding the distribution of segment types in the PHOIBLE data set is interesting because its language descriptions contain a genealogically and geographically diverse sample of languages. However, doing statistical inference to estimate the mean frequency in which a sound occurs across languages is not feasible on the PHOIBLE data set without some form of stratification. The problem with looking at just the frequencies of segment types in the data set is that its contents contain a genealogical and bibliographic bias, i.e. coverage is greater for certain language families (stocks and genera) due to the availability of language descriptions for those languages (and due to the resources that we chose).[12] For example, after compiling the PHOIBLE inventories I was intrigued by the relatively high frequency of the velar nasal /ŋ/, a sound I am familiar with through my work on West African languages, but one that is reportedly much less common in languages spoken in North and South America (Anderson, 2011). As will be discussed below, the higher frequency of velar nasals in the data set is due to the uneven geographic and genealogical make-up of the current data set. Of course an ideal segment inventory database would contain a theoretically uniform description of all segment inventories from all languages. This sample would represent the most complete population for investigating the distribution of segments and the shapes of segment inventories in the world's languages as they exist today. Note however that even if we had access to all those inventories (including undocumented languages), the range of possible human languages would not be represented in the current distribution of actual languages because today's languages are the result of the diffusion of typological features through shared descent and through areal effects due to geographic proximity. Thus statistical methods are used to control for genealogy so that we can attempt to account for the historical development of languages by assuming that there is a common trend within a language family and then we attempt to weight those groups accordingly.

To establish the probability through statistical inference that a language contains some typological feature, confounding factors like genealogical relatedness should be taken into

[12] There are several biases involved in creating a reliable sample to characterize the distribution of linguistic phenomena. See Section 2.3.2 for discussion. PHOIBLE's genealogical coverage is discussed in Section 4.4 and is illustrated in Figure 7.6 on page 302. Appendix A provides figures on PHOIBLE's genealogical coverage broken down by language family.

account when sampling typological databases.[13] As discussed in Section 2.3.2, several methods for choosing a typologically representative sample of languages have been proposed. A popular one is the diversity value (DV) sampling method developed in Rijkhoff et al. 1993 and then refined in Rijkhoff and Bakker 1998. Given a genealogical classification in a tree format as input and a typological data set, the DV method increases the probability that rare typological types will be represented in the language sample by adding together the change in the number of nodes at a given level in the tree. The DV method generates a variety sample to represent the diversity of phenomena that the researcher wishes to investigate and stratifies it to limit the influence of genealogical bias.

Although the DV sampling strategy is useful for generating a typological sample for exploratory typological research, I wanted a statistical method that would potentially allow me to incorporate as much data from PHOIBLE's inventories as possible, while also stratifying the sample.[14] To estimate the genealogical bias in the PHOIBLE sample, Taras Zakharko and I came up with and implemented a resampling technique in R that systematically recomputes a statistical estimate by randomly sampling from subsets within a data set. This technique averages the frequency of an element (e.g. segment types, consonant counts, etc.) over the number of iterations in which a segment inventory is randomly sampled from a chosen subset (e.g. language stock, language genus, geographical area, etc.).[15] In the experiment presented here, I use the language stocks from the Ethnologue 15th edition.[16] This procedure is run 1000 times and the frequency values are summed together and the mean is calculated. This method treats all subsets equally so that no bias from the inequality of subsets is introduced (the PHOIBLE data contains some big language families and some small families and the coverage of each varies from good to poor). If I were to just average over element counts for languages in a big family, they would be overrepresented

[13] Due to linguistic borrowing, areal bias is also a confounding factor. Experiments by Miestamo et al. (2011) show that areal stratification does not simply improve genealogical on sampling in producing a variety sample. They note that it is unclear why.

[14] In their experiment with WALS data, Miestamo et al. (2011) show that DV sampling does not fare much better than random sampling in capturing the diversity of typological variables (respectively 95% vs 94%).

[15] See Wu 1986 and Good 2006 for background on statistical resampling techniques.

[16] See Section 4.4 for discussion on these language families and how the data were collected.

in the results.

An assumption of this resampling method is that elements observed in each subset give a representative view of that subset. In this manner, I am essentially implementing Maddieson's quota sampling method during each iteration over the set of genealogical subgroups. But whereas Maddieson chose a representative sample for each small language family, I sample one representative from each group at random and assume that it is representative of the group in some way. An argument against this approach is that for small language families (e.g. those with singleton representatives, language isolates or families with only two or three members, etc.), it is not clear if these (surviving) languages should be representative of their prospective families. On the other hand, we want to get a representative estimate for each group and there is often only a limited set of data available. By controlling for genealogy with a resampling approach, we are assuming that there is a common trend in a language family group and that we are capturing some of those historical artifacts with the so-called *representative* for those languages. Therefore, the resampling procedure developed in this work also samples language isolates. Bond and Veselinova (2011) show that sampling with language isolates helps capture the distribution of sounds in different geographic areas, e.g. isolates in the Americas tend to lack voicing in fricatives, a feature that is considered an old world phenomenon (cf. Maddieson 2011c).[17]

Table 5.1 shows the 35 most frequent segments in the PHOIBLE data set, their genealogically controlled frequencies, their frequencies in the database and the difference between the two (ordered by controlled frequency).

Table 5.1: 35 most frequent segment types and their controlled frequencies by language stock

Segment	Controlled frequency (%)	Data set frequency (%)	Difference (%)
i	90.56	91.18	0.63
m	90.50	96.14	5.64
k	89.92	91.92	2.00

[17] Language isolates have an uneven geographical distribution. In some places like South America (and in particular Columbia), there are a disproportionately high number of language isolates.

Table 5.1: 35 most frequent segment types and their controlled frequencies by language stock

Segment	Controlled frequency (%)	Data set frequency (%)	Difference (%)
a	88.77	88.43	-0.34
p	87.16	84.85	-2.31
j	83.45	88.25	4.79
w	79.43	82.74	3.31
u	78.03	86.23	8.20
h	71.46	64.19	-7.28
n	70.65	80.99	10.34
s	69.37	77.13	7.76
t	65.64	73.37	7.73
ʔ	60.38	45.09	-15.29
b	59.67	73.00	13.33
l	55.01	70.16	15.15
ʃ	49.39	38.02	-11.38
g	47.38	65.66	18.28
o	42.67	61.16	18.48
e	41.14	59.41	18.28
d	40.90	56.01	15.11
ŋ	36.62	60.97	24.36
ɾ	30.52	21.76	-8.76
ɛ	29.93	47.75	17.82
o̞	28.43	16.25	-12.17
ts	28.04	21.58	-6.46
f	27.84	55.19	27.34
r	27.73	40.40	12.67
x	26.48	18.55	-7.93
ɔ	24.84	45.45	20.62
e̞	23.97	14.97	-9.00
ɳ	23.46	46.56	23.10
tʃ	22.97	28.01	5.03

Table 5.1: 35 most frequent segment types and their controlled frequencies by language stock

Segment	Controlled frequency (%)	Data set frequency (%)	Difference (%)
k^{h}	22.75	17.36	-5.39
p^{h}	22.59	17.36	-5.23
iː	22.19	28.01	5.82

If Maddieson's genealogically balanced quota sample used to construct UPSID is a valid predictor of the distribution of segments in the world's languages, then we would expect a genealogically stratified sample of a larger data set like PHOIBLE to concur with Maddieson's observations. One might also expect that the most frequent segment types across languages should remain relatively constant before and after stratification if the segments do indeed appear in most languages and that the language sample being probed has broad coverage.[18] In fact, there is an overlap of the eight most frequently occurring segments in both UPSID$_{451}$ and the controlled sample from PHOIBLE, shown in Table 5.2. PHOIBLE's genealogically controlled and uncontrolled segment frequencies differ by roughly plus (overrepresented) or minus (underrepresented) 5%. The segment /u/ is slightly higher at 8%. The results of resampling show that the controlled segment type frequencies in PHOIBLE line up (although not perfectly by rank) with the frequency of segment types found in Maddieson's quota sample.

Instead of looking at the most frequent segments, which show a relatively small difference between their controlled and uncontrolled frequencies, what happens if we look at the segments in PHOIBLE with the greatest overrepresentation? Continuing with the 35 most frequent segments in Table 5.1, the resampling method suggests that the frequency of segments / f, ŋ, ɔ, ɳ /, when stratified for language family, actually occur over 20% too frequently in the PHOIBLE data set. On the one hand, I suspected that some nasals would be

[18] Nearly half of all segment types found in language descriptions used in UPSID$_{451}$ and PHOIBLE occur only once in a segment inventory. Thus the most infrequent segment types are not a good place to compare data sets.

Table 5.2: Most frequent segments in UPSID$_{451}$ and a controlled PHOIBLE sample

	PHOIBLE (%)	UPSID$_{451}$ (% & rank)
i	90.56	87.10 (3)
m	90.50	94.20 (1)
k	89.92	89.40 (2)
a	88.77	86.90 (4)
p	87.16	83.20 (6)
j	83.45	83.80 (5)
w	79.43	73.60 (8)
u	78.03	81.80 (7)

overrepresented because of PHOIBLE's broad coverage of Niger-Congo languages. On the other hand, I was did not expect /f/ and /ɔ/ to stand out as outliers. Under closer inspection, however, /f/ and /ɔ/ in inventories in PHOIBLE occur most often in languages spoken in Africa. PHOIBLE contains a disproportionate number of inventories from languages spoken in Africa, which skews the frequency of segment types towards those inventories. Table 5.3 summaries these figures. The number after each geographic region indicates the total number of segment inventories in PHOIBLE for that region. The number of inventories in each geographic region that contain /f/ or /ɔ/ is given as a percentage. In this case, the genealogical resampling method led to an insight regarding the geographical skew present in the current PHOIBLE data set.

Finally, if we take a look at the most underrepresented segments from Table 5.1, i.e. those which occur less frequently in the PHOIBLE data set than what the genealogically resampling method indicates they should. These segments are: /ʔ, ʃ, o̞, e̞, ɾ/. I am not sure why /ʃ/ and /ʔ/ are underrepresented. Perhaps because the languages in North and South America are more likely to contain these sounds, but in general they are underrepresented in PHOIBLE? For the remaining three segments, there is a straightforward explanation.

Table 5.3: Frequency of segments /f/ & /ɔ/ by world region

PHOIBLE (1089)	/f/	/ɔ/
Africa (451)	71.4%	84.7%
America (248)	16.9%	23.4%
Asia (192)	39.6%	52.6%
Europe (61)	36.1%	49.2%
Pacific (137)	24.9%	27.7%

An individual linguist's transcriptions may be systematic and consistent, but linguists' transcriptions across language descriptions are not consistent with each other. One example is the use of the keyboard <a> for the low back unrounded vowel instead of the IPA <ɑ>. Another example is that linguists sometimes use the terms tap and flap indiscriminately. Those who do discriminate between the alveolar tap and alveolar flap tend to use the symbols <ᴅ> and <ɾ>, respectively. The problem with this distinction is that no language seems to contrast a tap and flap at the same place of articulation. Hence, the tap symbol <ᴅ> is not recognized by the IPA, which simply labels the manner of articulation as "Tap or Flap" and uses the symbol <ɾ>. Nevertheless, linguists may use either alveolar tap or alveolar flap in their descriptions of languages' phonological systems.[19] Thus the compilers of SPA and UPSID faced the challenge of reinterpreting original phonemic analyses from different language descriptions into a consistent standard or to keep the original analysis.[20]

[19] Ladefoged and Johnson (2010, 175-176) consider it is useful to make a distinction between taps and flaps. Although each is caused by a single contraction of muscles and two articulators making contact, a tap is made by moving the tongue tip up to the point of contact (teeth or alveolar region) and back down again. And a flap starts in a retroflex gesture (curled up and back) and then makes contact with the post-alveolar region. Therefore, the distinction between taps and flaps is somewhat bound with their place of articulation. However, note that the distinction referred to in UPSID is between alveolar taps and alveolar flaps.

[20] This is different from the approach that I have taken in this work in which I accept the linguist's analysis at face value and only reinterpret their phonetic symbols and descriptions into a consistent transcription standard for interoperability. As long as there is one language that purportedly has a contrast, I try to preserve it. Then I leverage the graph data model (discussed in Section 3.2.3) with distinctive features

In the case of UPSID, Maddieson kept the distinction between the alveolar tap and alveolar flap, although they do not co-occur (i.e. contrast) in any language in the UPSID$_{451}$ sample. In UPSID$_{451}$ there are 91 languages that have a voiced alveolar flap and seven languages with a voiced alveolar tap. These seven languages are also the only languages in the current PHOIBLE data set that contain a voiced alveolar tap, although there are a total of 234 languages that have a voiced alveolar flap.[21]

The genealogical resampling method indicates that the frequency of /ɾ/ is too low in the database. This makes sense if we consider that the voiced alveolar tap and flap should be treated as the same segment, and thus, the same symbol. In some cases the resampling technique will randomly choose a representative language from some language family that contains a language with an alveolar flap and in other cases it may choose a language from the same family that has alveolar tap. Of course the latter is much rarer, since there are only seven languages with a tap, but 234 languages with a flap. Nevertheless, if these two distinct symbols are collapsed into one, we would expect the underrepresentation of /ɾ/ to be less. And it is, as shown in Table 5.4. The impact is minimal, suggesting that the alveolar flap is still underrepresented in PHOIBLE.

A greater effect of the sort demonstrated by the tap and flap segments can be seen in the distinction made by SPA and UPSID between higher-mid vowels (/e/ and /o/) and mid vowels (/e̞/ and /o̞/).[22] These vowels fall into the mid-range vowels category (Maddieson, 1984, 123). With the exception of only five languages in SPA and UPSID$_{451}$, where there is a reported phonemic contrast between /e/ and /e̞/ or /o/ and /o̞/, this division of the

(discussed in Chapter 6) to encode these segments, so that users can query on a selection of features that leave aspects of the segment underspecified. In this manner it is not the linguist's analysis that is reinterpreted; it is stored as given by them in the database. It is the data structure underlying one view of the data that allows the user to manipulate the underlying data via the query.

[21] These include 71 records from SPA, which did not make a distinction between taps and flaps.

[22] There are no IPA letters that make a distinction between higher-mid and mid vowels. I chose to mark the mid-vowels with the lowered diacritic and to leave the higher-mid vowels unmarked because it followed the approach taken in UPSID$_{451}$: <"e> and <"o> denote mid and <e> and <o> higher-mid. Note that in both front and back vowels, there are more occurrences of mid than higher-mid in the inventories in UPSID$_{451}$. In SPA this is confusingly the other way around – there are more mid vowels than higher-mid vowels. UPSID borrowed heavily from SPA and therefore it includes many of the same inventories. When there exists a mid or higher-mid vowel in UPSID, I encoded the corresponding higher-mid and mid vowels equivalently in SPA, even though SPA labels its distinctions as simply "e" versus "mid-e".

Table 5.4: Controlled frequencies of segments

Segment	Controlled frequency (%)	Data set frequency (%)	Difference (%)
ɾ	30.52	21.76	-8.76
ɾ & ᴅ	30.76	22.31	-8.45
e	41.14	59.41	18.28
e̝	23.97	14.97	-9.00
e & e̝	65.11	74.38	-9.27
o	42.67	61.16	18.49
o̝	28.43	16.25	-12.17
o & o̝	70.84	77.41	-6.57

mid-range vowel space for front and back vowels splits languages into two large groups – those that have a front and/or back mid-vowel (/e/ and/or /o/) and those that have a front and/or back higher-mid vowel (/e̝/ and/or /o̝/). The number of languages that have an /e/ in SPA and UPSID$_{451}$ are 42 and 170, respectively. The number of languages with /e̝/ in SPA and UPSID$_{451}$ are 61 and 125. Out of all these languages, there is only one language in SPA (Lahu [lhu]) and two languages in UPSID$_{451}$ (Lahu [lhu] and Klao [klu]) that have a phonemic contrast between /e/ and /e̝/. Further, the number of languages that have /o/ in SPA and UPSID$_{451}$ are 73 and 131. And the number of languages containing /o̝/ in SPA and UPSID$_{451}$ total 47 and 181. Of these, only Bengali [ben] and Telugu [tel] in SPA and Breton [bre] and Klao [klu] in UPSID$_{451}$ contrast /o/ and /o̝/.

Casting the mid-range vowel space into two vertical dimensions and distributing them across a large number of mutually exclusive languages causes /e/ and /o/ to be overrepresented in PHOIBLE and /e̝/ and /o̝/ to be underrepresented. Table 5.4 shows the lower difference when the resampling method is rerun with each mid and higher-mid pair is treated as one symbol. Collapsing this distinction also makes linguistic sense (except for those handful of languages that phonemically contrast mid and higher-mid vowels), since we would expect many languages to make use of the mid-range vowel plane.

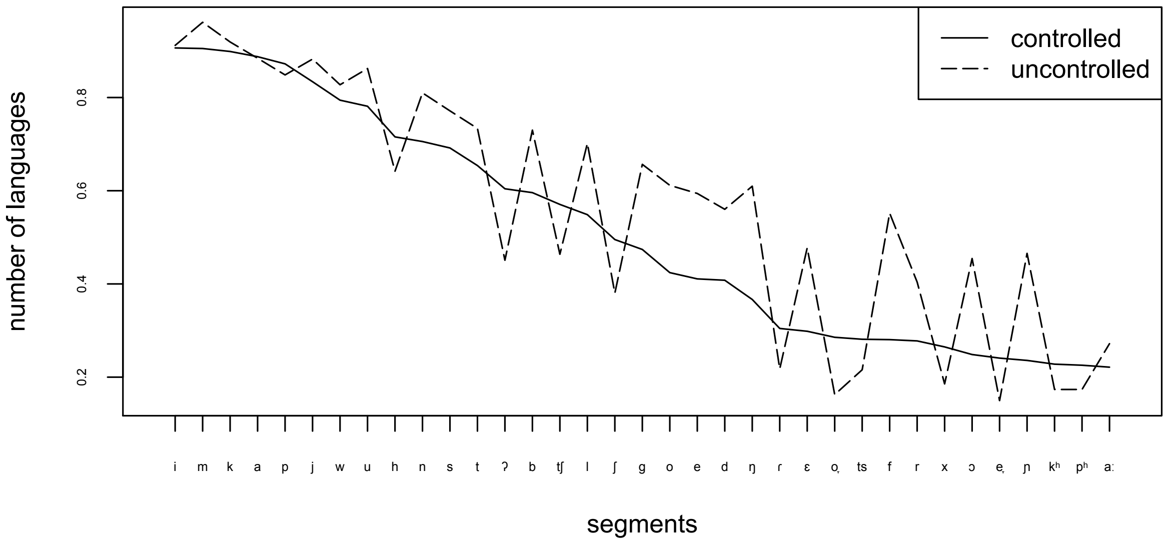
In summary, in this section I have examined the distribution of segment types in PHOIBLE and have shown that as the number of segment inventories increases, the number of segment types seems to be increasing in a quadratic curve with no asymptote in sight. I also investigated the frequency of segment types in PHOIBLE by implementing a resampling method that estimates genealogical bias. The resampling technique lets us infer the probable distribution of segment type frequencies by repeatedly sampling a random language representative from groups such as language families and systematically recomputing a statistical estimate by randomly sampling from subsets within a data set.

The most extremely overrepresented segments occur often in inventories of languages spoken in Africa, which is expected because PHOIBLE is genealogically skewed towards broad coverage of Africa. This is partly due to the inclusion of the 203 segment inventories from Alphabets of Africa (Hartell, 1993; Chanard, 2006). On the other hand, some fairly underrepresented segments are due to differences in phonemic analysis and factors of data collection. Lastly, segments that are not very overrepresented or underrepresented in the sample coincide with the frequency of segments found in UPSID$_{451}$, e.g. the most frequent segments in both PHOIBLE and UPSID$_{451}$ and segments like /p^h/ and /k^h/ that appear with nearly the same frequency in the genealogically controlled PHOIBLE sample (22.59% and 22.75%) and UPSID$_{451}$ (22.4% and 22.8%). A plot of the 35 most frequent segments controlled for genealogical factors via the resampling technique and their uncontrolled frequencies in PHOIBLE is given in Figure 5.3

5.4 Segment inventories

In this section I review some of the typological facts put forth by research undertaken with SPA (Crothers et al., 1979), UPSID$_{317}$ (Maddieson, 1984, 1986; Lindblom and Maddieson, 1988), UPSID$_{451}$ (Maddieson and Precoda, 1990; Maddieson, 1991) and WALS (Maddieson, 2008a,c,b). I present these data within a historical perspective by comparing the SPA, UPSID$_{451}$ and PHOIBLE inventories. I then apply the genealogical stratification method, discussed in the previous section, to the PHOIBLE data set. I show that Maddieson's findings generally still hold even as the size of a segment inventory databases increases.

Figure 5.3: Controlled and uncontrolled segments plotted against the number of languages they appear in

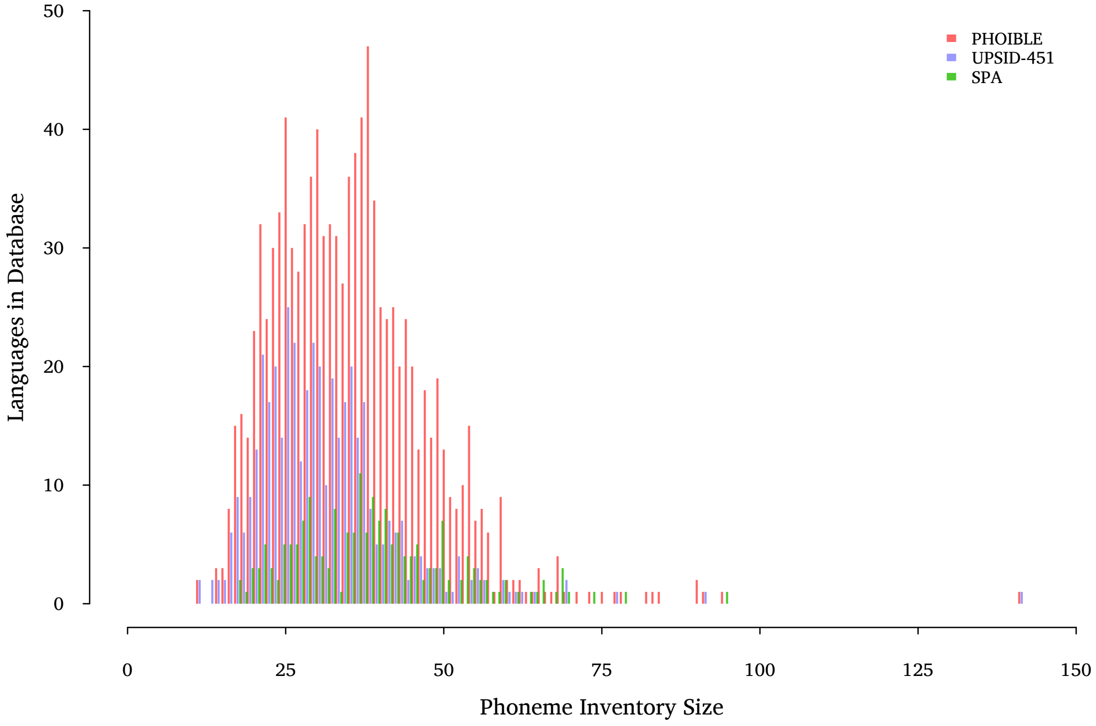


5.4.1 Inventory size

The size of phonological segment inventories varies widely, ranging from 11 to 141 total segments. This range was documented in the UPSID_{451} sample and still holds in the current PHOIBLE sample of 1089 languages. A histogram of phoneme inventory sizes offset with the contents of PHOIBLE, UPSID_{451} and SPA is given in Figure 5.4.

The smallest known segment inventories belong to Rotokas [roo] (North Bougainville; Papua New Guinea; Firchow and Firchow 1969b) and Pirahã [myp] (Mura; Brazil; Everett 1982; Rodrigues 1980). Each has only 11 contrastive sounds; both share /p, k, g, i, o̪, a/. However, Everett reports that Pirahã, as spoken by women, has 10 phonemes because /s/ is lacking; the phoneme /h/ is used instead, although not entirely consistently. Additionally, if tone is taken into account, the inventory size of Pirahã increases by two, and thus has either 12 or 13 total phonemes, depending on the gender of the speaker.

Figure 5.4: Histogram of phoneme inventory sizes in PHOIBLE, UPSID_{451} and SPA



The largest known segment inventory belongs to !Xũ [ktz] (Khoisan; Botswana), also known as !Xoon or !Xóõ, which has 141 segments (Snyman, 1970, 1975). As discussed in Sections 2.3.3 and 2.3.4, the size of a segment inventory is partially determined by the phonological theory and phonemic principles applied to an analysis of a particular language or dialect. Mielke (2009) reports the number of distinctive segments in !Xóõ at 160, based on an analysis of East !Xoon by Traill (1985).[23] Members of the DoBeS Taa project have analyzed the western dialect of Taa (West !Xoon) as having 164 segments (including 85-87 consonants and 43 clicks), making it the largest documented segment inventory to date.[24] However, Naumann (forthcoming) applies a cluster analysis to the Taa data, which sub-

[23] Mielke also notes that Central Rotokas in UPSID is distinct from the Aita dialect of Rotokas, the latter has more segments as described in Robinson 2006.

[24] http://www.mpi.nl/DOBES/projects/taa/project

stantially reduces the consonant inventory from 161-164 to 85-87.[25] This puts Taa much closer to the upper end of languages with very rich segment inventories, so that it is not so much an outlier as shown in Figure 5.4. Another example of the effect of different analyses is shown by the total number of segments in Hindu-Urdu [hin] as described in SPA (total of 94) and UPSID$_{451}$ (61). The former analysis includes geminates in the inventory, the latter does not.[26]

Half of all languages surveyed in UPSID$_{317}$ have between 11-28 consonants and vowels, and the other half have 29 or more. Thus the median inventory size in UPSID$_{317}$ is between 28 and 29 (Maddieson, 1984, 7). The mean segment inventory size is a little of over 31 and 70% of languages fall between 20 and 37 segments. In the expanded UPSID$_{451}$ data set, the mean inventory size rises to 30.97 and the mean is 29 segments. These values are close to Hockett's estimation that the average number of segments in languages is 27 ± 7 (Hockett, 1955; Maddieson, 1984).

In the entire genealogically uncontrolled PHOIBLE sample, the mean number of segments is 35 segments per language.[27] The median inventory size is 34 segments. In comparison to UPSID$_{317}$, only 58% of languages fall between 20-37 segments. Fifty percent of all languages in the PHOIBLE sample fall between 26 and and 41 segments. These results fall at the edge of Hockett's estimate.

When I apply the genealogical stratification sampling method to PHOIBLE, I get the figures provided in Table 5.5. This method takes into account the estimation errors of the data set by randomly sampling within language family stock and summing together segment inventory sizes and taking the mean by dividing by the number of language family stocks and then iterating this method a given number of times. For segment inventory size I ran two experiments: one with 1000 iterations and the other with 50,000 iterations.[28] The

[25] The cluster analysis classifies clicks as accompaniments with segments. It was initially suggested by Traill (1985) and Naumann's analysis builds on the work of Güldemann (2001) and Nakagawa (2006).

[26] However, note the comment in the UPSID$_{451}$ data: "All (or almost all) consonants appear geminate". See: `http://web.phonetik.uni-frankfurt.de/L/L2016.html`.

[27] These figures only take phonemes into account and not allophones.

[28] Using R64 on an iMac 2.7 GHz Intel Core i5 with 4GB of RAM, this process takes about an hour for 50,000 iterations.

figures are quite similar. The genealogically stratified mean is 31.6 and the median is also 31.6.

Table 5.5: Summary of average number of total segments using genealogical resampling

	1000x	50,000x
Min.	29.49	29.28
1st Qu.	31.17	31.18
Median	31.62	31.66
Mean	31.66	31.69
3rd Qu.	32.11	32.18
Max.	33.97	35.20

Figures 5.5 and 5.6 are density plots that show the weight of the probability mass from the results of genealogical stratification resampling. The higher the density, the larger the likelihood that the corresponding value in the x-axis will be selected. Within the randomized data and aside from genealogical influences, there seems to be a true average segment inventory size in the data, which can be seen in the curve of the density plots; they are roughly symmetrical and normally distributed.[29] If the data were actually very diverse, the results would not show a roughly normal distribution. Thus one might assume that there is a tendency for languages to converge on an optimal segment inventory size. However, there may be another explanation for this convergence. There may be no optimal inventory size, but instead there are simply multiple data points with the same frequency in the data set. Note that it is not the variation of segment inventories under observation, but their average size. Thus when controlling for genealogical factors in this way, if one picks a random language it is likely to have 31-32 segments.

[29] In future work I would like to investigate whether there exists a maximally optimal size to which segment inventories gravitate, or if an optimal size is influenced by other factors, such as language family.

Figure 5.5: Density plot of the average number of segments (1000x)

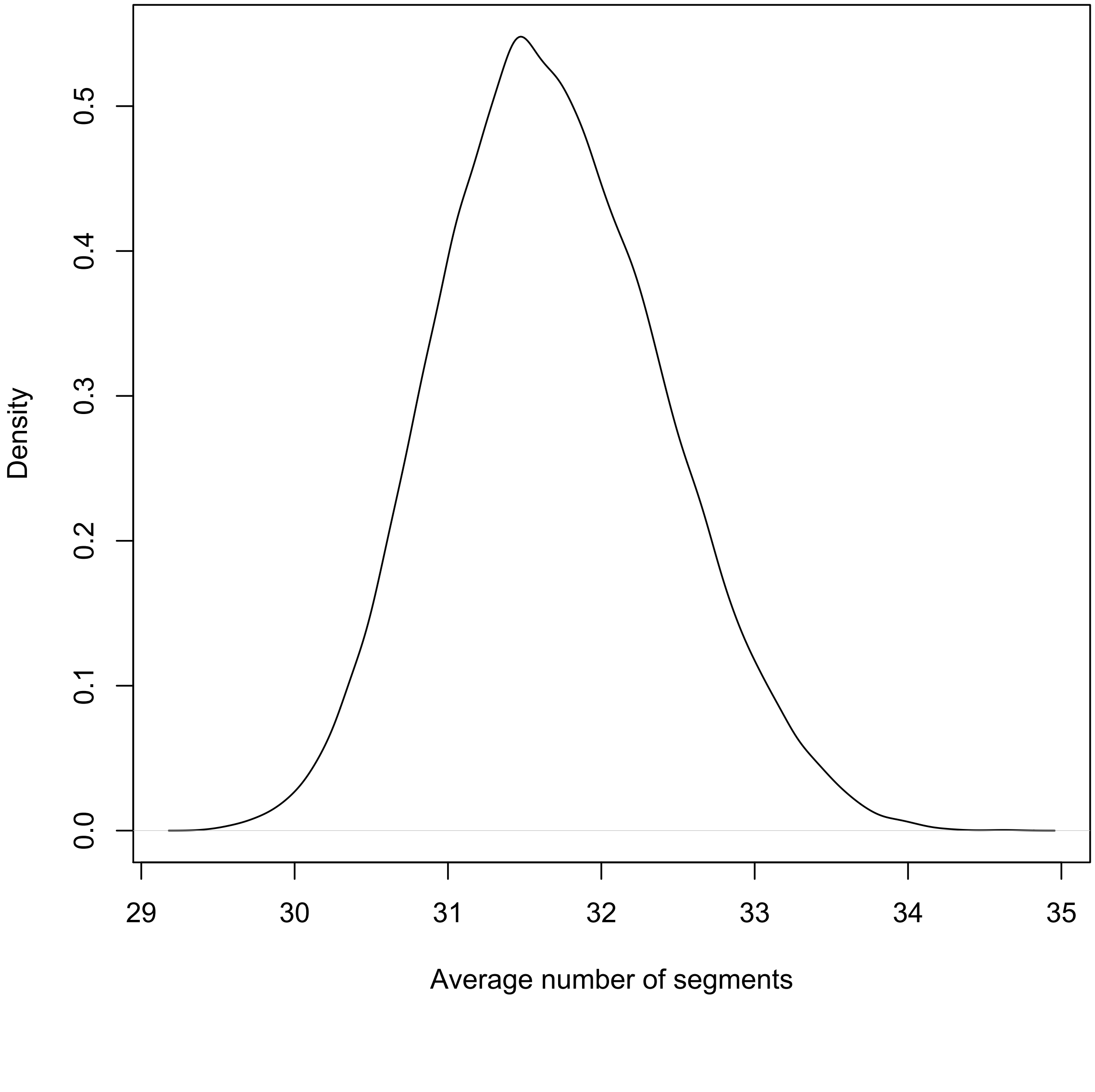
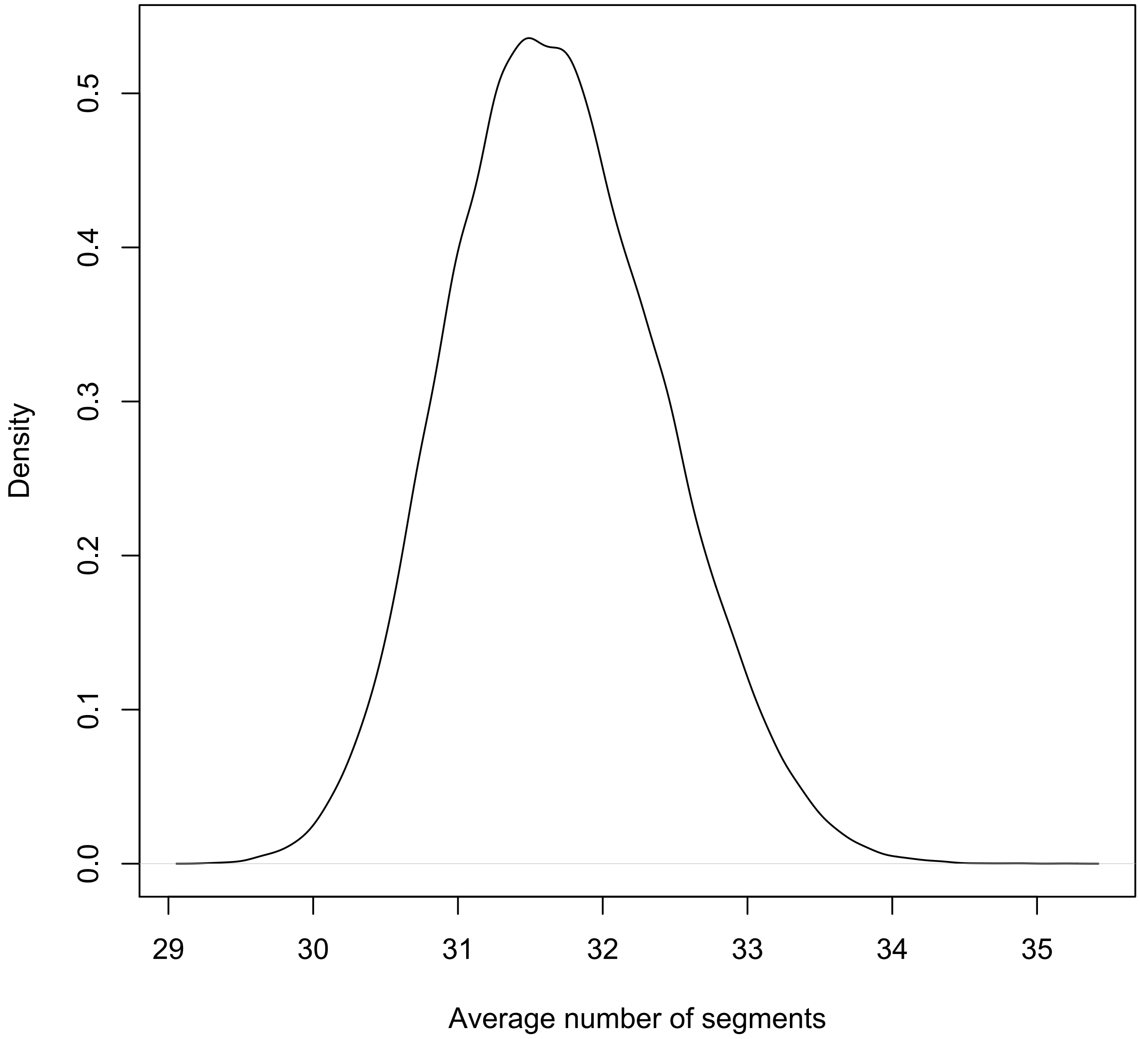


Figure 5.6: Density plot of the average number of segments (50,000x)

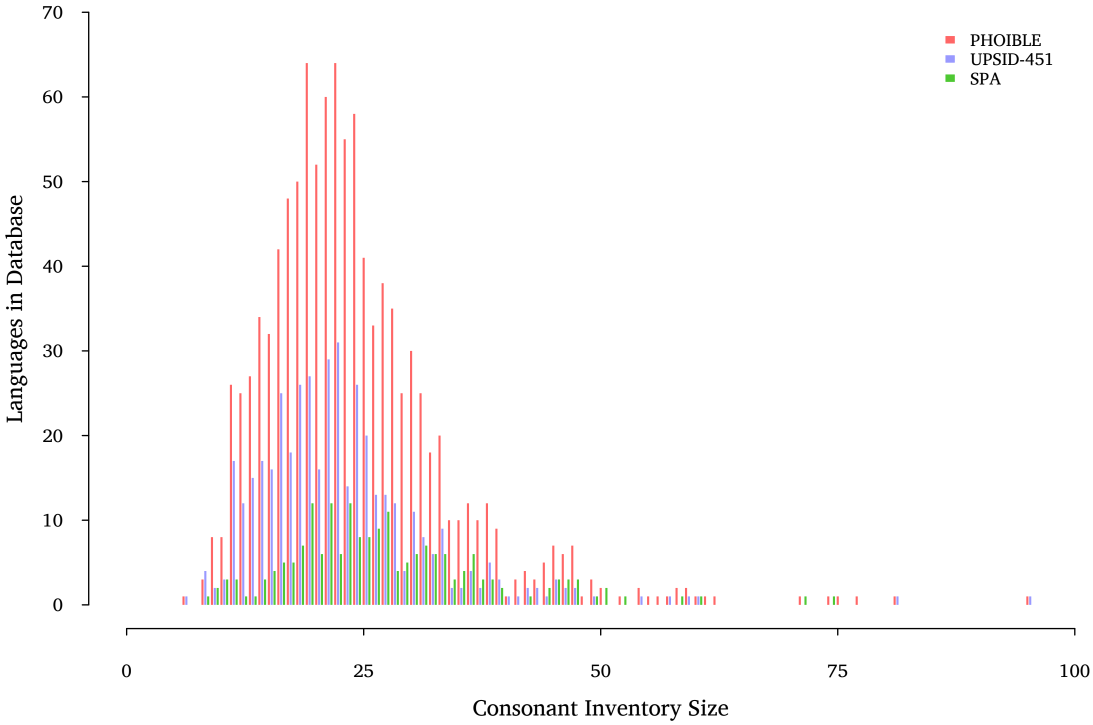


5.4.2 Consonants

If we examine the contents of inventories at a finer grained level, consonants in UPSID_{317} range between 6 and 95 segments (Rotokas and !Xũ, respectively) with a mean of 22.8 (Maddieson, 1984, 9). The range is unchanged by the inclusion of more segment inventories in both the UPSID_{451} and the PHOIBLE samples, both of which are also bounded by Rotokas and !Xũ. UPSID_{451} has a slightly lower mean for consonants at 22.45, with a median of 21.

Before genealogically stratifying the PHOIBLE inventories, the mean number of consonants is slightly higher at 24. Figure 5.7 is a histogram of the consonant counts for PHOIBLE, UPSID_{451} and SPA.

Figure 5.7: Histogram of consonant inventory sizes in PHOIBLE, UPSID_{451} and SPA



In the larger sample size of 562 languages in WALS, Maddieson (2008a) states that typical consonant inventory size is in the low twenties and that the mean of the sample is

22.7 and the median is 21. Although the specific consonant counts per language are not provided, Maddieson categorizes the average inventory as 22 ± 3, with the other categories divided into large ($\geq$ 34), moderately large (26-33), moderately small (15-18) and small (6-14).[30]

To the consonant inventories in PHOIBLE, I applied the genealogical resampling technique and ran 50,000 iterations, randomly choosing a representative language from each language family stock. A summary of the frequencies by quartiles, median and mean is given in Table 5.6.

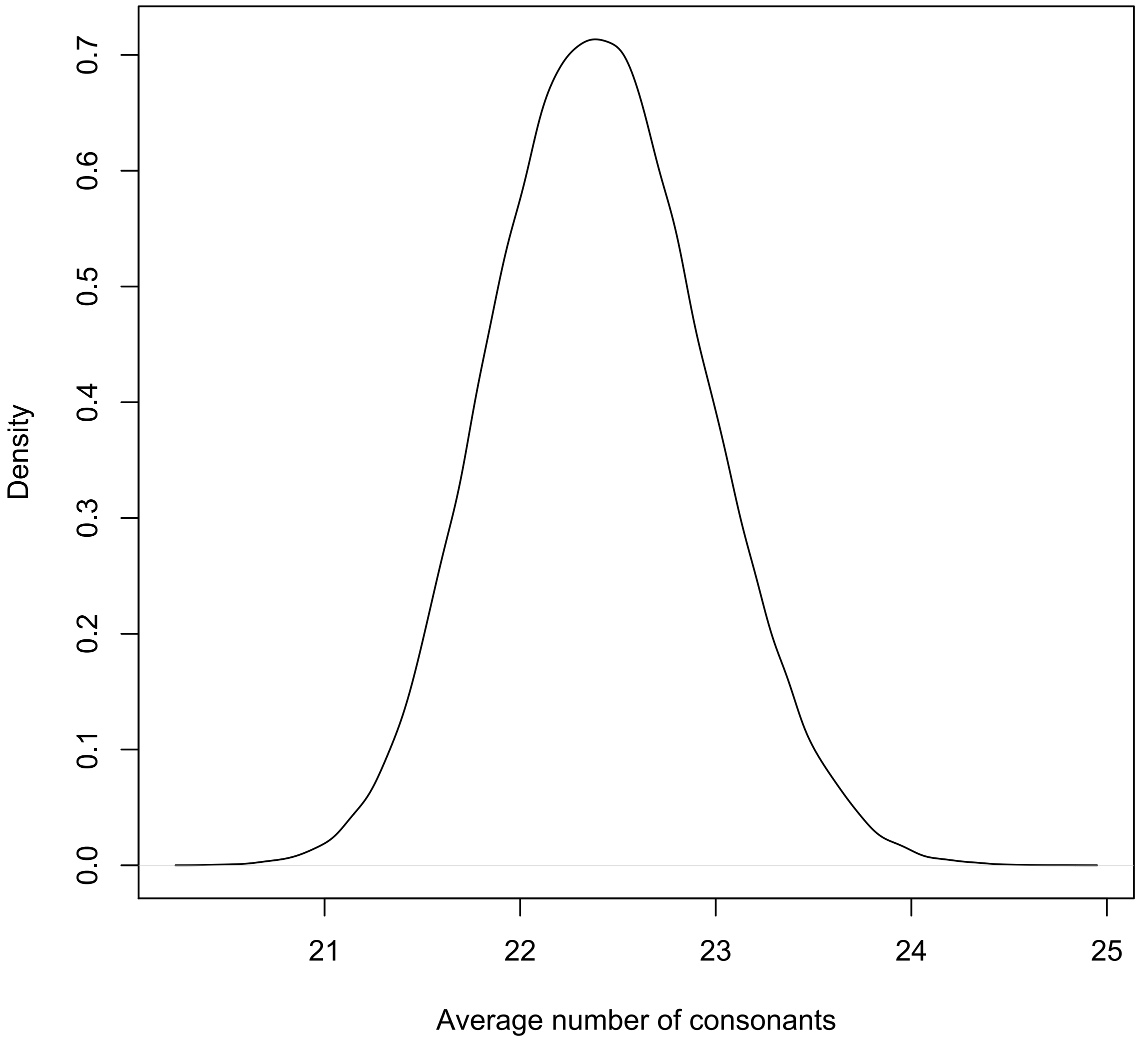
Table 5.6: Summary of average number of consonants using genealogical resampling

Min.	20.41
1st Qu.	22.02
Median	22.40
Mean	22.40
3rd Qu.	22.77
Max.	24.78

The 1st and 3rd quartiles are almost symmetrical around the mean, which is 22.4 and nearly the same as the mean consonant inventory size in UPSID$_{451}$. The median of the stratified sample is slightly higher, also at 22.4. Although the measures are not perfectly normally distributed, they are nearly symmetrical and there is a clearly pronounced mean, as shown in the density plot given in Figure 5.8.

[30] See: http://wals.info/chapter/1.

Figure 5.8: Density plot of the average number of consonants in inventories



5.4.3 Vowels

In UPSID_{317}, the number of vowels in a segment inventory ranges from 3 to 46 with a mean of 8.7 (Maddieson, 1984, 9). The range remains the same in UPSID_{451} and the mean is 8.5 and median is 7. In WALS these figures are calculated without the non-quality distinctions of vowel length, vowel nasalization and diphthongs (Maddieson, 2008c). Therefore the maximum number of vowels across languages drops to 14 (in German) and the overall average is fractionally below 6. The WALS sample also provides an increased sample size of 559 languages, roughly a quarter more languages than in UPSID_{451}. The increase in typological coverage results in four languages being included that only have two contrastive vowels.[31] Under one phonological interpretation, only two contrasting vowel qualities are employed in these languages.[32] In the PHOIBLE data set, two languages contain only two contrastive vowels: Zulgo [gnd] (from the AA sample) and Cuvok [cuv] (from the PHOIBLE inventories). Yimas [yee] and Abaza [abq], in WALS, are not among the segment inventories in the PHOIBLE sample; the inventories of Kabardian provided by SPA and UPSID_{451} both list 7 vowels (Crothers et al., 1979; Maddieson and Precoda, 1990). Figure 5.9 provides a histogram for the vowel inventory counts in PHOIBLE, UPSID_{451} and SPA.

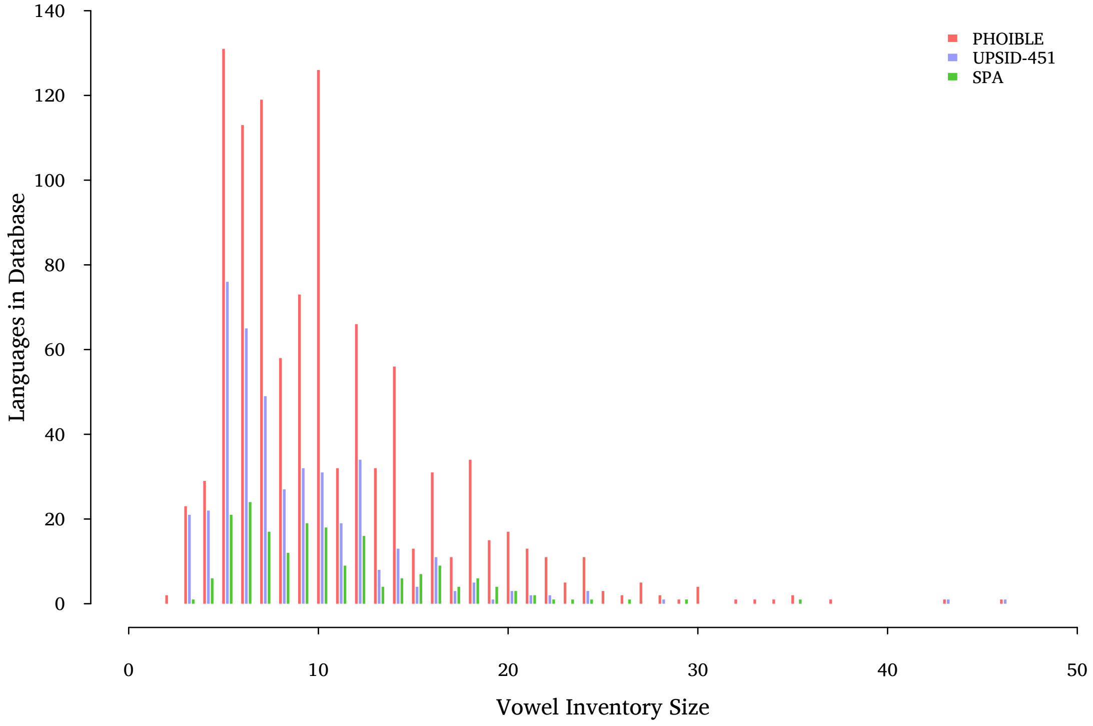
In the UPSID_{451} sample, languages most often have a five vowel system. This tendency is also noted in the WALS sample, in which it is reported that over 1/3 of the languages (188/559) have a five vowel system (Maddieson, 2008c). The next most frequent vowel system in the WALS sample is the six vowel system (17.8%).

Although in the SPA sample, a six vowel system is the most prevalent, the distribution of vowel inventories curves into a long tail like the UPSID_{451} sample. On the other hand, the distribution of vowels in the overall PHOIBLE sample does not present a nice curve. The PHOIBLE sample shows a ten vowel system to be most prevalent, followed closely

[31] Although two vowels analyses do not appear in UPSID_{317}, Maddieson noted that Kabardian [kbd] (Caucasian; Russia) and Abaza [abq] (Caucasian; Russia) had been analyzed elsewhere as having fewer than three vowel phonemes (Maddieson, 1984, 126).

[32] Yimas [yee] (Lower Sepik-Ramu; Papua New Guinea) is the only example mentioned in the text of Maddieson 2008c in WALS (Haspelmath et al., 2008). Identifying the languages that contain just two phonemic vowels is not currently possible because data in WALS is divided into categories of small, average and large (consonant, vowel, tone) inventories and not as individual figures on a per language basis.

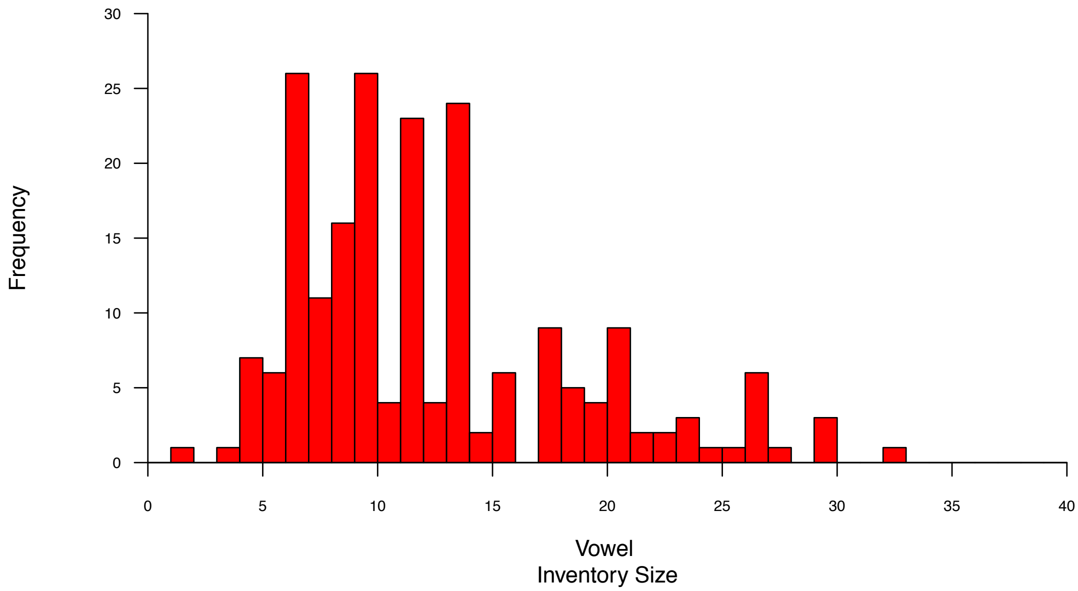
Figure 5.9: Histogram of vowel inventory sizes in PHOIBLE, UPSID_{451} and SPA



by a five vowel system. This difference in distribution is due to the fact that PHOIBLE subsumes a large number of African languages. For example, Figure 5.10 shows a histogram of the distribution of vowel inventory sizes in just the AA data set (Hartell, 1993; Chanard, 2006). It shows that the majority of the 203 language sample have either seven, ten, twelve or fourteen vowel systems.[33]

[33] Dan McCloy (p.c.) suggests this could have something to do with maximal dispersion or a preference for symmetry, e.g. perhaps a language that adds a lax version of a high front vowel is likely to add a lax high back vowel.

Figure 5.10: Distribution of vowel inventory sizes in AA



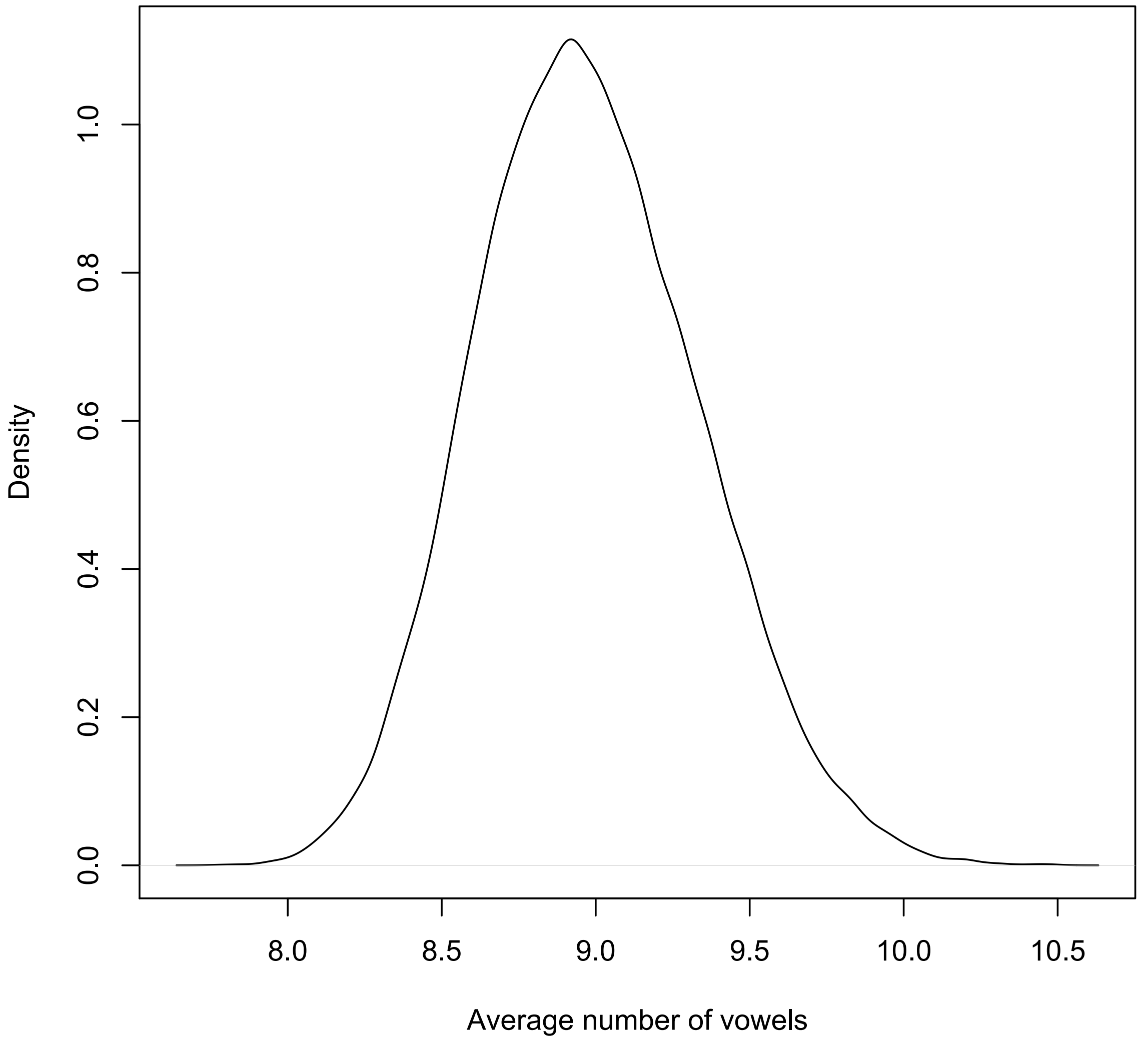
I applied the genealogical resampling technique to the vowel inventory data in PHOIBLE. I ran the sampling method for 50,000 iterations over language family stocks of which there are 96 for this experiment. A summary of the results is given in Table 5.7.

Table 5.7: Summary of average number of vowels using genealogical resampling

Min.	7.750
1st Qu.	8.719
Median	8.958
Mean	8.977
3rd Qu.	9.219
Max.	10.521

The mean number of vowels after genealogical stratification is 8.97, slightly higher than both the UPSID$_{317}$ (8.7) and UPSID$_{451}$ (8.5) data sets. The median vowel inventory size is 8.95, greater than the 7 in UPSID$_{451}$. Figure 5.11 shows the density plot of the average number of vowels in inventories. Again, the curve is roughly normally distributed and there is a clearly pronounced mean.

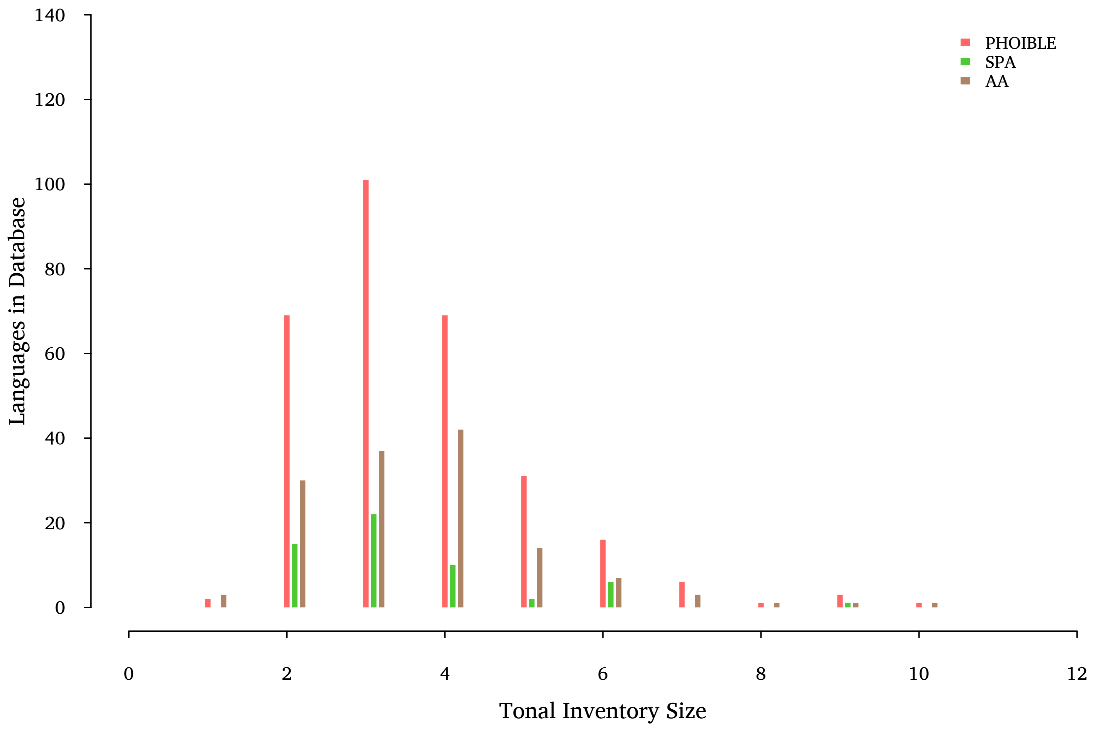
Figure 5.11: Density plot of the average number of vowels in inventories



5.4.4 Tone

Tone data, but not stress, is available in the SPA, AA and PHOIBLE inventories.[34] However, there is no comparable data from UPSID$_{451}$ because the inventories do not contain suprasegmentals. Starting with the current 1336 segment inventories in PHOIBLE and removing UPSID$_{451}$ leaves 885 inventories. Applying the trump hierarchy to these remaining inventories leaves 808 distinct languages. Of those 808 distinct inventories, 302 have tone, so slightly over 37%. Descriptions of these languages range in their number of tones from 1-10 and the mean number of tones per language is 3.5. Figure 5.12 show the distribution of tones in the inventories in PHOIBLE, SPA and AA.

Figure 5.12: Histogram of tone inventory sizes in PHOIBLE, SPA and AA



A comparison of the WALS and PHOIBLE samples with regard to tone is of little value.

[34] Unfortunately, languages with minimal pairs for stress are given the short shrift because stress seems to be rarely described as a phonemic contrast in language descriptions.

In the WALS 526 language sample, 220 languages are tonal (41.8%) (Maddieson, 2008b). However, Maddieson notes that this figure probably underrepresents the proportion of tonal languages because the sample is not proportional to the density of languages in geographic areas that contain languages with tone. Likewise, the PHOIBLE sample is geographically skewed and no effort was taken to gather a representative sample of tonal languages, which are concentrated in places like sub-saharan Africa, Southeast Asia, Papua New Guinea and scattered throughout the Americas.[35] In WALS and other work, Maddieson (2007, 2008b) investigates relationships between phonological properties like the number of consonants and vowels, syllable structures and simple and complex tone systems. I leave reevaluating these findings with PHOIBLE's data set for future work.

If we inspect the current types of reportedly contrastive tonal segments in the PHOIBLE data set, we get the following tones given in Table 5.8. High and low tones occur in equal numbers across languages in the sample. Mid tone is the next most frequent, followed by the contour tones HL (high-low) and LH (low-high). Rarer combinations follow.

5.4.5 Summary

In summary, in Sections 5.3 and 5.4 I investigated the distribution of segment types and the distribution of segment inventory sizes and their consonant, vowel and tone compositions. I applied a genealogical stratification technique with randomized data at the language family stock level to account for genealogical influence. As I have shown, the mean and median figures from the genealogically stratified PHOIBLE sample are similar to those given by Maddieson through his work with the UPSID$_{451}$, UPSID$_{317}$ and WALS samples.

5.5 Consonants and vowels

One area of typological interest in segment inventories is the balance between consonants and vowels across inventories. This may be partly driven by the assumption that all languages are equally complex (cf. Miestamo et al. 2008; Sampson et al. 2009), so investigating the distribution of consonants versus vowels might provide some insight into how languages'

[35] See Table 5.11 on page 250 for a geographic breakdown.

Table 5.8: Simple and complex tones in the PHOIBLE sample

Description	Symbol	Count
High	˦	129
Low	˨	129
Mid	˧	71
High-Low	˦˨	50
Low-High	˨˦	47
Mid-Low	˧˨	11
Low-Mid	˨˧	6
Extra-High	˥	6
High-Mid	˦˧	3
Mid-High	˧˦	3
Falling	˥˩	3
Rising	˩˥	3
Downstep	↓	2
Downstep-Extra-High	↓˦	1
High-Low-High	˦˨˦	1
Low-High-Low	˨˦˨	1

phonological systems vary to compensate for complexity in different subsystems.[36] This process is known as the compensation hypothesis, i.e. that a simplification or complication in one area of an inventory will be counterbalanced by the opposite somewhere else (Martinet, 1955). For example, Maddieson (1984, 21) examined suprasegmentals (tone and stress) in languages in UPSID$_{317}$ and reported that the "overall tendency appears once again to be more that complexity of different kinds goes hand in hand, rather than for complexity of one sort to be balanced by simplicity elsewhere". In order to answer the question of compensation, some method for measuring complexity is needed for empirical evaluation

[36] For a thorough review of issues and approaches to phonological complexity, see Pellegrino et al. 2009.

of the phonological system. The size of a phoneme inventory is one viable target. However, there is no agreement on how to measure the underlying probability distribution of typological variables (Cysouw, 2010). For instance, to describe phoneme inventory size, gamma (Lehfeldt, 1975) and log-normal distributions (Justeson and Stephens, 1984) have been proposed. Maddieson (2008a) also hints at a normal distribution (Cysouw, 2010, 30).

It is not my intent here to develop a complexity measure to describe phoneme inventory size (or even more ambitiously, to develop one for phonological systems).[37] However, I do want to illustrate, in some measurable and replicable fashion, the distribution of consonant and vowel inventories in the current PHOIBLE sample. Figure 5.13 shows a scattergram of languages by the number of consonants and number of vowels in their inventories.[38] Darker colored points represent overlapping languages that contain the same consonant and vowel ratio.

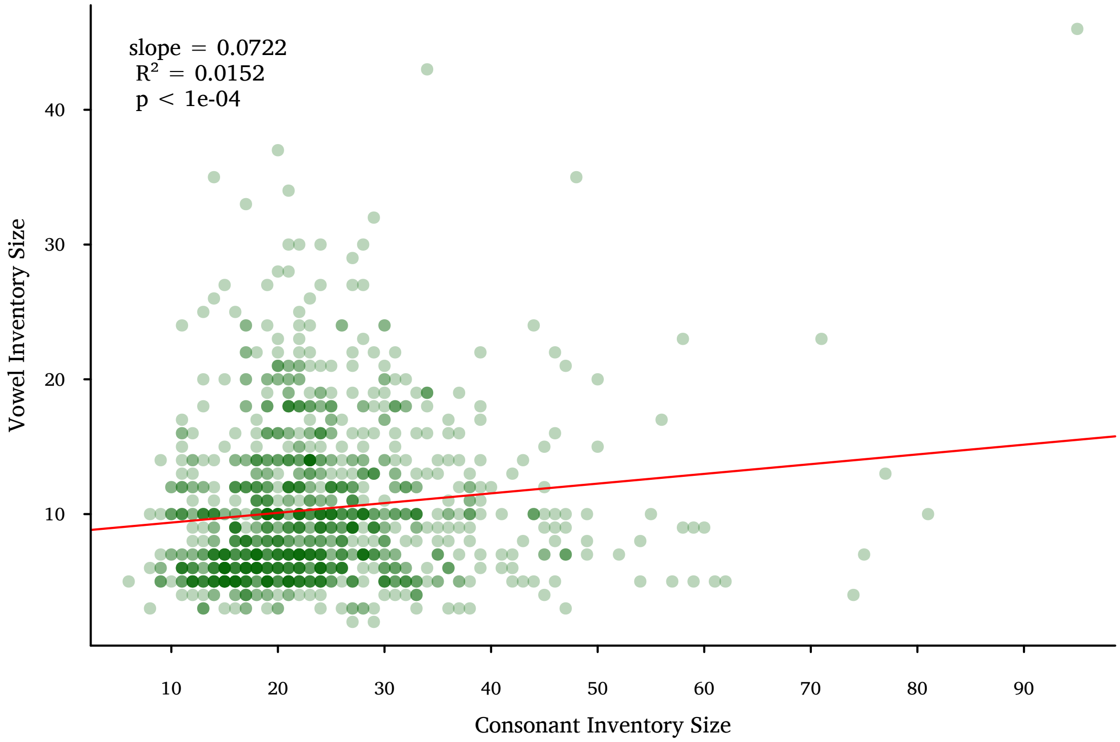
An early investigation into the purported correlation between consonant and vowel inventory size is given in Justeson and Stephens 1984. The authors come to the conclusion that there is no correlation between the number of consonants and the number of vowels in languages of the world based on a genealogically stratified sample of 50 languages. Calculating consonant and vowel ratios with the PHOIBLE segment inventory data shows that for each increase in roughly 13 or 14 consonants there is an increase in one vowel (slope = 0.0738, R2 = 0.0143, $p < .0001$). The p-value suggest that the hypothesis is robust, but the correlation is weak, if it is even reliably there.

What we know is that certain aspects of phonological complexity may not be captured by simple consonant and vowels counts (cf. Shosted 2006). There may be some other aspect of phonology driving simplification or complication. How to measure linguistic complexity is an active area of current research; see for example work in Miestamo et al. 2008 and Sampson et al. 2009 (in particular Deutscher 2009). McWhorter (2001, 135) discusses complexity of phoneme inventories and defines their complexity through markedness of segment types, i.e. a phonemic inventory is more complex than another inventory if it has more marked

[37] The reader is referred to work on phonological complexity in Maddieson 2006; 2007 and typological complexity in Cysouw 2005; 2010.

[38] Note that the aspect ratio is not perfectly square, so the line looks steeper than it actually is.

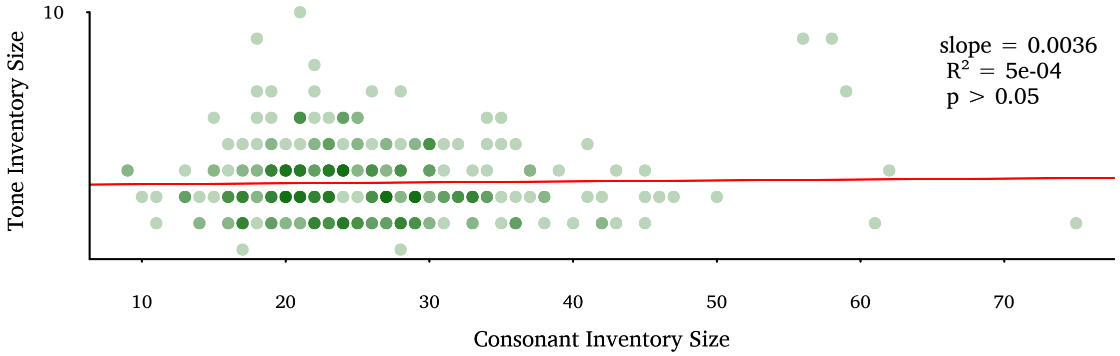
Figure 5.13: Scatterplot of the number of consonants and vowels per inventory



segments. Marked segments are calculated by their crosslinguistic distribution. McWhorter develops an empirical approach that allows him to measure the complexity of two linguistic systems and compare them to support his thesis that creoles have the world's simplest grammars. Phonetic similarity metrics have also been proposed for measuring the distance between phones for comparability purposes, e.g. Frisch 1997, Kondrak 2003, and Mielke 2004. In regard to measuring complexity in linguistic (sub)systems and addressing the compensation hypothesis, an important consideration is that a replicable empirical approach be taken to evaluate the differences among languages. Ideally, the approach also attempts to find a reason for the particular distribution. For example, Justeson and Stephens (1984) claim that the probability distribution that describes both consonant and vowel inventories is log-normal. Their argument is that phoneme inventories are rooted in distinctive phonological features. Given a set of n distinctive features, the phoneme inventory is maximally bound to 2^n phonemes.[39] If feature inventories are normally distributed, then the logarithm of phoneme inventory size is also normally distributed. Thus the authors argue that the probability distribution of phoneme inventory size is rooted in phonological factors.

The PHOIBLE data shows a weak correlation between the number of consonants and vowels. Figures 5.14 and 5.15 show scatterplots of the number of consonants, and vowels, versus the number of tones per inventory. Both plots show, at least in the current PHOIBLE data set, that there is no correlation between the number of consonants and tones in languages, nor is there a correlation between the number of vowels and tones in languages. These data provide just a preliminary study into issues of inventory complexity, but the PHOIBLE data set includes much more information, both linguistic and non-linguistic, to further explore these issues. In Chapter 7, I will look at the purported correlation between phoneme inventory size (which has been used as a measure of phonological complexity) and population size.

Figure 5.14: Scatterplot of the number of consonants and tone per inventory

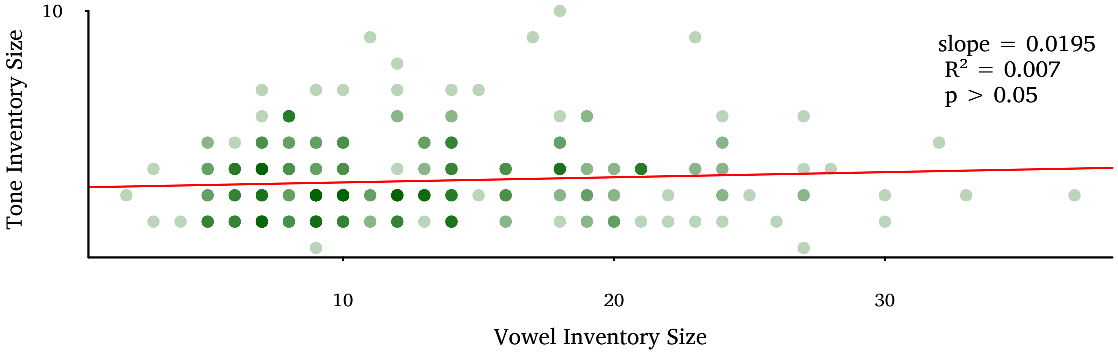


5.6 Implications in vowels systems

Pioneered by Greenberg (1963), the implicational universal is a tool used by linguists to express typological generalizations of the sort: if a language has x, then it has y. An

[39] Clements (2003a,b) terms this phonetic feature principle *feature bounding*.

Figure 5.15: Scatterplot of the number of vowels and tones per inventory



implicational hierarchy consists of a chain of implicational universals (Croft, 1990). An early example for segment inventories is the vowel hierarchy by Crothers (1978, 133).[40]

Based on the segment inventory data in UPSID_{317}, Maddieson (1984, 13-14) gives a list of implicational hierarchies for segment inventories, but as he notes, few are without exception in his data set:[41]

1. /k/ does not occur without /t/ (one exception).

2. /p/ does not occur without /k/ (four exceptions).

3. Nasal consonants do not occur unless there are stops or affricates at the same place of articulation (five exceptions).

4. Mid vowels only occur when high and low vowels also occur (two exceptions).

5. Voiceless nasals and approximants only occur when a language has their voiced counterparts.

[40] An illustration of Crothers's vowel hierarchy is given on page 176. In Section 6.5, I use PHOIBLE to test several proposed descriptive universals of phonological systems.

[41] Exceptions are listed in Maddieson 1984, 13-14.

6. Rounded front vowels only occur with unrounded front vowels of the same basic height (two exceptions).

In general, the problem with stipulating implicational universals is that the (exceptionally high or low) frequency of a given phenomenon in a sample is not necessarily indicative of anything. It is the deviation from the statistical expectation and not absolute number of occurrences that is relevant (Cysouw, 2003).[42] For example, in the PHOIBLE data set /m/ occurs 1047 times in 1089 unique segment inventories, so 95%. Additionally, /n/ occurs 883 times across the same set of unique inventories (although not necessarily in the same set of languages that /m/ occurs), so 80%. So taken together, the frequency of occurrence of /m/ and /n/ is maximally roughly 76% (.95 * .80). However, because both /m/ and /n/ occur very frequently in languages, the significance of their occurrence together tell us very little about the probability of /n/ given /m/ and vice versa. Simply, the conditional probability of /n/ given /m/, and vice versa, is not a good measure of whether /m/ (or /n/) is an interestingly good predictor of /n/ (or /m/). To garner statistical significance of implicational co-occurrences, some type of different approach is needed.

Multidimensional scaling (MDS) is a collection of statistical methods that are often used for data analysis and to visualize similarities and dissimilarities of the underlying structure of relations between entities (Borg and Groenen, 2005). MDS starts with a distance matrix and plots locations according to a proximity measure for variables in an N-dimensional geometric space. The visualization shows the distance of entities in the structure of the data. A stress majorization function is used in the reduction of the n-dimensional space into two dimensions.

Table 5.9 shows a small portion of a distance matrix between segment inventories in PHOIBLE. It was created by calculating the Jaccard index (or Jaccard similarity coefficient) between sets of segment types. The Jaccard index measures the similarity of two sets by dividing the intersection size by the size of the union (Jaccard, 1901). The formula is given in 5.1.

[42] Cysouw (2003) argues that implicational universals should be interpreted as bidirectional statistical correlations. Rebuttals are given in Maslova 2003a, Plank 2003 and Dryer 2003.

(5.1) $J(A, B) = \frac{|A \cap B|}{|A \cup B|}$

Table 5.9: Distance matrix of PHOIBLE segment inventories[43]

	bvr	qvh	alh	roo	ald	ale
bvr	0.0	0.6667	0.6176	0.72	0.6098	0.8
qvh	0.6667	0.0	0.8125	0.7647	0.4773	0.7917
alh	0.6176	0.8125	0.0	0.8436	0.8077	0.8723
roo	0.72	0.7647	0.8436	0.0	0.7949	0.8825
ald	0.6098	0.4773	0.8077	0.7949	0.0	0.8113
ale	0.8	0.7917	0.8723	0.8825	0.8113	0.0

This is a very coarse grained approach that gives a numerical distance between two segment inventories by calculating their shared segments. A distance matrix can be calculated at the level of segments, phonetic features or at the level of distances of segment types (this would require some notion of similarity between segment types, which for example can be derived from the shared/not-shared characters or phonetic features between segment types). For example, Table 5.10 shows a partial PHOIBLE data dump of segment inventories by language code and segment type. This type of matrix can be read as input into R and functions can be used to calculate Jaccard distance, Pearson correlation, etc., matrices that can then be used to do MDS.

Figure 5.16 is an MDS plot of the 75 most frequent vowel types in PHOIBLE using Classical Multidimensional Scaling, also known as Principal Coordinates Analysis (Gower, 1966).[44] A distance matrix using the Jaccard index was the input for the MDS. These figures were generated using the *cmdscale* function in the R software package (R Development Core

[43] Key: Burarra [bvr] (Australia), Quechua [qvh] (Peru), Alawa [alh] (Australia), Rotokas [roo] (PNG), Alladian [ald] (Côte d'Ivoire), Aleut [ale] (US).

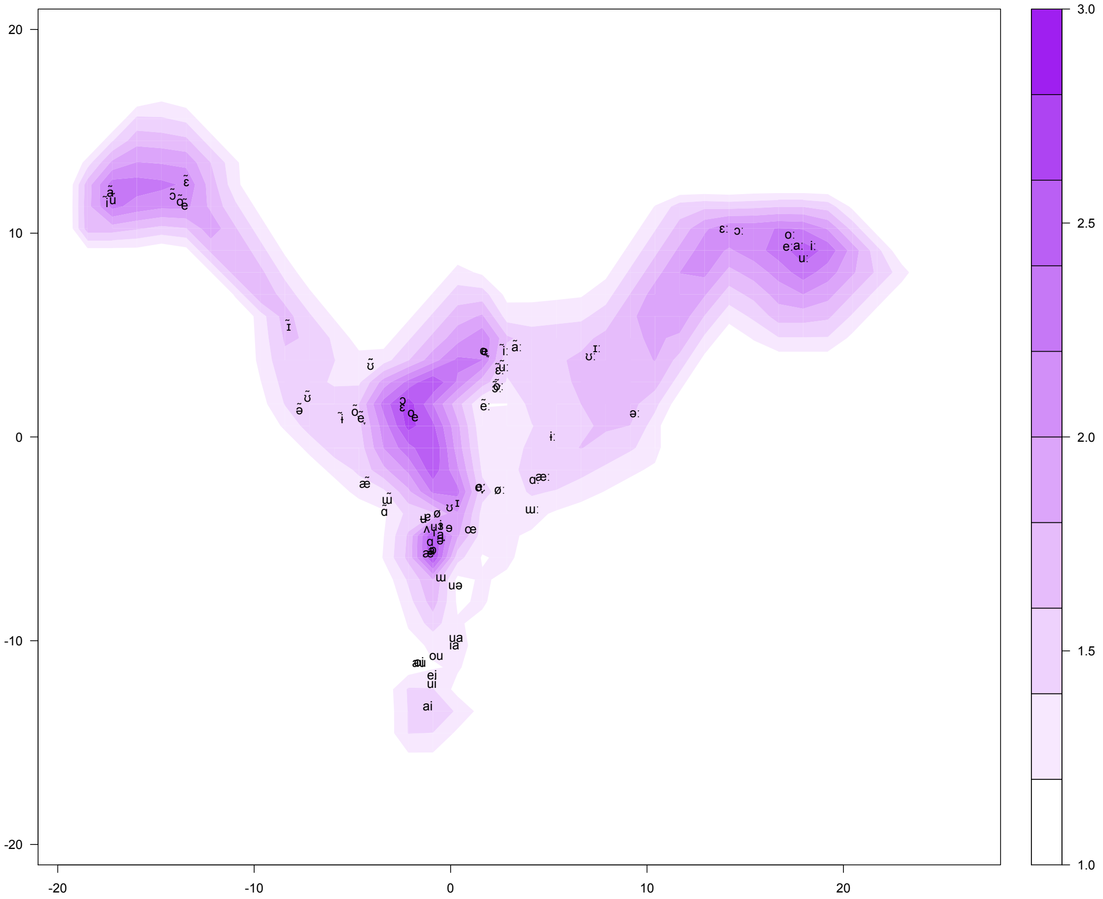
[44] Although there are 495 vowel segment types, a limit of 75 was chosen as a matter of convenience – vowels in the range of 75-495 occur exceedingly rarely (in less than 3% of languages in the sample, with 270 of them occurring in only one inventory, i.e. in less than .001% of the languages in the sample).

Table 5.10: PHOIBLE segment inventories by language code and ngram

	m	ɱ	mb	mː	ŋm
xan	0	0	0	0	0
bud	1	1	0	0	1
oca	1	0	0	1	0
kwd	1	0	1	0	0

Team, 2011). The x and y axes are the first two dimensions of MDS, i.e. unnamed dimensions of variation deemed important by the MDS. The bar on the right represents the frequency.

Figure 5.16: MDS plot of 75 most frequent vowels in PHOIBLE

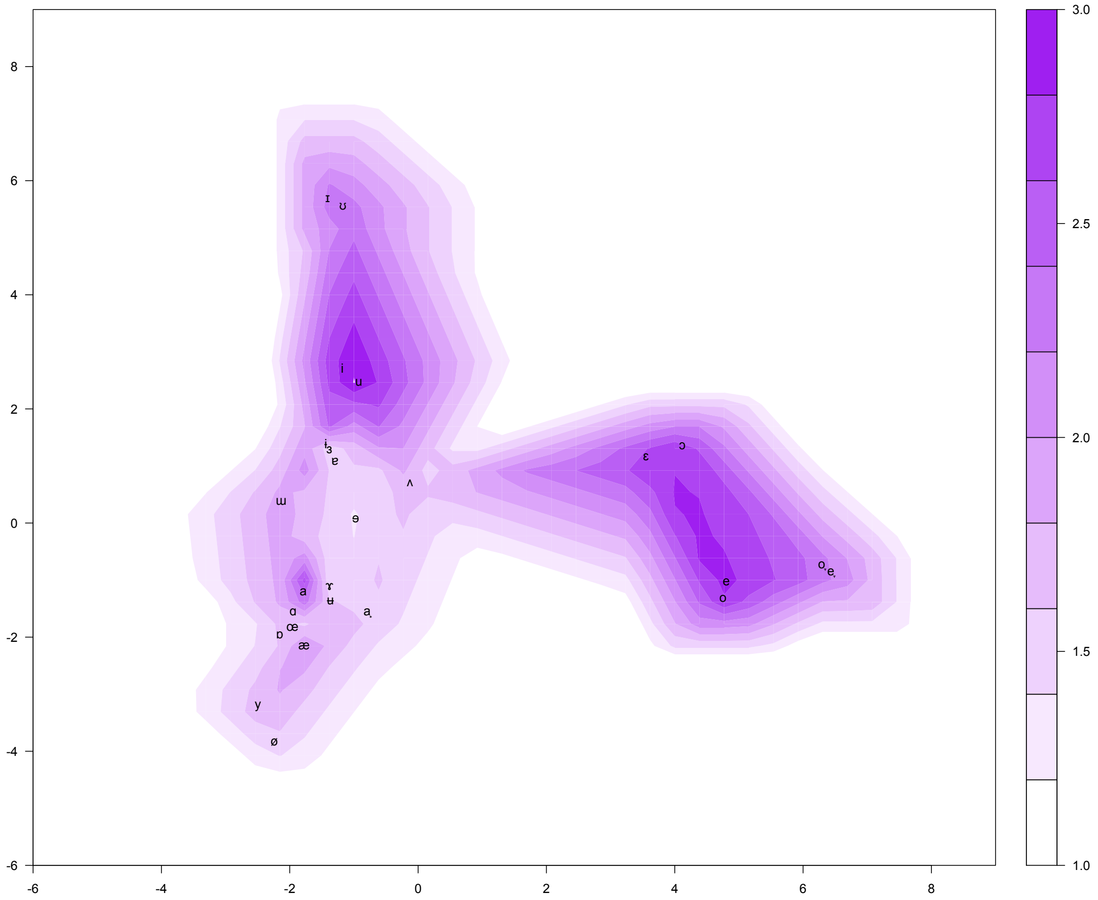


Although the stress function has flattened the cluster of cardinal vowels around the center due to the frequencies and complexity of the relations between them in segment inventories, there is a clear tendency for vowels systems in the PHOIBLE sample, once cardinal vowels are in place, to make one of three decisions for expansion. To the upper left there is a high frequency cluster of nasalized vowels, with nasalized /i, a, u/ being the most frequent in segment inventories, and /o, e, ɔ, ɛ/ and then /ɪ, ʊ, ə/ being less frequent. On the other hand, another choice is for the vowel system to use vowel length to employ contrast outside of the cardinal vowels. These can be seen in the upper right corner. Again /i, a, u/ are the most frequent of the splinter group, then /o, e/, /ɛ, ɔ/, and then /ʊ, ɪ, ə/. Between the two frequency nodes, one can see the set of nasalized and lengthened vowels, which are less frequent than either set independently. Towards the center bottom of the MDS image, a peak of diphthongs is clearly visible. Thus, according to this classical multidimensional scaling technique, once languages expand their vowel inventories beyond cardinal vowels, they tend to do so by either nasalization or lengthening, and to a lesser extent by adding diphthongs to the inventory.

Figure 5.17 focuses more specifically on the cardinal vowels space by sampling the 26 most frequent vowels (those occurring in greater than 10% of inventories) and their co-occurrences in the same data sample. The vowels /i/ and /u/ clearly cluster frequently together, with /a/ being a bit below in its own peak.[45] From those cardinal vowels, /e/ and /o/ are the next most frequently co-occurring. Then there is /ɛ/ and /ɔ/ very frequently occurring together, as well as the pair /ɪ/ and /ʊ/. The MDS image shows that these cardinal vowels typically occur in front/back pairs at the various height levels.

[45] The position of /a/ may be the influence of transcription effects. See Section 2.3.5.

Figure 5.17: MDS plot of 26 most frequent vowels in PHOIBLE



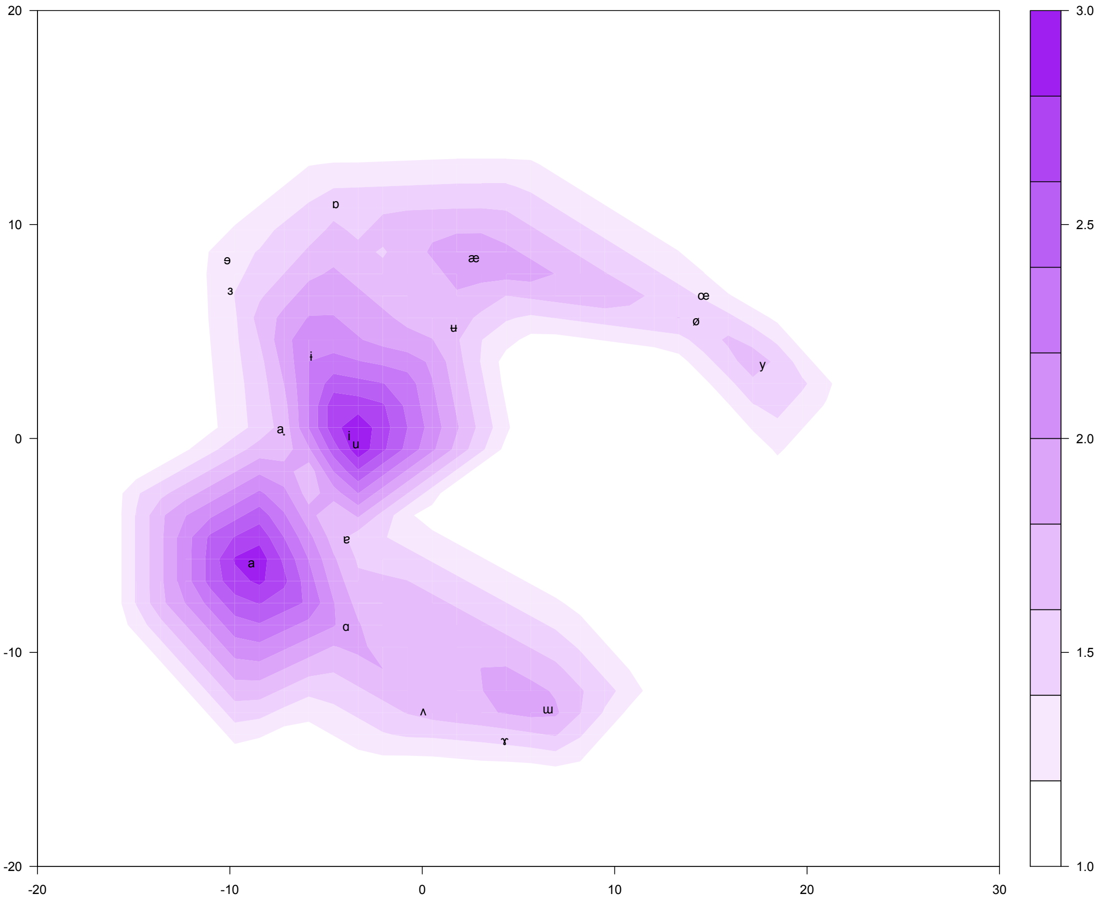
In this experiment, I have not standardized the PHOIBLE data but have gone with the keep-all-data approach.[46] Note that the distinction between /e/ and /e̞/, and /o/ and /o̞/, are from the SPA and UPSID_{451} data and they indicate a distinction between "higher-mid" and "mid" vowels. There are only three inventories in SPA and UPSID_{451} that have

[46] I also did not apply genealogical sampling beforehand because I was not trying to estimate standard error due to genealogical bias. Instead I am interested in looking at possible patterns in all of the data. I suspect resampling would not change the results very much, because certain vowels patterns tend to occur regardless of genealogical origin, e.g. front and back vowel pairs, sets of cardinal vowels that are also lengthened, etc.

“higher-mid” and “mid” contrastive pairs of vowels.[47] Nevertheless, they show the same vowel space patterning: a front/back pair along the same height.

Figure 5.18 focuses on the 18 most frequent vowels (occurring in 17% or more segment inventories) in the PHOIBLE data set. It clearly shows that /i, a, u/ are the most likely vowels to occur in a language, in line with Crothers’s (1978) claim using the SPA database in the 1970s.

Figure 5.18: MDS plot of 18 most frequent vowels in PHOIBLE



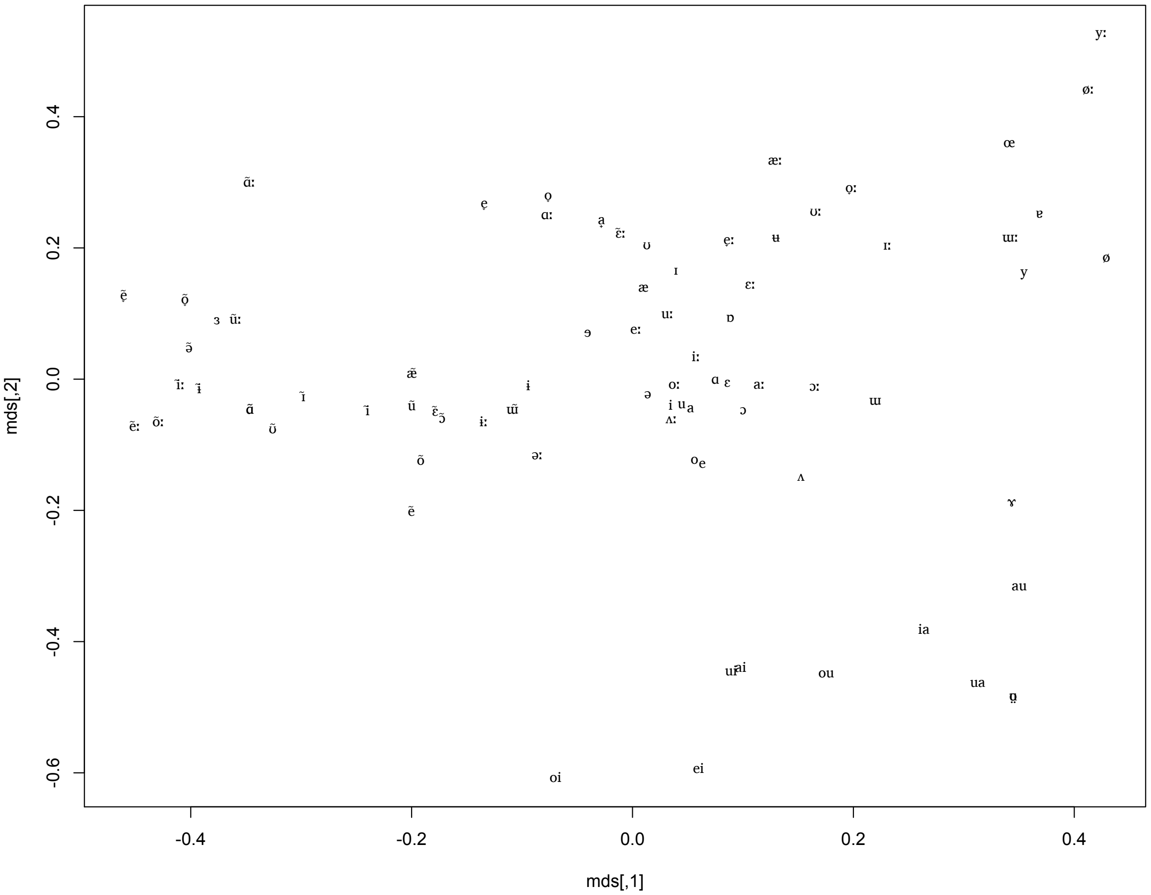
After looking at the output of MDS using the Jaccard index, I decided to also see what

[47] See discussion in Section 5.3.

kind of visualization would occur if I used a distance matrix produced by a different metric. Pointwise mutual information (PMI) is an information theoretic approach that measures the mutual dependence of two variables. PMI is a measure of association and I used it to calculate the distances between segments. I then produced an MDS plot based on the PMI distance matrix. Keeping with the same 78 most frequent vowels for sake of consistency, a PMI plot is given in Figure 5.19. As can be seen, there is a separation on the x-axis between nasalized vowels and all other vowels. On the y-axis, there is a separation between diphthongs and all other vowels. Since the PMI study is only preliminary, I have not yet investigated what appears when only the more frequent vowels are used. I leave this topic for future research.

To summarize, I have looked at relationships that hold among vowel systems in the PHOIBLE data set by creating distance matrices using the Jaccard index, and preliminarily pointwise mutual information, and then visualizing these through multidimensional scaling. MDS visualizes some of the patterns that are inherent in the vowel space of inventories in PHOIBLE, e.g. vowel systems seem to grow by non-vowel quality distinctions like nasalization, lengthening and diphthongization. They then tend to pattern in front and back pairs. The smallest vowel systems tend to start with /i, a, u/.

Figure 5.19: Pointwise mutual information



5.7 Conclusion

In this chapter I have revisited some of the descriptive typological facts about segment inventories and vowel systems put forth in works like Maddieson 1984 and Crothers 1978. I did so by comparing the SPA, UPSID$_{451}$ and PHOIBLE data sets, each of which includes substantially more languages and genealogical diversity than its predecessor. Since PHOIBLE is a convenience sample, I implemented a method to genealogically stratify its contents to estimate the standard deviation of segment type frequencies and counts due to genealogical bias. What I found is that although the PHOIBLE data set has more than twice the number of languages in UPSID$_{451}$ and greater typological coverage, in general the segment frequencies and the mean for inventories and their segment makeup remain close to those put forth in previous work by Maddieson.

The PHOIBLE data set is genealogically skewed towards certain language families, such as Niger-Congo. I implemented a statistical technique that resamples groups to calculate the controlled frequencies of the distribution of segment types in PHOIBLE. This technique shows that segment types frequently found in most languages tend to be not far off from their frequency in the PHOIBLE and UPSID$_{451}$ databases. Table 5.11 shows the number of inventories in PHOIBLE per geographic region and their mean number of segments.

Table 5.11: Geographic area and mean of segment inventories in PHOIBLE

Area	Languages	PHOIBLE count	Mean of inventories
Africa	2,110	451 (21.4%)	39.6
Americas	993	246 (26.4%)	31.7
Asia	2,322	192 (8.3%)	35.6
Europe	234	61 (26.1%)	39.8
Pacific	1,250	137 (11%)	23.7

Underrepresented segments in PHOIBLE are found in segment types like those that may be considered spurious across descriptions (e.g. higher-mid vs mid vowels in languages in

SPA and UPSID$_{451}$) and overrepresented segments are shown to occur in nasals and other common sounds in Niger-Congo languages (which are overrepresented in PHOIBLE).

In this chapter I have also shown that as the number of segment inventories in PHOIBLE increases, the number of distinct segments also continues to increase quadratically. More than 50% of these segment types occur language-specifically, i.e. they occur in one language. With the addition of more inventories, we will see if this curve flattens out before all languages are added to PHOIBLE or if the number of rare segment types will continue to increase as new descriptions of languages are added to the data set.

Finally, I used multidimensional scaling to investigate implications in vowel systems. I show that Crothers's (1978) observation that vowel system typically have /i, a, u/ holds. Furthermore, when vowel systems grow beyond the basic cardinal vowels, they seem to do so first by length and nasalization, and then diphthongization. In the next chapter, I develop the computational architecture needed to probe segment inventories at the level of distinctive features.

Chapter 6
DISTINCTIVE FEATURES

6.1 Introduction

In the previous chapter, I revisited some of the typological facts of segment inventories at the segment level. In this chapter, my aim is to examine segment inventories at the level of features. To do so, I begin with a brief discussion of segments and features in Section 6.2 and then I show in Section 6.3 that distinctive feature sets in general lack the typological representation needed to straightforwardly map each segment type in PHOIBLE to a set of features. Therefore, in Section 6.4 I investigate the different types of segments and I outline how to compositionally encode features by combining feature vectors and assigning them to segment types. The segment types and their features vectors are modeled in an RDF/OWL knowledge base, which provides the functionality for the user to query across segment inventories at the feature level.[1] The user can query by feature, by sets of features that define natural classes, or by omitting features in queries to utilize the underspecification of segment types. The RDF/OWL model also provides structure that allows for the hierarchical organization of features into a feature geometry, which can be used to query inventories, and the model provides additional functionality to use logical operators and constraints in queries. My intent is to build a computational tool to allow researchers to undertake typological comparisons of segment inventories at the level of features. The system I have built does not rely on any particular feature set and the technologies I use allow users to plug other distinctive feature sets into the PHOIBLE architecture by mapping feature vectors to segment types, defining them in RDF, and merging the graphs. I use the system in Section 6.5 to investigate descriptive universals of phonological systems, such as "all languages have coronals" and "every phonological system has at least one front vowel

[1] In Sections 3.2 & 3.3, I described how I model segments and features in an RDF/OWL knowledge base. More examples of how the knowledge base can be queried are given in Section 6.5 in this chapter.

or the palatal glide /j/" (Hyman, 2008).

6.2 Background

In Section 2.2 I gave a brief overview of the linguistic theories that underly segmental phonology and distinctive feature theory. To summarize, features can be thought of as atoms that combine compositionally to form a segment. A glyph is used to graphically encode a segment, i.e. a language-specific phoneme or allophone. A segment may also be used to encode an abstract class of phonemes that may pattern in similar ways across languages. I call the former, language-particular segments, *segment tokens*. The latter, abstract sense, are *segment types*. For example, by consensus of the descriptive linguistics literature, the segment <u> is typically used to encode the articulatory features of an acoustic signal that is a high back (IPA "close") rounded vowel. In a particular feature set, say Hayes 2009, the segment <u> (either allophonically [u] or phonemically /u/) is shorthand for an unordered set, or vector, of binary features: {+high, +back, +round, ...}. The segment type /u/, i.e. the contrastive phoneme characterized by those features, is found in many different languages. In fact, there are 939 segment tokens of /u/ that occur in 1089 segment inventories in PHOIBLE. Thus, the segment type frequency of /u/ is 86% in the PHOIBLE data set.[2]

Statements regarding the distribution of segment types, however, conceal multiple layers of abstraction.[3] What does it mean to state that 86% of languages in some data set have a contrastive /u/? If features figuratively resemble atoms, then in the acoustic speech signal, formants analogize to quarks. Spectrogram analysis shows that every utterance is unique. If every utterance of [u] is unique, then every [u] in every language must be unique. Therefore every /u/ is unique, unless some level of abstraction is introduced for cross-linguistic language-level analysis.[4]

I distinguish between three levels of abstraction for speech sounds and their symbolic representations, summarized in Table 6.1. At the utterance-level, allophones are an abstrac-

[2] When weighted for genealogical, as discussed in Section 5.3, /u/ occurs with a frequency of 78%.

[3] Ideas in this section benefitted from discussions with Dan McCloy.

[4] See discussion in Section 2.3.1.

tion that glosses over minor variation in the acoustic speech signal (even though realizations of a given allophone may vary greatly). An additional abstraction at the language-level is introduced when systematic allophonic variants are collapsed into an abstract phoneme, symbolized by language-level segment tokens. However, if we want to compare phonemes cross-linguistically, how can we be certain that the segment tokens are all representing something similar enough to justify making cross-linguistic claims? In the case of /u/, a three-vowel system that contains a non-low back vowel will likely permit much more variation in the acoustic space to the sounds represented by /u/, than a vowel system that contrasts /ɯ, u, ɤ, o/. As discussed in Section 2.3, there is an inherent problem in comparing languages at the phonemic level, since varying levels of abstraction will be present from analysis to analysis. Additionally, the level of detail varies greatly from language description-to-language description. If representations are not comparable, they cannot be counted as two instances of the same thing. Thus, some type of comparative concept is needed to undertake typology (Haspelmath, 2010).

Table 6.1: Speech sounds and symbolic representations at different levels

	Speech sounds	Symbolic representations
Utterance level [ ]	(allo)phone	segment token
Language level //	phoneme	segment token
Cross-linguistic level	comparative concept	segment type

Is it legitimate to generalize from language-specific tokens to cross-linguistic types when it comes to phonemes? There are arguments for and against.[5] A major problem is that different linguists typically reach different conclusions on what a set of phonemes is for a particular language.[6] For example, if one linguist's phonemic analysis of a language leads

[5] See Section 2.3.1.

[6] I have undertaken preliminary analysis on how often two descriptions of the same language's phoneme inventory in PHOIBLE are described differently by two authors. Using a very strict segment to segment comparison on a set of 217 pairs of inventories, only two matched precisely. The mean Jaccard index

him or her to posit an /u/ phoneme from the allophones [ɯ, u, ɤ, o], but another linguist posits /ɯ/, how can we typologize vowel systems by high back vowels? Searching for all languages with /u/ will not return results like Ocaina [oca], which has the vowel system /i, ĩ, ɛ, a, ã, o, õ, ɯ, ɯ̃/ (Agnew and Pike, 1957; Maddieson and Precoda, 1990). Alternatively if one searches for languages with /ɯ/, the /u/ results are missed. Of course one can search for languages that have either /u/ or /ɯ/, but this is just a simple example in a rather complex system. Of the 216 languages that have multiple segment inventories in PHOIBLE, nearly all of them differ in some aspect of their phonemic inventories. For example, Tuva [tuv] as described in UPSID$_{451}$ has 29 phonemes with a nine vowel system /y, i, e, ø, ɛ, a, ɤ, o, u/ that contains a rounding distinction in front and back mid vowels and high front vowels (Maddieson and Precoda, 1990).[7] On the other hand, Harrison (2000a) posits a segment inventory consisting of 37 phonemes and 16 distinctive vowels: /y, yː, i, iː, e, eː, ø, øː, a, aː, o, oː, ɯ, ɯː, u, uː/. Harrison's analysis treats length as contrastive. Like the example of rounding, other vowel features like tense, length and nasalization vary widely from description to description. The answer to this search problem lies in mapping features to segments and then underspecifying features in a query to match classes of segments. For example, to capture all high back vowels regardless of rounding, underspecify the feature [round]. If tense should not be taken into account, underspecify the feature [tense].[8]

The many flavors of the phoneme /t/ is another example of why some form of feature underspecification is desirable. For example, querying the PHOIBLE knowledge base for the segment token /t/ returns 800 inventories. However, that query does not return the 172 inventories that have a voiceless dental plosive /t̪/, or the 12 inventories that contain

across inventories is only roughly 57%. This analysis does not yet take into account phonetic distance of segments (e.g. one author posits phonemic /u/ and the other /ʊ/) or differences inherent in the data sources (e.g. UPSID$_{451}$ does not contain tone; SPA does).

[7] In the UPSID$_{451}$ data, Maddieson and Precoda (1990) note that "Accounts of the Tuva vowel system differ widely. The system given here is that of Song (1982) since this makes the basis of vowel harmony clear: front and back vowels belong to different sets. All vowels occur long and maybe nasalized. Song mentions that older speakers distinguish a series of vowels with tense phonation. Place of articulation is based on Seglenmej (1979)." Note that vowel length is a contrastive feature used in UPSID$_{451}$ inventories, although it isn't marked in the Tuva segment inventory.

[8] There are 15 languages in PHOIBLE that have the vowels /i, a, ʊ/, but not /u/ (however they include /uː/). Eleven languages have /i, a, ʊ/, but not /u/ or /uː/.

a voiceless palato-alveolar plosive /t̠/, or the 92 inventories from UPSID$_{451}$ that leave the dental/alveolar /t̪|t/ place of articulation underspecified.[9] I try to avoid the problem of different linguists representing different sounds with the same segment by instantiating the ability to query at the level of features. Using features and feature geometry allows us to underspecify our queries within a given feature theory, so that statements like N% of languages have at least one coronal stop can be answered by the knowledge base (see Section 6.5). For example, querying on the Hayes 2009 features [+coronal] and [−delayed release] will return all coronal stops, including /t, t̪, t̠, t̪|t/.[10] To attain this functionality, however, we must have complete typological coverage of features for all segment types in PHOIBLE.[11]

6.3 Typological coverage

In recent years, a different hypothesis, that language learners acquire and classify features and constraints instead of picking them out from a predefined Universal Grammar (UG), has led to emergent theories of phonology and distinctive features (Blevins, 2004; Mielke, 2004; Mohanan et al., 2009). In an emergent approach, features and constraints emerge from the learner's experience and not from mapping the target language to a set of inherent features. Thus, one implication is that features are language-specific. If features are language-specific, then there is no limit on the set of features used across languages. If there is no UG of features, should we abandon all feature-based cross-linguistic comparison (Mohanan et al., 2009)? The answer is no. Mohanan et al. (2009, 151) suggest that what is needed for typological comparison in phonology is "a theory of feature emergence that expresses the family resemblances of features, connecting the concrete aspects of the articulation and perception of speech to a cross-linguistically shared set of features".

If features are indeed emergent, and therefore language-specific, one would expect segment inventories to contain random segments. However, one striking observation is that

[9] We can be reasonably sure that nobody will use <t> to represent a labial or velar sound, etc.

[10] The segment /t̪|t/ stands for an underspecified dental /t/ or alveolar /t/. This construction appears in UPSID$_{451}$. See Sections 2.3.4 and 4.3.2.

[11] There are over 1700 segment types in PHOIBLE. See Section 5.3.

segment inventories, particularly consonant systems, tend to exhibit symmetry in their structure (Clements, 2003b). To constrain the range of phonetic possibilities, features are grounded in concrete physical terms and are involved in structuring inventories of contrastive sounds. There seems to be general principles controlling phonological systems, like feature economy, which apply at the level of distinctive features and not segments. According to feature economy, languages tend to maximize the ratio of sounds over features (Clements, 2003a,b, 2009). For example, by introducing a non-quality feature into the vowel system, such as length or nasalization, a language can increase its number of contrastive vowels with a single feature. On the other hand, to introduce the same number of contrastive vowels by using vowel quality distinctions may introduce (more) asymmetry into the vowel space.[12]

The question of whether features are innate, or if they emerge through language learning and use, is an important question to investigate, but it is not within the scope of this work. Regardless of whether there is a set of universal features or if features are emergent, to undertake typological comparisons of segment inventories at the feature level and to investigate feature-based principles structuring phonological inventories, a cross-linguistic set of features is needed that has full typological coverage over the data set. Mohanan et al. (2009, 151) state, "What is needed is a cross-linguistically valid currency of distinctive features: such a currency can obtain without reference to a set of features stipulated in UG." The assignment of features to segment types in this work is considered a computational challenge because there is no feature set that has complete coverage of segment types that appear for all inventories in PHOIBLE. Therefore, one must be created. As a first step, I chose to investigate the segment type typological representation of two feature sets aimed at wide typological coverage: Hayes 2009 and Maddieson and Precoda 1990.[13]

Most of the language documentation that I encountered during PHOIBLE's development

[12] See Section 5.6.

[13] Hayes's feature set is available in electronic form (Microsoft Excel) online at: `http://www.linguistics.ucla.edu/people/hayes/120a/index.htm`. Note that this resource is not in Unicode. Bill McNeill, Dan McCloy and I converted the segments to Unicode IPA. The Maddieson & Precoda features are also available electronically online: `http://www.linguistics.ucla.edu/faciliti/sales/software.htm`. Note that the UPSID_{451} segments are in an ASCII encoding.

used the IPA for transcription or some dialect of the IPA, e.g. APA, IPA with Africanist conventions, IPA with idiosyncratic changes, etc. Therefore, I decided to target the Hayes 2009 feature set because it likely has the most complete IPA coverage of any feature set. Hayes's feature set includes 141 segments and 18 diacritics, which can be combined together compositionally to assign features to segment types that are not explicitly defined.[14] In Hayes, features are either binary or not applicable ("+" means has feature; "−" does not have feature; "0" signals not applicable). Segments are defined by 28 features. Hayes defines features for diacritics, for four complex segments (pt, bd, kp, gb) and for 30 contour segments (e.g. pf, ts, dʒ, ...).[15] Table 6.2 shows a few segments and their feature vectors.

Table 6.2: Hayes 2009 feature set (selected segments)

	i	u	a	p	v	h	ɾ	n	pf	ʤ
syllabic	+	+	+	−	−	−	−	−	−	−
stress	−	−	−	−	−	−	−	−	−	−
long	−	−	−	−	−	−	−	−	−	−
consonantal	−	−	−	+	+	−	+	+	+	+
sonorant	+	+	+	−	−	−	+	+	−	−
continuant	+	+	+	−	+	+	+	−	−	−
delayed release	0	0	0	−	+	+	0	0	+	+
approximant	+	+	+	−	−	−	+	−	−	−
tap	−	−	−	−	−	−	+	−	−	−
trill	−	−	−	−	−	−	−	−	−	−
nasal	−	−	−	−	−	−	−	+	−	−
voice	+	+	+	−	+	−	+	+	−	+
spread glottis	−	−	−	−	−	+	−	−	−	−
constricted glottis	−	−	−	−	−	−	−	−	−	−
labial	−	+	−	+	+	−	−	−	+	−
round	−	+	−	−	−	−	−	−	−	−
labiodental	−	−	−	−	+	−	−	−	+	−

[14] Hayes (2009, 94) notes that "Many sounds absent from the charts can have their features deduced by looking up a similar sound and changing the most obvious features".

[15] I leave out the combining tie bar in this text.

Table 6.2: Hayes 2009 feature set (selected segments)

	i	u	a	p	v	h	ɾ	n	pf	ʤ
coronal	−	−	−	−	−	−	+	+	−	+
anterior	0	0	0	0	0	0	+	+	0	−
distributed	0	0	0	0	0	0	−	−	0	+
strident	0	0	0	0	0	0	−	−	0	+
lateral	−	−	−	−	−	−	−	−	−	−
dorsal	+	+	+	−	−	−	−	−	−	−
high	+	+	−	0	0	0	0	0	0	0
low	−	−	+	0	0	0	0	0	0	0
front	+	−	−	0	0	0	0	0	0	0
back	−	+	−	0	0	0	0	0	0	0
tense	+	+	0	0	0	0	0	0	0	0

Whereas the Hayes feature set is compositional, i.e. a set of features is assigned to each IPA segment and those segments can be used as building blocks for other segments, the UPSID feature set is non-compositional. Each segment type in UPSID was specifically assigned a set of feature values from a set of pre-defined features and these segment types cannot be combined to specify feature vectors for additional segment types. I chose to evaluate the typological coverage of the UPSID_{451} feature set because of its broad coverage of languages' segment inventories and because Maddieson (1984) and Maddieson and Precoda (1990) faced the same challenges of assigning a vector of features to each segment type in their database. These mappings allowed Maddieson to report on the distribution of segment types. I was also interested in the segment type coverage of UPSID_{451}'s features on the expanded inventories in PHOIBLE. This shows to what degree the range of segment types vary from a cross-linguistic segment inventory database of 451 inventories to one with over 1000 inventories.

Maddieson and Precoda (1990) use a set of 64 binary features to define each of the 921 segment types in UPSID_{451}.[16] Some example segments are given in Table 6.3.[17]

Table 6.3: UPSID_{451} feature set (selected segments)

Feature	uo	a	b	t̪|t	dⁿ	ʃ	qʷʔ	ɡ!	kx	ɱ	mb
plosive	0	0	1	1	1	0	1	0	0	0	1
implosive	0	0	0	0	0	0	0	0	0	0	0
ejective stop	0	0	0	0	0	0	0	0	0	0	0
click	0	0	0	0	0	0	0	1	0	0	0
fricative	0	0	0	0	0	1	0	0	0	0	0
ejective fricative	0	0	0	0	0	0	0	0	0	0	0
affricate	0	0	0	0	0	0	0	0	1	0	0
ejective affricate	0	0	0	0	0	0	0	0	0	0	0
affricated click	0	0	0	0	0	0	0	0	0	0	0
unspecified r-sound	0	0	0	0	0	0	0	0	0	0	0
tap	0	0	0	0	0	0	0	0	0	0	0
flap	0	0	0	0	0	0	0	0	0	0	0
trill	0	0	0	0	0	0	0	0	0	0	0
approximant	0	0	0	0	0	0	0	0	0	0	0
nasal	0	0	0	0	0	0	0	0	0	1	0
simple vowel	0	1	0	0	0	0	0	0	0	0	0
diphthong	1	0	0	0	0	0	0	0	0	0	0
lateral	0	0	0	0	0	0	0	0	0	0	0
sibilant	0	0	0	0	0	1	0	0	0	0	0
bilabial	0	0	1	0	0	0	0	0	0	0	1
labiodental	0	0	0	0	0	0	0	0	0	1	0
linguolabial	0	0	0	0	0	0	0	0	0	0	0
dental	0	0	0	0	0	0	0	0	0	0	0
unspecified dental or alveolar	0	0	0	1	0	0	0	0	0	0	0
alveolar	0	0	0	0	1	0	0	1	0	0	0
palatal-alveolar	0	0	0	0	0	1	0	0	0	0	0

[16] There is no non-applicable "0" feature in UPSID_{451}. Values are either true or false.

[17] In the original UPSID_{451} feature table, features values are denoted with "TRUE" or "FALSE". They are represented here with "1" and "0", respectively.

Table 6.3: UPSID$_{451}$ feature set (selected segments)

Feature	uo	a	b	t̪|t	dⁿ	ʃ	qʷʔ	ɡ!	kx	ɱ	mb
retroflex	0	0	0	0	0	0	0	0	0	0	0
palatal	0	0	0	0	0	0	0	0	0	0	0
velar	0	0	0	0	0	0	0	0	1	0	0
uvular	0	0	0	0	0	0	1	0	0	0	0
pharyngeal	0	0	0	0	0	0	0	0	0	0	0
glottal	0	0	0	0	0	0	0	0	0	0	0
labialized	0	0	0	0	0	0	1	0	0	0	0
palatalized	0	0	0	0	0	0	0	0	0	0	0
velarized	0	0	0	0	0	0	0	0	0	0	0
pharyngealized	0	0	0	0	0	0	0	0	0	0	0
nasalized	0	0	0	0	0	0	0	0	0	0	1
nasal release	0	0	0	0	1	0	0	0	0	0	0
prestopped	0	0	0	0	0	0	0	0	0	0	0
lateral release	0	0	0	0	0	0	0	0	0	0	0
high	0	0	0	0	0	0	0	0	0	0	0
higher mid	0	0	0	0	0	0	0	0	0	0	0
mid	0	0	0	0	0	0	0	0	0	0	0
lower mid	0	0	0	0	0	0	0	0	0	0	0
low	0	0	0	0	0	0	0	0	0	0	0
front	0	0	0	0	0	0	0	0	0	0	0
central	0	0	0	0	0	0	0	0	0	0	0
back	1	0	0	0	0	0	0	0	0	0	0
nonperipheral	0	0	0	0	0	0	0	0	0	0	0
rounded	1	0	0	0	0	0	0	0	0	0	0
unrounded	0	1	0	0	0	0	0	0	0	0	0
lip-compressed	0	0	0	0	0	0	0	0	0	0	0
r-colored	0	0	0	0	0	0	0	0	0	0	0
backing	0	0	0	0	0	0	0	0	0	0	0
lowering	1	0	0	0	0	0	0	0	0	0	0
rounding	0	0	0	0	0	0	0	0	0	0	0

Table 6.3: UPSID$_{451}$ feature set (selected segments)

Feature	uo	a	b	t̪|t	d^{n}	ʃ	q^{w}ʔ	ɡ!	kx	ɱ	mb
voiceless	0	0	0	1	0	1	1	0	1	0	0
voiced	1	1	1	0	1	0	0	1	0	1	1
aspirated	0	0	0	0	0	0	0	0	0	0	0
laryngealized	0	0	0	0	0	0	1	0	0	0	0
long	0	0	0	0	0	0	0	0	0	0	0
breathy	0	0	0	0	0	0	0	0	0	0	0
overshort	0	0	0	0	0	0	0	0	0	0	0
preaspirated	0	0	0	0	0	0	0	0	0	0	0

I used two rough measurements to evaluate the typological representation of the Hayes 2009 and Maddieson and Precoda 1990 feature sets on the segment types in the PHOIBLE inventories. The first is a full string match for segment types (there are 1780 segment types in the combined PHOIBLE inventories). I use this method to evaluate the typological coverage of both feature sets on the range of PHOIBLE segments. The second measurement splits PHOIBLE segment types into their component Unicode characters, and then checks for a feature vector for each character. This evaluation method is not applicable for the UPSID$_{451}$ feature set; I use it to evaluate the compositional nature of features in Hayes 2009. In Section 6.4, I discuss an algorithm that extends this second measurement by then compositionally combining feature vectors and assigning them to a segment type.

When using a simple segment type match, the coverage for Hayes 2009 is poor, covering roughly 7% of the segment types in PHOIBLE inventories. This is to be expected, since the Hayes feature set is like the IPA in the sense that each segment is a potential building block for segment types, so it will only cover the non-compositional IPA segments in inventories in PHOIBLE. UPSID$_{451}$ defines the feature vectors for 951 segment types. Its segment type coverage was considerably higher at nearly 46%.

At the compositional level, the typological coverage of Hayes 2009 is much higher than its segment type coverage, which should also be expected. Hayes defines feature vectors for

159 segments and diacritics that can be combined to create feature vectors. For example, the feature vector for aspiration <^{h}> { +spread glottis, −constricted glottis } overwrites the applicable features in the base segment it combines with. The segment <p> is (among other features) { −spread glottis, −constricted glottis }. Combining the features of <p> and <^{h}> would result in { +spread glottis, −constricted glottis } for <p^{h}>. However, even when I decomposed all PHOIBLE segment types into their component Unicode characters and took a unique list of those characters, Hayes 2009 only accounts for 71% of the characters in PHOIBLE. This is due to the lack of feature definitions for tones, clicks, implosives, some IPA-sanctioned segments (open-mid central unrounded vowel [ɜ], epiglottal plosive [ʡ], voiceless epiglottal fricative [ʜ], voiced epiglottal fricative [ʢ]),[18] a non-sanctioned IPA segment (voiced retroflex implosive [ᶑ]), and some IPA-sanctioned diacritics (half-length [ˑ], lateral release [ˡ], nasal release [ⁿ], extra short [ŏ], centralized [ö], advanced tongue root [o̘], retracted tongue root [o̙], raised [o̝], lowered [o̞], non-syllabic [o̯], more rounded [o̹], less rounded [o̜], apical [t̺], laminal [d̻]).[19] Finally, during the construction of PHOIBLE I added some segments that appear in SPA, UPSID$_{451}$ or in grammars from which I extracted inventories and there exists no IPA-sanctioned symbols: [ᴅ] (used to represent a tap as distinguished from flap in UPSID$_{451}$), [ʱ] (breathy marker for stops), [ᶣ] (a palatalized diacritic [ʲ] plus rounding), [ᶾ] (slightly palatalized while also being slightly labialized; see Heath 2005a), [x͈], (tense diacritic used for SPA), [x̦] (lax marker used for SPA), and [x͓] (fricated marker used for UPSID$_{451}$). All segment types are defined in Hayes′, an extended version of the Hayes 2009 feature set that I discuss in Section 6.4.

To summarize, there are benefits to both the Hayes 2009 and Maddieson 1984 & Maddieson and Precoda 1990 approaches to assigning feature vectors to segment types. Using the UPSID approach insures that there is a feature vector for each segment type in the data set, but it isn't computationally scalable to new segment types because each new segment type must be manually assigned a vector of features. On the other hand, Hayes's approach outlines a methodology for compositionally generating a feature vector for new segment

[18] The voiceless epiglottal fricative [ʜ] and voiced epiglottal fricative [ʢ] segments do not occur in any inventory currently in PHOIBLE.

[19] For illustration purposes, diacritics are given with a base segment [o], [t] or [x].

types. If a new segment type is encountered, there is an explicitly defined formulation of how existing segments and diacritics combine to create a feature vector. However, the combination of feature vectors for complex and contour segment types, which is not discussed in Hayes 2009, has to be addressed to reach full typological coverage of all segment types in the PHOIBLE inventories.[20]

6.4 Challenges and implementation

The IPA is not designed as a catalog of possible phonemes, but as a catalog of building blocks for describing the sounds and contrastive sounds in the world's languages through the combination of articulatory features.[21] The combination of segments into segment types comes in three different kinds: simple, complex and contour. I refer to each kind as a *segment class*. These different segment classes pose challenges in assigning a vector of features from a given feature set to a particular segment type. Addressing these challenges is important because in order to query across every segment in all segment inventories at the feature level, each segment type must have a vector of features assigned to it. In the previous section, I showed the need for an explicit definition of all segment types by evaluating the segment type coverage of two typologically diverse feature sets against the segment type diversity found in segment inventories described in PHOIBLE. If we used just those feature sets, our feature level queries would miss many matches.

Traditionally, there is a distinction between three segment types (Sagey, 1986; Clements and Hume, 1995):

1. **simple** segments consist of a single segment (plus optional diacritics) and are characterized by one oral articulator feature; they can be described with a vector of distinctive features, e.g. [p] is { +labial, −voice, −delayed release, −velar }

2. **complex** segments consist of two or more roughly simultaneous oral tract constric-

[20] If a new segment is added to the IPA, it would also have to be assigned features and added to feature sets like Hayes 2009.

[21] The IPA consists of 114 speech sounds (86 consonants, 28 vowels) and 31 modifying diacritics (International Phonetic Association, 2005).

tions; they can also be described with a vector of distinctive features, e.g. the dually articulated labial-velar stop /kp/ is { +velar, −voice, −delayed release, +labial } (Ladefoged, 1964) or the labial-aveolar stop /tp/ (Ladefoged and Maddieson, 1996, 344)

3. **contour** segments represent a temporal movement in phonetic features from a preceding segment to the following segment; they cannot be captured in a single tier of distinctive features, e.g. a prenasalized stop like [nt] is composed of the conflicting features in [n] { +coronal, +voice, +nasal } and [t] { +coronal, −voice, −nasal }

All three segment classes behave as individual phonemic elements in segment inventories in the PHOIBLE database. Each segment type requires features to be assigned to it so that all segments in the PHOIBLE knowledge base can be queried via feature categories.

Simple segments are fairly straightforward to assign features to algorithmically. Any simple segment is assigned the set of features as defined for it in a given input feature matrix, such as Hayes 2009. Table 6.4 shows a partial feature matrix of several simple segments and a diacritic. Each simple segment [p, b, t, d] is assigned a vector of binary features from a row in the matrix. Following Hayes 2009, a simple segment plus a diacritic would first be assigned a vector of features for the base segment and then the diacritic feature(s) overwrite any of the base segment's features where applicable.[22] For a segment plus diacritic, there are the logical possibilities given in Table 6.5.

[22] Implementing this algorithm is a bit more complex because certain diacritics can also precede the base segment, such as preglottalized stops, e.g. [ʔp], so this has to be accounted for when merging vectors.

Table 6.4: Partial feature matrix

segment	continuant	delayed release	voice	spread glottis	constricted glottis	labial	coronal
p	−	−	−	−	−	+	−
b	−	−	+	−	−	+	−
t	−	−	−	−	−	−	+
d	−	−	+	−	−	−	+
h				+	−		
p^h	−	−	−	+	−	+	−

Table 6.5: Logical consequences of merging binary features

segment	+	+	−	−	0	+	0	−	0
diacritic	+	−	+	−	+	0	−	0	0
combined segment	+	−	+	−	+	+	−	−	0

Complex segments can be straightforwardly accounted for as well, if they are defined in the input feature matrix. For example, Hayes's feature set includes feature vector definitions for complex segment types like the dually articulated segments [kp, gb, pf, pt, bd] as well as a number of common affricates. Complex segments are assigned a feature vector, and if they occur with a diacritic, the same principle of overwriting features applies to the base segment.[23] However, can we algorithmically assign feature vectors to complex segments that are not pre-defined in a compositional feature set?

I pointed out in Section 6.3 that the typological coverage of feature sets does not cover all segment types that appear in the inventories in the PHOIBLE data set. My aim is to automatically generate feature vectors for segment types that are encountered in language descriptions, but that are not pre-defined in a given feature set. However, the problem is that assigning a feature vector to a complex segment type can be ambiguous given the features of its component simple segments. Table 6.6 illustrates some simple segment feature vectors and their corresponding complex segment feature vectors as pre-defined in Hayes 2009.

Although the labial velars [kp] and [gb] are separately defined, applying the logical consequences of merging the binary features from [k] & [p] and [g] & [b] would actually result in the correct feature vector assignments for these complex segments. However, this is not the case with the labiodental [pf] as defined in Hayes 2009, 95. Notice the ambiguity in feature assignment in the continuant and delayed released columns. The simple segment

[23] The algorithm has to also take into account factors like performing a complex segment type match instead of compositionally assigning features to the segment type from its component segments, e.g. <gb:> as <gb> + <:> instead of <g> + <b> + <:>. This is accomplished by matching longer segments first.

[p] is {−continuant, −delayed release} and [f] is {+ continuant, + delayed release}. However, the complex segment [pf] is {−continuant, +delayed release}; the resulting complex segment's feature vector cannot be derived from the simple segments' features by simply overwriting − with +. Therefore, the feature vector assignment must be undertaken by someone with expert knowledge of phonetics because the logical combinations given above are not always dependent on the particular combination of segments. If feature assignment cannot be derived logically from its constituent segments, then a feature vector for each complex segment type has to be manually assigned, just as in the Maddieson and Precoda 1990 feature set. Thus, I manually created feature vectors for about 3% of segment types.

Next, look at the strident features for [p], [t] and [pt]; respectively 0, − and +. These also seem to follow the case of [pf]. However, if we assume that the assignment of [+strident] to [pt, bd] is actually a typo, then this system of automatically assigning features can be used. According to Hayes's feature chart, strident is only defined for [+ coronal] sounds. It only gets a + value for sibilant fricatives and affricates. All coronal stops are [−strident], so all dually-articulated stops should be [−strident] when one of the constrictions is coronal. Therefore, I decided to change the Hayes's feature set to reflect this by implementing an extended feature set called Hayes′ ("Hayes Prime"), discussed below. My implementation keeps [t, d] as [−strident], which overwrites the "0" feature of the [p, b] segments.

Contour segments pose a different problem because they are temporal in nature. Whereas simple and complex segments' feature vectors are static, contour segments encode a changing signal. Merging two feature vectors to reflect temporal movement is not a method that is explicitly defined by Hayes (2009), so I have implemented two computational approaches.

The first approach defines ordered sets within the set of features per segment type. An advantage of this approach is that all feature values from each segment are mapped to the segment type. Table 6.7 illustrates an example with the diphthongs and triphthongs found in segment inventories in PHOIBLE.

Table 6.6: Simple and complex segment feature resolution

segment	labial	coronal	continuant	delayed release	anterior	distributed	strident	dorsal	high	low	front	back
k	−	−	−	−	0	0	0	+	+	−	−	−
p	+	−	−	−	0	0	0	−	0	0	0	0
kp	+	−	−	−	0	0	0	+	+	−	−	−
g	−	−	−	−	0	0	0	+	+	−	−	−
b	+	−	−	−	0	0	0	−	0	0	0	0
gb	+	−	−	−	0	0	0	+	+	−	−	−
p	+	−	−	−	0	0	0	−	0	0	0	0
f	+	−	+	+	0	0	0	−	0	0	0	0
pf	+	−	−	+	0	0	0	−	0	0	0	0
p	+	−	−	−	0	0	0	−	0	0	0	0
t	−	+	−	−	+	−	−	−	0	0	0	0
pt	+	+	−	−	+	−	+	−	0	0	0	0
b	+	−	−	−	0	0	0	−	0	0	0	0
d	−	+	−	−	+	−	−	−	0	0	0	0
bd	+	+	−	−	+	−	+	−	0	0	0	0

Table 6.7: Contour segment feature vectors

segment	labial	round	high	low	front	back
u	+	+	+	−	−	+
i	−	−	+	−	+	−
a	−	−	−	+	−	−
o	+	+	−	−	−	+
e	−	−	−	−	+	−
ui	{+, −}	{+, −}	{+, +}	{−, −}	{−, +}	{+, −}
iu	{−, +}	{−, +}	{+, +}	{−, −}	{+, −}	{−, +}
iau	{−, −, +}	{−, −, +}	{+, −, +}	{−, +, −}	{+, −, −}	{−, −, +}
uai	{+, −, −}	{+, −, −}	{+, −, +}	{−, +, −}	{−, −, +}	{+, −, −}
iou	{−, +, +}	{−, +, +}	{+, −, +}	{−, −, −}	{+, −, −}	{−, +, +}
uei	{+, −, −}	{+, −, −}	{+, −, +}	{−, −, −}	{−, +, +}	{+, −, −}

A second approach is to fill in the feature cells with decimals by dividing the number of "+" features over the total number of features, shown in Table 6.8. On the one hand this is useful because it gives us a method to calculate a rough similarity matrix of contour segment types. This type of output can then be read into the statistical software package R as input for creating distance matrices.[24] On the other hand, this method does not capture the ordering or unique temporal properties of contour segments.

Table 6.8: Contour segment feature vectors with fraction values

segment	labial	round	high	low	front	back
u	1	1	1	0	0	1
i	0	0	1	0	1	0
a	0	0	0	1	0	0
o	1	1	0	0	0	1
e	0	0	0	0	1	0
ui	.5	.5	1	0	.5	.5
iu	.5	.5	1	0	.5	.5
iau	.33	.33	.5	.33	.33	.33
uai	.33	.33	.66	.33	.33	.33
iou	.66	.66	.66	0	.33	.66
uei	.33	.33	.66	0	.66	.33

My process of assigning feature vectors to segment types is illustrated in Figure 6.1. The process begins by preprocessing the PHOIBLE phoneme level data into a unique list of segment types. This list is input into a feature vectors generation script that also takes as input: simple, complex/contour and diacritic feature vector specifications. Complex and contour segments' feature specifications have been split away from simple segments because they must be consulted first when evaluating segment types in the PHOIBLE data set, i.e. if features are to be assigned to [pf], then it should receive the pre-defined features for

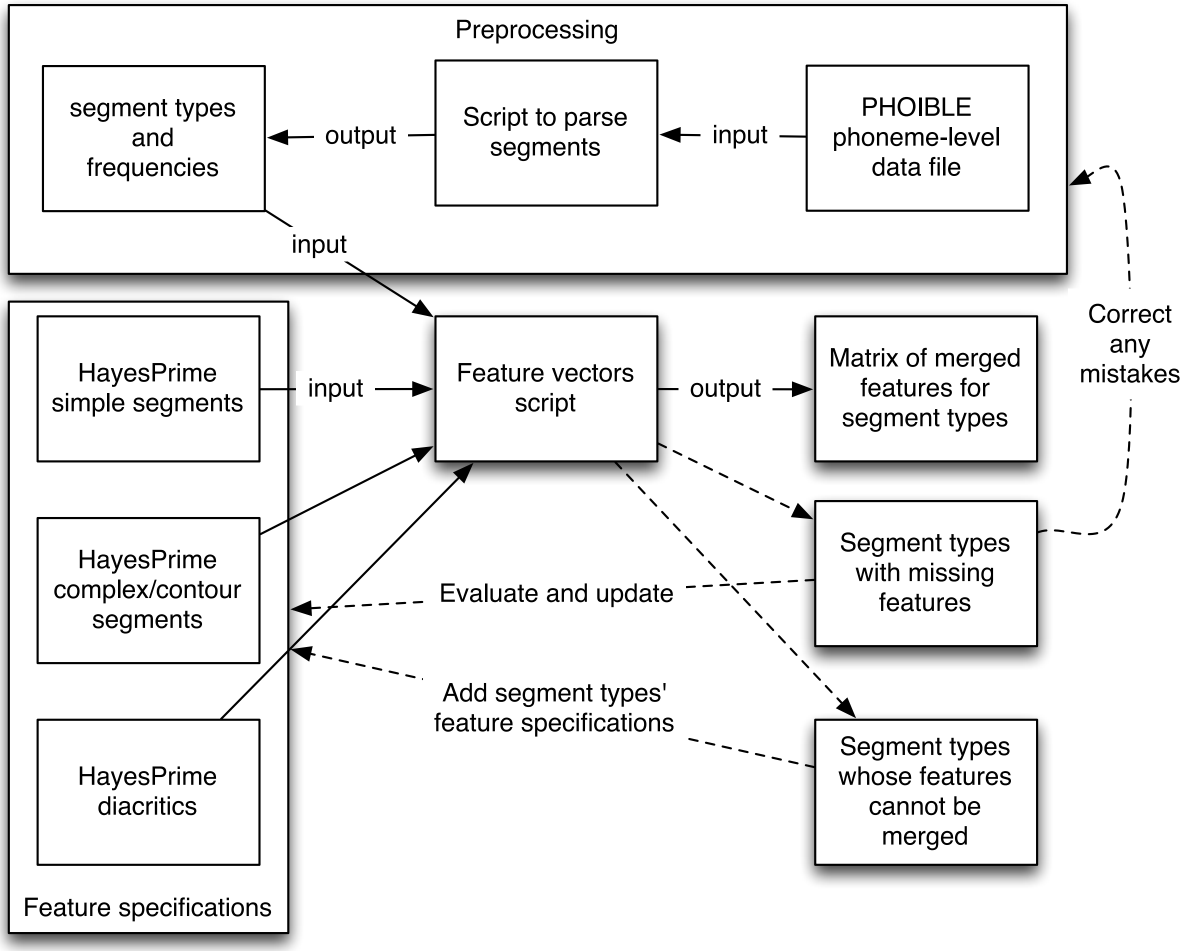
[24] See Section 5.6 for discussion.

[pf]. The feature vectors script then evaluates the input and it outputs minimally a matrix of successfully merged features for segment types. If the output includes segment types with missing features and/or segment types whose features cannot be merged, then the results must be evaluated. Complex/contour feature vectors may have to be assigned to segment types manually and any mistakes in the phoneme level data must be corrected.[25] As additional inventories are added to PHOIBLE with new segment types, this process can be rerun and the results reevaluated. It may be the case that additional characters, diacritics, their feature vectors (or even new features) will have to be added to the input feature set when new segment types are encountered.

Now that I have defined how to compositionally combine segments' feature vectors for the three segment classes, I can create a feature set that has complete segment type coverage of the PHOIBLE inventories. There are, however, several other challenges to address in this process. I take as my starting point the Hayes 2009 feature set and expand and adapt it as Hayes′. The first task is to identify characters in the PHOIBLE segment types that are not in the Hayes feature set and to add them. For some of the segments that are not in Hayes's feature set, but are in the IPA, this process is straightforward. For example, the open-mid central unrounded vowel [ɜ] receives the features of the open-mid central rounded vowel [ɞ], but instead it is specified [−round]. Assignment of features to implosives is also straightforward. Each implosive receives the features of its voiced plosive counterpart (e.g. [ɓ] and [b]) with the additional feature [+implosive]. The feature [implosive] is added to the entire set of features in Hayes′ and all sounds that are not implosive are marked [−implosive]. For clicks, a similar approach is taken. Each click has two parts. The first part, the [k], [g] or [ŋ] that precedes the click character determines the segment type's voicing or whether it is nasal. The click character specifies the segment type's place of articulation, e.g. the [ʘ] in [kʘ] is bilabial. The features of the click's plosive counterpart are assigned to the click segment type (in this example the features of the bilabial plosive are assigned to the segment). Finally, an additional feature [click] is added to the Hayes′ feature set. All clicks receive [+click]. All other segments are [−click].

[25] I use dotted lines to represent possible output and post-processing.

Figure 6.1: Process for creating feature vectors


Preprocessing
segment types and frequencies
output
Script to parse segments
input
PHOIBLE phoneme-level data file
input
HayesPrime simple segments
input
Feature vectors script
output
Matrix of merged features for segment types
Correct any mistakes
HayesPrime complex/contour segments
Segment types with missing features
Evaluate and update
Add segment types' feature specifications
HayesPrime diacritics
Segment types whose features cannot be merged
Feature specifications

For diacritics that were added to Hayes′, I followed Hayes's approach and I define which features of the base segment should be overwritten by a combining diacritic. For example, apical and laminal diacritics overwrite their base segment's feature [distributed]; apical segments receive [−distributed] and laminal segments receive [+distributed]. For segments that are advanced or retracted tongue root, another feature was added to Hayes′, the feature [atr]. The feature [atr] is specified not applicable for consonants and is specified "−" for all vowels unless otherwise overwritten by the advanced tongue root diacritic. For the SPA-specific "tense" and "lax" features, I added the feature [fortis] to Hayes′. In each case where another feature was added that is denoted by a diacritic, all other features remain non-applicable in feature assignment, i.e. a diacritic only overwrites the feature that it specifies.

Tone is a bit trickier. So far my approach has been to create a [tone] feature and to specify that all tones get [+tone] and all other segments are [−tone]. All other features are non-applicable to tonal segments. At this time, contour tones and downstep are treated the same as single tones, i.e. they receive [+tone]. They are not yet specified for additional features. Whether we need features for tone and if tones have features are issues raised in Clements et al. 2010 and Hyman 2010a. I leave the matter of what to do with tones and features for segment types in PHOIBLE for future work.

Another issue is how to handle archiphonemes that are underspecified for place of articulation, e.g. /N/ occurs often in segment inventory descriptions in West African languages. To tackle these, I underspecify the place of articulation features with a non-applicable “0”, as shown in Table 6.9. Although this reuses the symbol “0” for underspecified and non-applicable, in practice there is no ambiguity here about which interpretation is intended.

Table 6.9: Underspecified nasal segment /N/

	nasal	labial	round	labiodental	coronal	anterior	distributed	strident	dorsal	high	low	front	back
m	+	+	−	−	−	0	0	0	−	0	0	0	0
ɱ	+	+	−	+	−	0	0	0	−	0	0	0	0
n	+	−	−	−	+	+	−	−	−	0	0	0	0
ɳ	+	−	−	−	+	−	−	−	−	0	0	0	0
ɲ	+	−	−	−	+	−	+	−	+	+	−	+	−
ŋ	+	−	−	−	−	0	0	0	+	+	−	−	−
ɴ	+	−	−	−	−	0	0	0	+	−	−	−	+
N	+	0	0	0	0	0	0	0	0	0	0	0	0

Table 6.10 shows the feature specifications in Hayes's feature set that describe major natural classes of sounds. These feature combinations can be used to query the PHOIBLE knowledge base to investigate universals in phonology, which I show in Section 6.5.

Table 6.10: Feature specifications for natural classes of sounds

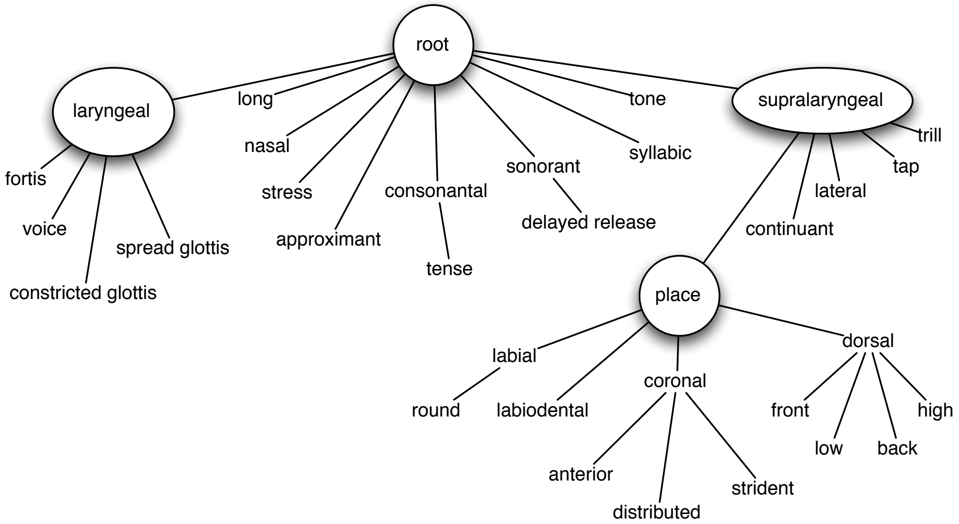
Class of sounds	Feature specification
Vowels	[+syllabic] [−consonantal]
Vowels & Syllabic Consonants	[+syllabic]
Glides	[−syllabic] [−consonantal]
Liquids	[+consonantal] [+approximant]
Nasals	[+sonorant] [−approximant]
Fricatives	[−sonorant] [+continuant]
Affricates	[−continuant] [+delayed release]
Stops	[−delayed release]
Stops & Affricates	[−continuant]
Liquids & Glides	[−syllabic] [+approximant]
Liquids, Glides, & Nasals	[−syllabic] [+sonorant]

So far I have described the specification of features in a matrix. In Chomsky and Halle 1968, distinctive features were organized into a two-dimensional matrix, where columns were functions that assigned segments to feature values and rows were phonetic features.[26] The implications of this matrix structure are that there are no overlapping features between segments, no ordering of features, and no internal hierarchical structure of features within segments. Although the matrix approach captures the existence of natural classes of sounds, there is also abundant evidence for an internal structure of features. A classic example is place assimilation, a phonological process that occurs widely cross-linguistically, e.g. nasal consonants often assimilate in place of articulation to the following consonant, but they tend not to change in their manner of articulation or to lose their nasality feature, etc.

[26] I gave an example in Table 2.2 on page 31.

Clements (1985) proposed hierarchically ordering features into a feature "geometry" to address such deficiencies of modeling features in matrices so as to handle temporal processes like assimilation and contour segments. Thus in a feature geometry, such as the one given in Figure 6.2, only features under the place node may be affected by a phonological process like nasal assimilation.

Figure 6.2: Hayes Prime feature geometry



Feature geometries have been proposed in various works, including: Clements 1985, Sagey 1986, Halle 1992, and Clements and Hume 1995. Figure 6.2 shows the Hayes′ feature geometry, which is informed from the logical relations that hold in the Hayes′ feature set. For example, any segment that is [−coronal] (e.g. all vowels), will have "0" for the features below the coronal node ([anterior], [distributed] and [strident]), since those features are non-applicable to [−coronal] segments. Other features, like [tone], [syllabic], [long], etc. that come off the root node apply to the entire segment. For example, a segment is either tonal or it is not.

At this point we can use OWL to model this taxonomy of features by defining the relationships among the features in the Hayes′ feature geometry. We can then model a feature geometry into the knowledge base and use it to query over classes of features.

Features modeled in RDF and hierarchically structured in OWL provide researchers with a mechanism for querying PHOIBLE's segment inventories at the level of features, natural classes and different levels in the feature geometry such as the "place" node that is not usually included as a feature in the feature set. The ability to query segment inventories at the level of segments *and* features allows us to investigate some of the claims made about phonological universals.

6.5 Investigating universals in phonological systems

In this section, I use the RDF/OWL knowledge base of PHOIBLE segment inventories and distinctive features from Hayes′ to revisit some of the universals of phonological inventories stated in Hyman 2008. Hyman distinguishes between descriptive universals, which minimize the effects of different theoretical frameworks, and analytical universals, which are theory-dependent. I will address the descriptive universals regarding segment inventories and features.[27]

In Sections 3.1.3 and 3.2.3 I provided an overview of RDF graphs and how they may be merged to combine data sets for querying. To investigate segment inventories at the level of features, the approach I take here is to combine two RDF graphs, namely the PHOIBLE segments and distinctive features graphs, into one combined RDF graph for querying. A portion of just the PHOIBLE segments graph is illustrated in Figure 6.3.[28]

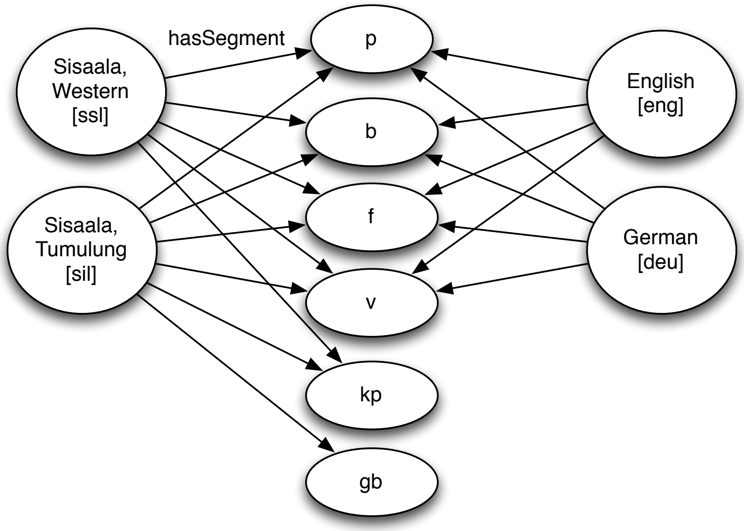
To review, if someone wants to query for the segments of a particular language, he or she could use a query like the one given in Example 6.1.

(6.1)
```
SELECT ?segments
WHERE { ssl hasSegment ?segments }
```

[27] With the computational tool for typological comparisons that I have built and its limitations, I cannot yet address other theory-dependent architectural universals (e.g. statements made within Optimality Theory) or universals dealing with tendencies above the segment level, e.g. universals regarding syllable structure (cf. Hyman 2010b). Future extensions that include theory-dependent information would allow us to investigate architectural universals.

[28] In this section I use a set of 1089 distinct languages and their inventories from the PHOIBLE data set using the trump hierarchy: PHOIBLE > SPA > UPSID > AA. Note that my results also hold on the entire PHOIBLE data set even when competing inventories for the same language are taken into account.

Figure 6.3: PHOIBLE segments graph



This query would return the segments { p, b, f, v, kp }, as illustrated in Figure 6.4, in which the matching segments are highlighted. If a user wants to search for languages that have a particular segment, this query can be stated as in Example 6.2. The result is illustrated in Figure 6.5.

(6.2)
```
SELECT ?languages
WHERE {
?languages hasSegment gb
}
```

Figure 6.4: PHOIBLE segments graph illustrating query results for segments in [ssl]

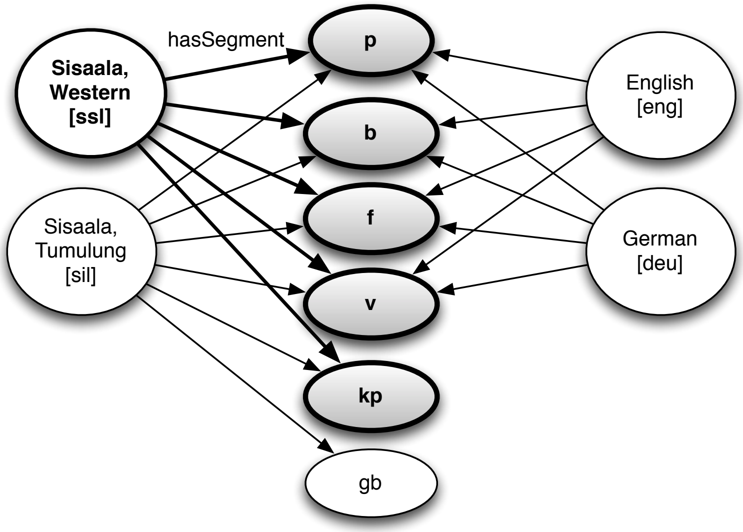
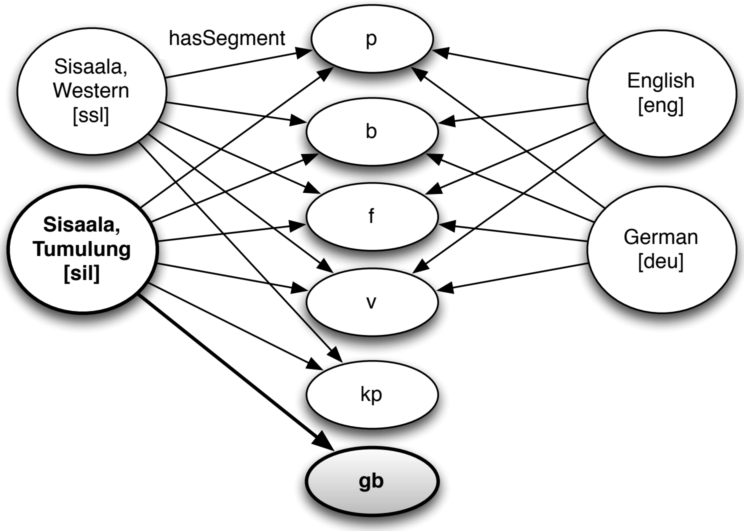
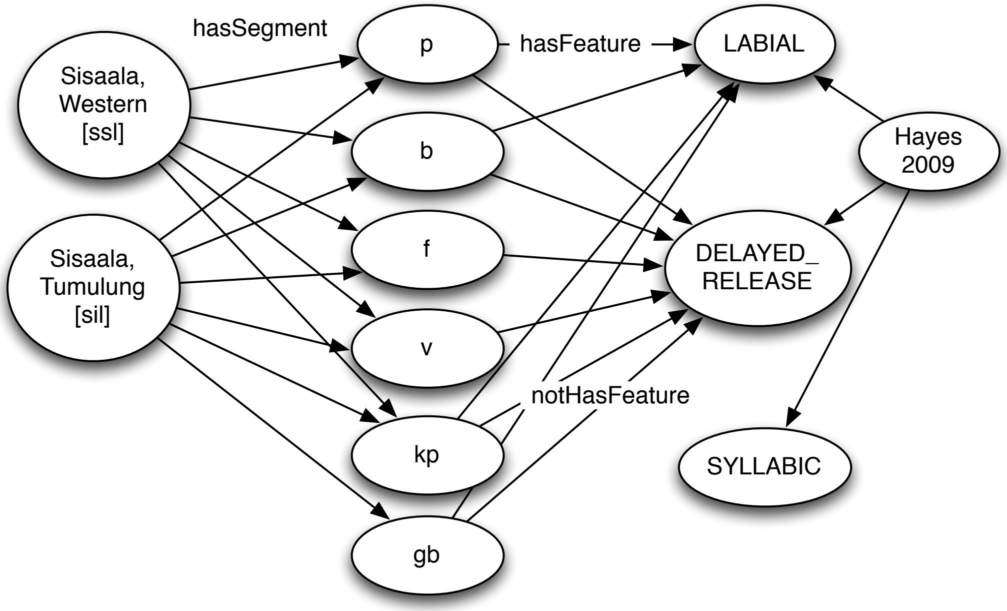


Figure 6.5: PHOIBLE segments graph illustrating query results for /gb/



These are some basic queries at the segment level. To expand the query functionality to the level of features, the PHOIBLE RDF/OWL graphs for segments and features are merged, as illustrated in Figure 6.6. Now a user can also query the merged graphs at the level of features.

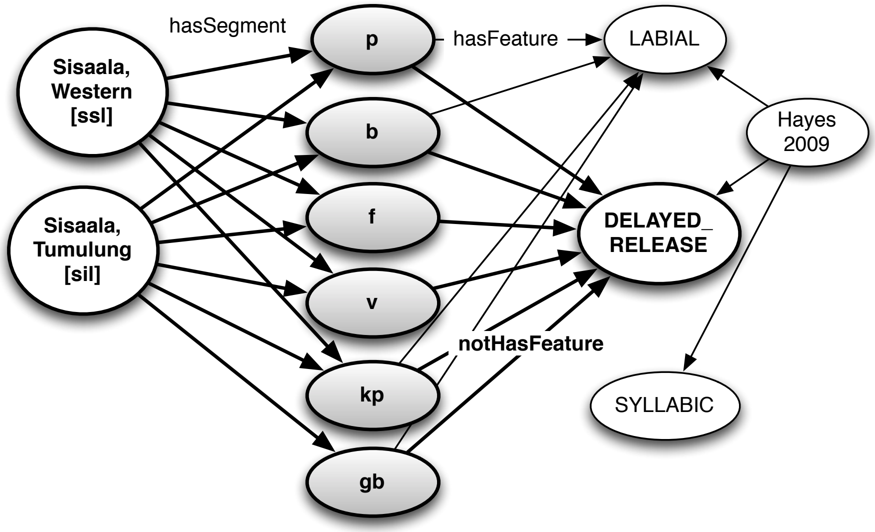
Figure 6.6: Merged PHOIBLE segments and features graph



Example 6.3 shows a SPARQL query to select languages that have a particular class of sounds. In this example, stops, which are [−DELAYED_RELEASE], are queried by selecting languages that have segments that are specified via the predicate *notHasFeature*, which connects segments to features that they do not have. The result of the query is illustrated in Figure 6.7.

(6.3)
```
SELECT ?languages
WHERE {
?languages hasSegment ?segments .
?segments notHasFeature DELAYED_RELEASE
}
```

Figure 6.7: Query result for stops on the merged segments and features graph



With the functionality to query segment inventories at the level of segments and features, we can easily investigate the proposed descriptive universals of phonological systems tested by Hyman (2008).[29] Let's start with universals in consonant systems. Hyman (2008, 92-94) posits that "every phonological system has stops" and that "every phonological system has coronal phonemes".

The SPARQL query already given in Example 6.3 queries for the first universal by selecting all languages that have segments that have the feature [−DELAYED_RELEASE], i.e. the class of all stops. Since every inventory in PHOIBLE has at least one segment that contains the feature [−DELAYED_RELEASE], all languages in the current data set have at least one stop. Therefore, the proposed universal that all languages have at least one stop holds in the PHOIBLE data set.

Next, the query in 6.4 searches the PHOIBLE data set for all languages that have a coronal phoneme.

[29] Hyman (2008) uses the UPSID$_{451}$ data in testing proposed phonological system universals. The inventory data were taken from Henning Reetz's online version of UPSID$_{451}$, at: `http://web.phonetik.uni-frankfurt.de/UPSID.html`.

(6.4)

```
SELECT ?languages
WHERE {
?languages hasSegment ?segments .
?segments hasFeature feature:CORONAL
}
```

This query follows the same pattern: it inspects all segments in all languages for the feature [+coronal] as specified in the predicates that connect segments and features. Again, if the number of results returned do not equal the number of total languages in the PHOIBLE data set, then there exists at least one language that does not adhere to the proposed universal. Indeed, the PHOIBLE data set contains counter-evidence to the universal, found in the segment inventory of Northwest Mekeo [mek] (Jones, 1995, 1998), which has the consonants: / p, β, m, w, ɡ, ŋ, j / but no coronals. In the UPSID$_{451}$ data set, all languages contain at least one coronal. Blevins (2009) was the first to report that Northwest Mekeo lacks coronals.

Another area to investigate descriptive universals is in vowel systems. Hyman (2008, 98) asks if "every phonological system has at least one unrounded vowel" and reaches the conclusion that no language in UPSID$_{451}$ has less than two unrounded vowels. A query to probe the data set for this universal is formulated in the SPARQL query given in example 6.5, using features from Hayes′.[30]

(6.5)

```
SELECT ?languages
WHERE {
?languages phoible:hasSegment ?segments .
?segments phoible:hasFeature feature:SYLLABIC .
?segments phoible:notHasFeature feature:CONSONANTAL .
?segments phoible:notHasFeature feature:ROUND
}
```

This query selects all languages that have segments that have the features [+syllabic], [−consonantal] and [−round], i.e. unrounded vowels. All 1089 languages are returned.

[30] Refer to Table 6.10 on page 276 for combinations of features that result in natural classes of sounds.

Therefore the universal "every phonological system has at least one unrounded vowel" still holds on the expanded PHOIBLE data set. The query can also be modified to return all languages and their segments, shown in example 6.6. This query returns all languages and their unrounded vowels.

(6.6)
```
SELECT ?languages ?segments
WHERE {
?languages phoible:hasSegment ?segments .
?segments phoible:hasFeature feature:SYLLABIC .
?segments phoible:notHasFeature feature:CONSONANTAL .
?segments phoible:notHasFeature feature:ROUND
}
```

Hyman (2008, 98) also postulates that "every phonological system has at least one back vowel". Again, querying the RDF graph is straightforward, as shown in Example 6.7. This universal also holds on the expanded number of inventories in the PHOIBLE data set.

(6.7)
```
SELECT ?languages ?segments
WHERE {
?languages phoible:hasSegment ?segments .
?segments phoible:hasFeature feature:SYLLABIC .
?segments phoible:notHasFeature feature:CONSONANTAL .
?segments phoible:hasFeature feature:BACK
}
```

Another universal investigated by Hyman (2008, 98) is that "every phonological system has at least one front vowel or the palatal glide /j/".[31] This can be asked of the PHOIBLE knowledge base by using the SPARQL UNION operator to query all languages that have segments of a particular feature make-up ([+SYLLABIC, −CONSONANTAL, +[ROUND]) *or* the segment /j/. This universal also holds in the PHOIBLE data set. The addition of a logical

[31] Note that Hyman uses the symbol <y> for the palatal glide. Here I use the IPA symbol <j>.

operator in the query illustrates just one of the many features of the SPARQL language that can be used to query the PHOIBLE data set.[32]

(6.8)
```
SELECT DISTINCT ?languages
WHERE {
?languages phoible:hasSegment ?segments .
?segments phoible:hasFeature feature:SYLLABIC .
?segments phoible:notHasFeature feature:CONSONANTAL .
?segments phoible:hasFeature feature:FRONT .
UNION {
?languages phoible:hasSegment segment:j
 }
}
```

To summarize, in this section I have shown how the PHOIBLE segment and feature RDF/OWL graphs are merged and how they can be queried at the level of segments and features. I use the SPARQL graph query language to investigate descriptive universals proposed by Hyman (2008). The results are given in Table 6.11.

Table 6.11: Descriptive universals in phonological systems

Every phonological system has...	UPSID$_{451}$	PHOIBLE
stops	yes	yes
coronal phonemes	yes	no
at least one unrounded vowel	yes	yes
at least one back vowel	yes	yes
one front vowel or the palatal glide /j/	yes	yes

[32] See the SPARQL documentation for a full list of operators, functions, modifiers, etc.: `http://www.w3.org/TR/rdf-sparql-query/`.

6.6 Conclusion

In this chapter I have set out to develop a mechanism for examining segment inventories at the level of distinctive features to investigate claims of descriptive universals in phonological systems. To do so, I began by summarizing the issues regarding segments, features and the lack of typological representation of features sets in regard to segment types. I developed a method for compositionally combining feature vectors to automatically derive features for segment types in the PHOIBLE data set that are not defined in Hayes 2009, which has the most comprehensive coverage of the IPA. I have shown that complex segments pose a serious challenge to automatic feature vector assignment because their component segments' feature vectors may not logically combine the way that segments and diacritics do. Contour segments also pose a challenge due to their temporal encoding of features changing through time. I proposed two ways of encoding contour segments for analysis.

Lastly, I used the system I have developed in this chapter to query segment inventories at the level of segments and features to revisit some of the descriptive claims that have been made about universals in phonological systems. I have shown that at least one of these claims, namely that all phonological systems contain at least one coronal phoneme, does not hold on the extended PHOIBLE data set.[33] There are other assumptions about segment inventories that are also important to test. However, I have not yet undertaken these studies. For example, one assumption is that languages with more fricatives will have a higher number of consonants overall. The data to address this question are easily attained from the PHOIBLE knowledge base by querying inventories for fricative and consonant segment types: for each inventory, get its number of fricatives and consonants by querying for all segments that are [−SONORANT], [+CONTINUANT] and [+CONSONANTAL], respectively. These queries would be difficult, or at least time-consuming, at the level of segments because a list of all fricative and consonant segment types in PHOIBLE would have to be identified and there are currently over 1700 segment types. Thus with PHOIBLE's inventories and the Hayes′ feature set, the technological infrastructure is in place for researchers to investigate

[33] This of course also boils down to a question of analysis. If an inventory in PHOIBLE is listed as not having some set of segments or features, then the output of a query will show just that.

many aspects of phonological systems and how they pattern. Going forward, certain claims are gaining attention nowadays that have to do with proposed correlations between certain aspects of phonological systems and non-linguistic factors. PHOIBLE is also an appropriate tool and data set to revisit these claims, as I will show in the next chapter.

Chapter 7

CASE STUDY: PHONEME INVENTORY SIZE AND POPULATION SIZE[1]

7.1 Introduction

Studies of the relationship between linguistic systems and the environment in which they are spoken date back at least a century. Sapir (1912) delineated environmental influences on language by physical (e.g. topography, climate, flora and fauna, etc.) and social factors (e.g. religion, ethics, politics, etc.). He showed that certain non-linguistic contexts clearly favor enrichment of the lexicon, evidenced by the uneven distribution of domain-specific vocabulary in languages in relation to the importance of their environments (cf. Nettle 1999b).[2] Apart from environmental influences on vocabulary, however, Sapir reported linguistic structure is not shown to be directly affected by environmental influences.

Recently an increasing amount of research utilizing statistical methods and typological data sets has challenged the view that changes in language structure are not purely linguistically driven, i.e. through language contact or recurrent processes of linguistic change. A controversial line of research associates changes in linguistic structure with ecological or demographic factors (Nettle, 2007). This research suggests that some typological patterns (be they synchronic or diachronic) may be related to (or even a consequence of) environmental or societal factors. For example, Nettle (1996) argues that the degree of ecological risk plays a role in shaping linguistic diversity in West Africa and presents evidence that correlates linguistic diversity with topography. By correlating climate information with phonological systems, researchers have claimed to have shown that languages spoken in warm climates use relatively more high-sonority sounds than languages spoken in cold climates (Munroe et al.,

[1] A version of this chapter will appear in *Language* as “Revisiting population size vs. phoneme inventory size” (Moran et al., to appear).

[2] For example, the Dogon in Mali distinguish between about 20 local grasshopper species. *Kraussaria angulifera* is an especially tasty variety when salted and roasted (Jeff Heath, p.c.).

1996; Munroe and Silander, 1999; Fought et al., 2004; Ember and Ember, 2007; Munroe et al., 2009). Ember and Ember (1999, 730) argue that the "degree of baby-holding is more predictive of CV scores [the percentage of CV syllables in the average word] than either climate or literacy". Lupyan and Dale (2010) find that languages with smaller groups of speakers have more complex inflectional morphology than languages spoken by large groups. And Hay and Bauer (2007) claim that there exists a robust correlation between population size and phoneme inventory size.[3]

What these studies have in common is that they use small and biased data sets that limit the type of statistical methods that can be used.[4] For example, Hay and Bauer (2007) use a convenience sample of 216 languages that includes coverage for 46 language families, but 38 of those contain five or fewer languages; most include just one or two.[5] Since a large number of groups (language families) in their sample include just one language, it is difficult to apply statistical mixed models to their data. Furthermore, their sample has radically unequal group sizes, which is problematic for many statistical tests, e.g. ANOVAs (Stevens, 2009, chap. 6). In this chapter I argue against the findings presented in Hay and Bauer 2007 by using PHOIBLE, a much larger and more diverse sample of the world's languages, which allows for more nuanced statistical techniques. Using a hierarchical linear model (a mixed model that is appropriate for nested data), I show that the correlation between population size and phoneme inventory size does not hold once the genealogical relatedness of languages is accounted for.

The PHOIBLE data set can also be used to assess other claims that I have mentioned. For example, Fought et al. (2004) and Munroe et al. (2009) use a very small sample of 60 languages to report that languages spoken in warm climates use relatively more sonorant sounds than those spoken in cold climates. As discussed in Chapters 3 and 6, the PHOIBLE data set contains geographic data for each segment inventory and each segment is associated

[3] If there exists a correlation, direction of causation is a valid question. However, it seems unlikely that language structure influences the environment (Kaye, 1989) or phonemic inventory size affects population size.

[4] Lupyan and Dale 2010 is an exception. The authors' data set includes an impressive 2236 languages.

[5] The remaining eight families are represented by the following number of data points: 6, 6, 7, 8, 11, 17, 26, 50.

with a vector of distinctive features. The geographic data can be used in coordination with the coding of climate in Fought et al. 2004 and Munroe et al. 2009 to determine which climate a language belongs to. The distinctive features can be used to categorize segments in the sonority classes proposed in these same works.

It is important to question the findings of studies that use small sample sizes because their claims may influence, or even become axioms, for further research. For example, a recent (and popular) proposal by Atkinson (2011, 346) begins, “The number of phonemes – perceptually distinct units of sound that differentiate words – in a language is positively correlated with the size of its speaker population [Hay and Bauer 2007] in such a way that small populations have fewer phonemes.” Atkinson goes on to report a negative correlation between phoneme inventory size of a language (what he calls “phonemic diversity”) and its geographic distance from West Africa, which he argues supports a single language origin in Africa via a repeated founder effect that accompanied the migration of modern humans. However, this claim crucially depends on a positive correlation between phonemic inventory size and speaker community size.

This chapter is structured as follows. In Section 7.2, I provide an overview of the studies on population size and phoneme inventory size that led up to Hay and Bauer 2007. I then discuss Hay & Bauer’s study and their findings in Section 7.3. In Section 7.4, I give an overview of the materials and method used in my study and I give my analysis and results. Section 7.5 compares my study with Hay & Bauer’s and it provides a discussion of methodological considerations in regards to typological data sets and quantitative methods. My concluding remarks are given in Section 7.6.

7.2 Previous studies

Previous studies that investigate a correlation between population size and phoneme inventory size are either speculative (they suggest a correlation based on some examples), computer simulated through models (population size affects rate of linguistic change and thus can affect the size of a language’s phonemic inventory), or empirical (population figures and phoneme inventory sizes are fed into a statistical model and examined for correlations). The initial studies were speculative.

A correlation between the size of a phoneme inventory and the number of speakers of that language is suggested at least as early as Haudricourt 1961. Haudricourt argued that small inventories are the product of impoverishment that is characterized by monolingualism, isolation, and/or by non-egalitarian bilingualism (Haudricourt, 1961; Trudgill, 1997, 2002; Hay and Bauer, 2007).

This issue was revived in Trudgill's (1996, 1997) studies on the effect of community size on linguistic structure, in particular on aspects of phonology. Trudgill (1997, 356) proposes a typology of three situations that lead to different sizes of segment inventories:

1. Isolated low-contact languages such as, to take the most extreme case, Hawai'ian, with small inventories.

2. High-contact languages where contact is long-term and involves child bilingualism such as, to take the most extreme case, Ubykh, with large inventories.

3. High-contact languages where contact is short-term and/or involves imperfect language-learning by adults such as, to take the most extreme case, pidgins, with small inventories.

Noting that his approach is speculative, Trudgill suggests that the distribution of typological characteristics may be affected by certain social characteristics of societies, such as their social network structure, the amount of shared information among speakers, and community size.[6] These factors are theorized to affect linguistic change, which in turn leads to observable differences in languages, e.g. the prediction that isolated communities have smaller inventories.

Whereas Trudgill's work is theoretical and speculative, Nettle (1999a,c) is the first to investigate the effect of community size on language change empirically by creating computer simulations. Nettle (1999a,c) designed a conceptualization of the process of language

[6] This work built on previous work by the same author. Trudgill (1974) introduced the gravity model from geography to dialectology, quantifying the amount of diffusion between two dialects as proportional to the product of two populations divided by their distance squared. In Trudgill's model, diffusion of linguistic change cascades from large population centers to smaller ones and so on.

evolution by modeling a population that learns one of two competing variants of the same grammatical item. His model draws on an adapted version of Social Impact Theory (SIT) (Latané, 1981; Nowak et al., 1990) in which the simulation of language change in social networks is measured by the percentage of individuals who adopt one of the grammatical item variants. Nettle (1999c, 115) manipulates settings in the SIT model, including the rate of mutation, weighting of social distance and the effect of majority consensus on impact and concludes that "changes are adopted because some speakers are much more influential than others as social models". Based on his simulations, Nettle argues that as a population gets larger, borrowing and the emergence of marked structures are less likely to occur. The rate of language change is therefore slower. The underlying idea is that an innovation can spread more easily and more quickly over a small group of speakers than within a large group.

Wichmann et al. (2008) revisit Nettle's results, but whereas Nettle's simulation modeled competition between only two languages or linguistic features (the original and the novel forms), Wichmann et al. used a simulation model that allowed several competing languages, each with several linguistic features, to compete simultaneously. Their study used an extended language model, which is the Schulze model (Schulze and Stauffer, 2005; Schulze et al., 2008) combined with a network as described in Barabási and Albert 1999. Wichmann et al. also analyzed a sample of 2140 languages with data from the World Atlas of Language Structures (WALS; Haspelmath et al. 2005) and language statistics, including population figures, from the Ethnologue 15th edition (Gordon, 2005). They estimated the stability of each of 134 WALS features and used the stability of features to estimate the rate of linguistic change for each language. The results from their study suggest that speaker population has no correlation with rate of linguistic change. The simulations showed both the presence and absence of some correlation, depending on whether linguistic diffusion was allowed to be global or if it was constrained to near neighbors in the social network. In more recent work, Wichmann and Holman (2009, 272) test several different empirical data sets and statistical methods and their findings, "strongly indicate that the sizes of speaker populations do not in and of themselves determine rates of language change". Compared to other factors involved in language change, they report that population size has a negligible effect. In light of these conflicting results of whether population sizes affects language change, we are still left with the question of whether speaker population and phoneme inventory size are correlated.

Trudgill (2002, 2004a) investigates societal features (contact, social network structure and stability) and their effects on linguistic patterning. In his words, “The issue at hand is whether it is possible to suggest that certain linguistic features are more commonly associated with certain types of society or social structure than others” (Trudgill, 2002, 708). Trudgill (2004a) investigates if there is any connection between the relative isolation of speakers of Austronesian languages and loss of consonants in those languages. As Austronesians expanded further into uninhabited Pacific islands, isolation and small community size are suggested as two factors that decrease phoneme inventory size. Small community size leads to tight social networks, implying greater shared background information, thus “a situation in which communication with a relatively low level of phonological redundancy would have been relatively tolerable” (Trudgill, 2004a, 315). On the other hand, as Trudgill points out, small isolated communities like the !Xũ speakers display extremely large phonemic inventories.

Noting the absence of large-scale typological databases for empirical study, Trudgill reaches the following tentative conclusions regarding the effects on phoneme inventory size due to language contact, isolation and community size (Trudgill, 2004a, 317):

1. long-term language contact that involves child language acquisition and high degrees of language contact may lead to larger phoneme inventories through borrowing

2. medium-sized phoneme inventories are favored by situations involving adult language contact (“i.e. inventories which are not so large as to be difficult for adolescents and adults to remember and acquire, but not so small as to cause confusability of constituents and high word length”)

3. situations with low degrees of language contact may lead to small inventories (“because the memory load difficulties caused by confusability and word length will not be relevant, since post-critical threshold learning is not involved”) or large inventories

because "the memory load difficulties caused by the acquisition of large numbers of phonemes will not be relevant"

4. large community size favors medium-sized phoneme inventories because such inventories "are not so small as to cause communicative difficulties as a result of a low degree of redundancy"

5. languages spoken by small communities may lead to very small inventories because "lower degrees of redundancy can be tolerated because of the large amounts of shared information present" or they may lead to very large inventories ("because of the ability of such communities to encourage continued adherence to norms from one generation to another")

Both Bakker (2004) and Pericliev (2004) test Trudgill's claims empirically. Bakker targets Trudgill's claims about the effects of language contact on phonological inventories. His study is effectively a series of case studies designed to shed light on outliers with respect to Trudgill's hypotheses. Bakker concludes that although a language learned by a group of second language learners, and subsequently passed down to new generations, loses some of its grammatical complexities and irregularities, there may not be any simplifying effect on the phoneme inventory because processes like pidginization and creolization do not significantly decrease segment inventories. Bakker is skeptical of Trudgill's thesis.

Pericliev (2004) takes aim at Trudgill's claims about correlations between community size and phonological inventories.[7] He strikes directly at Trudgill's explanation and methodology, using a well-defined approach, targeting these two specific claims (Pericliev, 2004, 376):

1. Large community size favours medium-sized phonological inventories.

2. Small (=non-large) community size favours either small phonological inventories or large inventories (but not medium-sized ones).

[7] Pericliev (2004, 377) focuses on consonant inventories because "judging from the context and the examples Trudgill gives, by 'phonological inventory' he means the consonantal inventories of languages (rather than inventories including both consonants and vowels)".

If the universe of all inventories is exhaustively split into three groups (small, medium and large), and all language-speaking communities are divided into small and large, then some categorization of the data should allow testable hypotheses. Pericliev turned to the UPSID$_{451}$ database to test these claims cross-linguistically and augmented its inventories with population figures from the Ethnologue.[8]

Trudgill (2004b) does not define numerically the range for small, medium or large speaker communities or phoneme inventory sizes, thus the claims are not well defined for testing. Pericliev (2004, 378) decides to investigate the claims in two ways. In the first, he redefines Trudgill's two claims (above) as:

1. Community sizes $>$ 5,000 speakers (large ones) favour inventories between 13 and 31 consonants inclusive (i.e. medium-sized ones).

2. Community sizes $\leq$ 5,000 speakers (small ones) favour either less than 13 consonants (small inventories) or more than 31 consonants (large inventories).

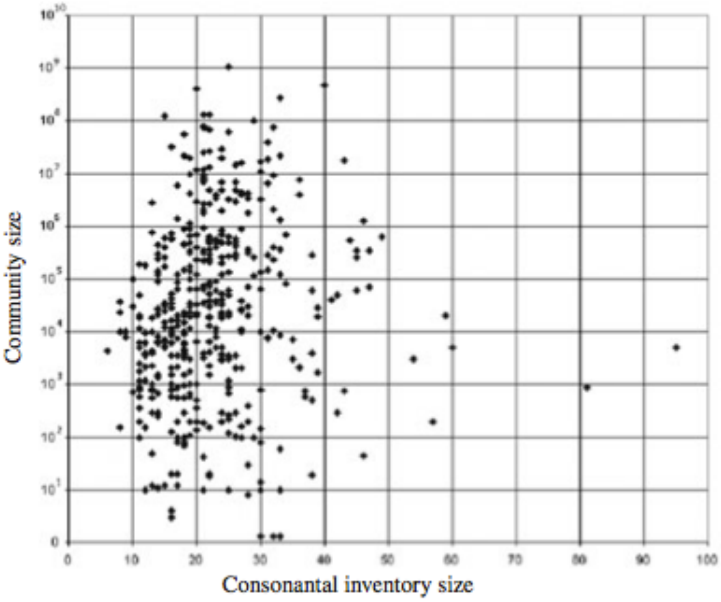
These figures are derived by taking the mean of consonants in inventories in the UPSID$_{451}$ data (22) and one standard deviation (9), so 22 ± 9 is considered an average size for a consonant inventory. For community size, Pericliev split small and large communities at 5000 speakers. He then uses these demarcations to randomly select languages from the UPSID$_{451}$ sample and test Trudgill's claims. He finds that the results based on a suite of random tests are valid around or below 50% of the time, which suggests there is no linguistic preference of the types suggested by Trudgill.

Pericliev's second approach uses a graphical test that plots languages from the UPSID$_{451}$ sample in an xy scatter diagram, reproduced in Figure 7.1. Each point on the graph represents the size of the consonantal inventory (x axis) by population size. The graphical test shows no trace of three distinct regions corresponding to small, medium or large inventories. Pericliev juxtaposes Figure 7.1 against a graphical representation in which he generates an artificial language sample that conforms to Trudgill's claims, reproduced in Figure 7.2. He

[8] Pericliev's sample size did not include 23 languages from UPSID$_{451}$ because they were either extinct or population figures did not exist in the Ethnologue (Pericliev does not cite a specific version of Ethnologue).

concludes there is no correlation between the size of a community of speakers and the size of the consonant inventory in that language.[9] Both studies by Bakker and Pericliev cast serious doubt on the patterns Trudgill hypothesizes.

Figure 7.1: Distribution of languages by consonant inventory size and community size (Pericliev, 2004, 382)

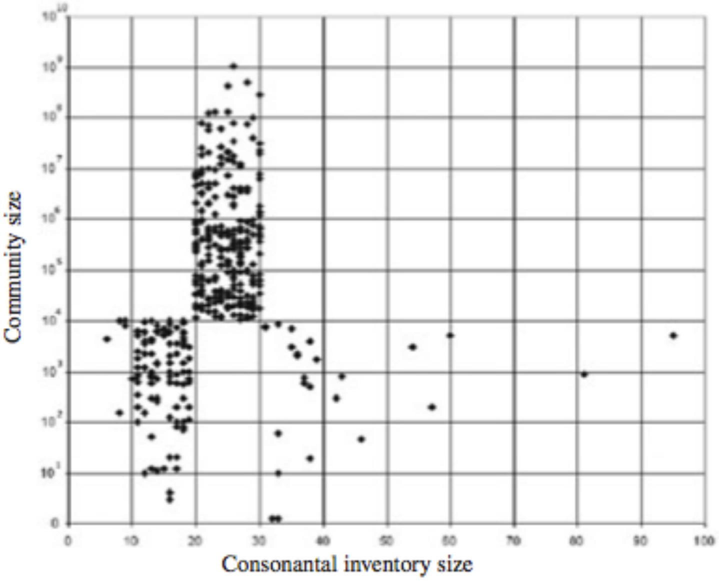


7.3 Hay & Bauer

In contrast to Pericliev's conclusion, Hay and Bauer (2007) find a correlation between phoneme inventory size and population size. Their data set is drawn from Bauer 2007, which includes a list of 250 languages and information regarding where the language is spoken, its genealogical affiliation, number of speakers and typological features. Since the data source is a textbook aimed at linguistics students, the sample is purposely not random and includes major and well-described languages, as well as some near extinct languages

[9] Pericliev reports that preliminary tests with whole inventories, i.e. consonants and vowels, also do not correlate with Trudgill's hypotheses.

Figure 7.2: Distribution of an artificial language sample confirming to Trudgill's claims (Pericliev, 2004, 381)



including isolates and languages of linguistic interest (Bauer, 2007, 221).[10]

Hay & Bauer's analysis does not include languages without living speakers, so the sample size represents a total of 216 languages. Hay & Bauer removed two extreme outliers, !Xũ [ktz] for total consonants and Acooli [ach] for total monophthongs because their values were more than four standard deviations above the mean. They used the log of the population to minimize the effect of outliers in speaker populations (Hay and Bauer, 2007, 389). Each language in Bauer 2007 is associated with a language family (its stock and sometimes also genus). The sample, the genealogical coverage of which is illustrated in Figures 7.5 and 7.6 on pages 301 and 302, is biased towards Indo-European and Pacific languages. Nonetheless, the data set presents a geographically diverse sample of the world's languages.

Hay & Bauer find correlations between speaker population and various measures of phonological inventory size. They use LOWESS (locally weighted scatterplot smoothing)

[10] Information in the textbook, such as population figures that often diverge by 100% or more as reported in different sources, should be thoroughly rechecked for testing hypotheses (Bauer, 2007, 222-224).

for curve-fitting with significance assessed by Spearman's rank correlation coefficient (Spearman's rho). The correlations they report are modest; the Spearman's coefficients range from 0.17 to 0.37. Figure 7.3 shows the significant correlations of log population size with the inventory size of obstruents, sonorants, consonants and total phonemes.

Figure 7.3: Association between population size and inventories (Hay and Bauer, 2007, 390)

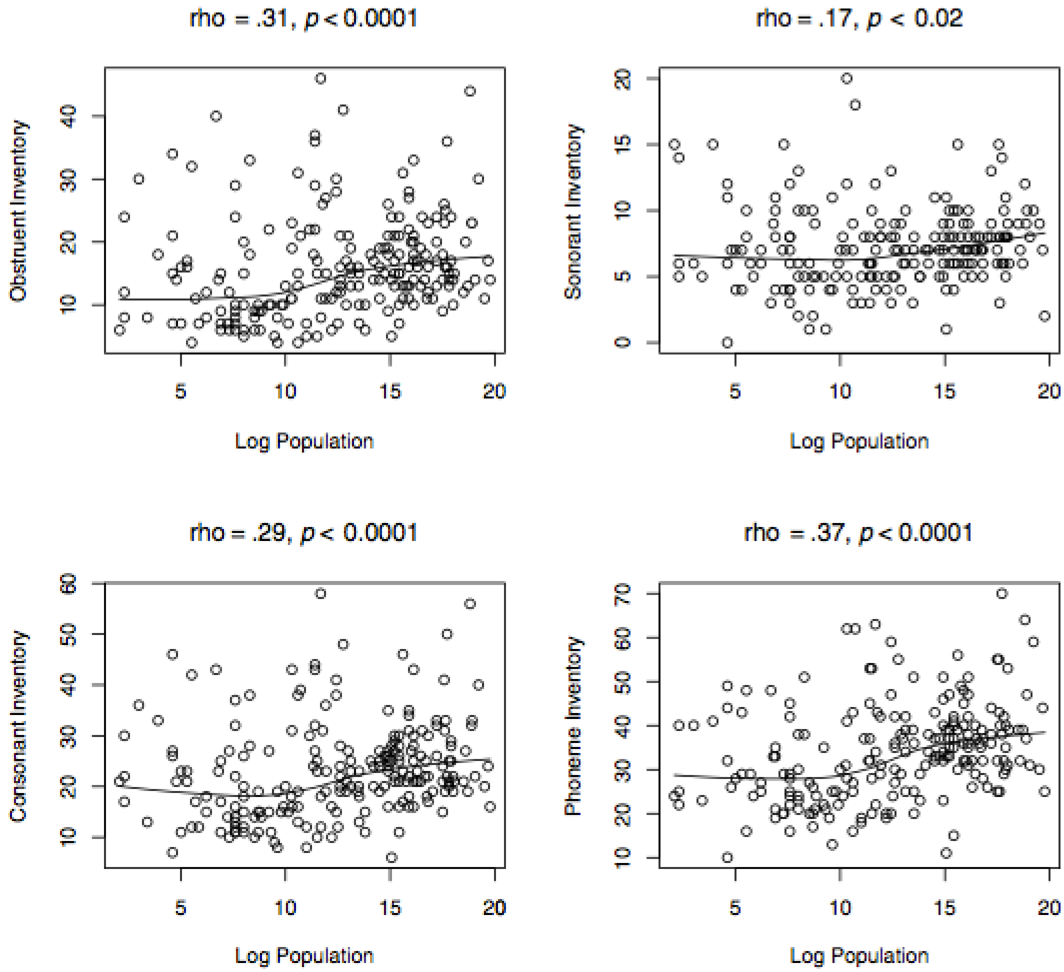
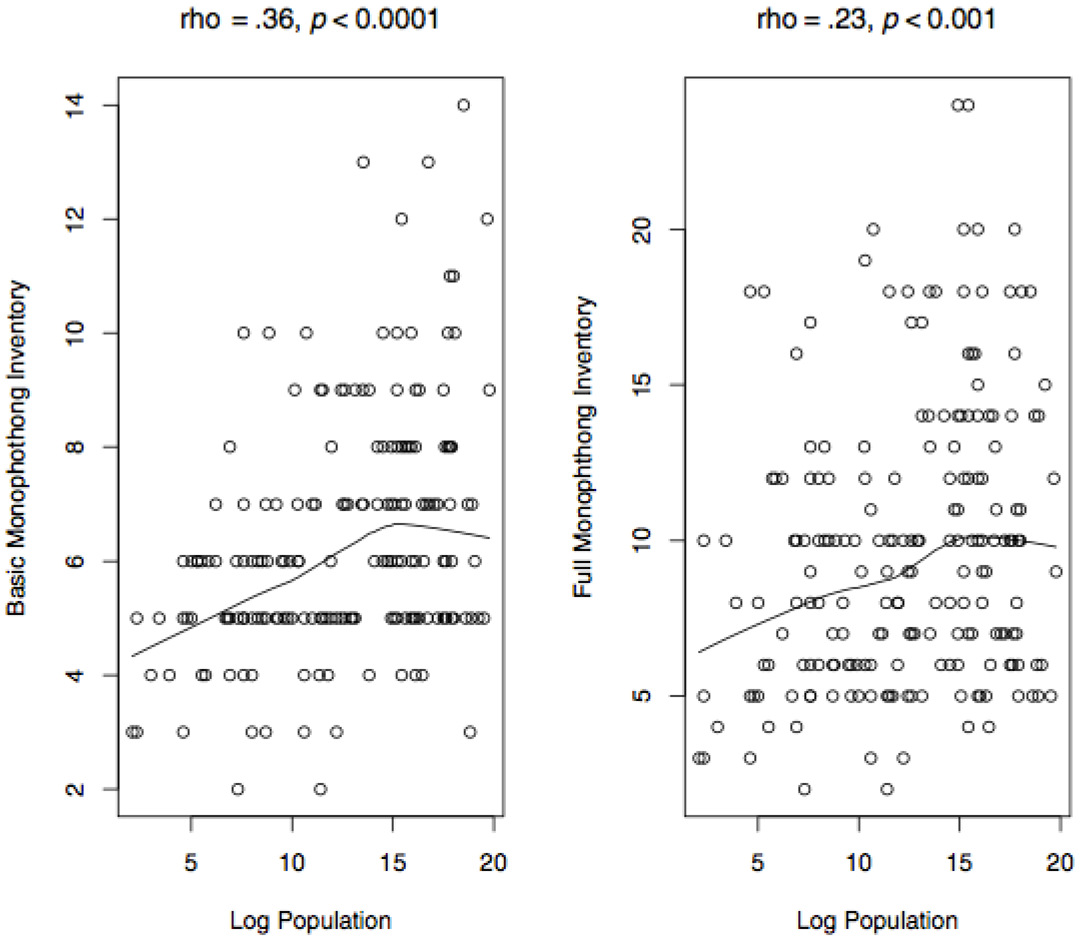


Figure 7.4 shows the positive association between the log population of speakers and vowel inventory. The left panel includes only basic monophthongs and the right panel includes the full monophthong inventory.[11] The tighter correlation in the left figure may be

[11] Hay & Bauer distinguish between basic monophthongs and extra monophthongs (i.e. vowels consisting

because monophthongs are more likely to be consistent across different researcher's analyses.

Figure 7.4: Association between population size and vowel inventory (Hay and Bauer, 2007, 390)



By using LOWESS curve-fitting with significance assessed by Spearman's rho, Hay & Bauer's method assumes that languages are independent. However, with regards to independence of observations, languages within a given language genus or stock are much more likely to have similar inventories than languages drawn from different families. Thus, their method does not take into account the problem of data nesting. Hay & Bauer attempt to

of nonquality distinctions such as length, nasalization, etc.) because linguists are more consistent in their descriptions of monophthongs. See Section 2.3.4 for discussion.

control for nested data by running two additional statistical tests. First, each family in their data set with sufficient representation was added to a multiple linear regression model as a categorical predictor. Hay & Bauer choose seven languages as the minimum cutoff for inclusion to preserve the needed degrees of freedom in their model. This results in five language families as predictors: Altaic, Austronesian, Indo-European, Niger-Congo, and Penutian. Their results show the Austronesian family as a significant predictor of phoneme inventory size. The variance seen in other language families, however, is too great to conclude if language family is a significant predictor. In Hay & Bauer's model, population size is a separate significant predictor. In multiple linear regression analysis, however, when the assumption of independence is violated, the analysis may be incorrect or misleading (Stevens, 2009). Furthermore, the language family groups are unequal in size: 23% of the languages in Hay & Bauer's sample are Indo-European[12] and 44% fall into their "other" category. This overrepresentation may have biased their results. The authors try to account for these biases by random regression resampling of the data, which they run 200 times. Although this may have removed any bias due to individual languages, resampling alone is likely insufficient to remove the strong Indo-European bias in their data set.

Hay & Bauer's second statistical test to control for data nesting attempts to account for the influence of language family. Each family is reduced to a single data point, which is comprised of the average speaker population and the average phoneme inventory size from each language family present in their data. This reduces their sample from 216 languages to 46 language family stock-level data points. The independence of observations is irrelevant here because no genealogical relationships have been established between stocks. Therefore, the issue at hand is sampling bias. What is the representative coverage of the sample? How many language families are included? Which ones? Within each language family, how many languages are represented? And which ones? In Hay & Bauer's sample, for example, the Austronesian language family includes only Malayo-Polynesian languages, excluding all Formosan languages.[13] Formosan languages go against the correlation under investigation.

[12] Only 6.4% of all languages in the Ethnologue are listed as Indo-European.

[13] Thanks to Dan McCloy for pointing this out.

They have large phonemic inventories, but small populations of speakers.

To summarize so far, effects of population size on phonemic diversity are equivocal. There are arguments for a correlation between population size and phoneme inventory size (Haudricourt, 1961; Trudgill, 1997, 2002, 2004a) and in recent years the correlation has been tested empirically with computer simulations and with statistical methods on typological data sets. A relationship between population size and rate of language change, which could lead to patterning of different sized phoneme inventories, has been shown to exist and not exist (Nettle 1999a and Wichmann and Holman 2009, respectively). And a correlation between population size and phoneme inventory size has also been shown to exist (Hay and Bauer 2007) and not exist (Bakker 2004; Pericliev 2004). Inspired by Hay & Bauer's unexpected results, I decided to retest their findings on a much larger data set to test whether the correlation is an artifact of their statistical method.

7.4 Materials and analysis

For this study, data was drawn from the PHOIBLE database. Some languages are represented in the database multiple times, either as descriptions of different dialects of the same language, different analyses of the same dialect, or different interpretations of the same linguistic description.[14] Therefore, duplicate inventories were removed using a "trump hierarchy".[15] After duplicates were removed, 1089 unique languages were grouped into 100 top-level language families (stock) available from the Ethnologue (Gordon, 2005) and retrieved via Multitree.[16] I have excluded pidgins, creoles, and ancient, extinct and mixed languages.[17] Additionally, languages for which there is no population figure available are not included. This left 984 languages which are used in my analysis. Figures 7.5 and 7.6 illustrate the genealogical coverage of the PHOIBLE and Hay & Bauer samples against the

[14] See discussion in Chapter 4, specifically Section 4.3.4.

[15] See Section 3.2.2.

[16] See Section 4.4 for details.

[17] Nineteen mixed languages are listed in the Ethnologue. A mixed language is the product of the fusion of two languages by speakers fluent in both languages, e.g. Michif [crg]. Different definitions of "mixed language" include or do not include pidgins and creoles. See: `http://www.glottopedia.de/index.php/Mixed_language`.

Ethnologue.[18] The PHOIBLE sample is a better representation of languages, especially for the Niger-Congo family.

[18] For ease of readability language families are ordered by increasing representation in PHOIBLE. I use the Multitree four-letter language family codes, except for language isolates, which are ISO 639-3 codes preceded by an underscore.

Figure 7.5: Percentage of Ethnologue entries represented in PHOIBLE and Hay & Bauer 2007

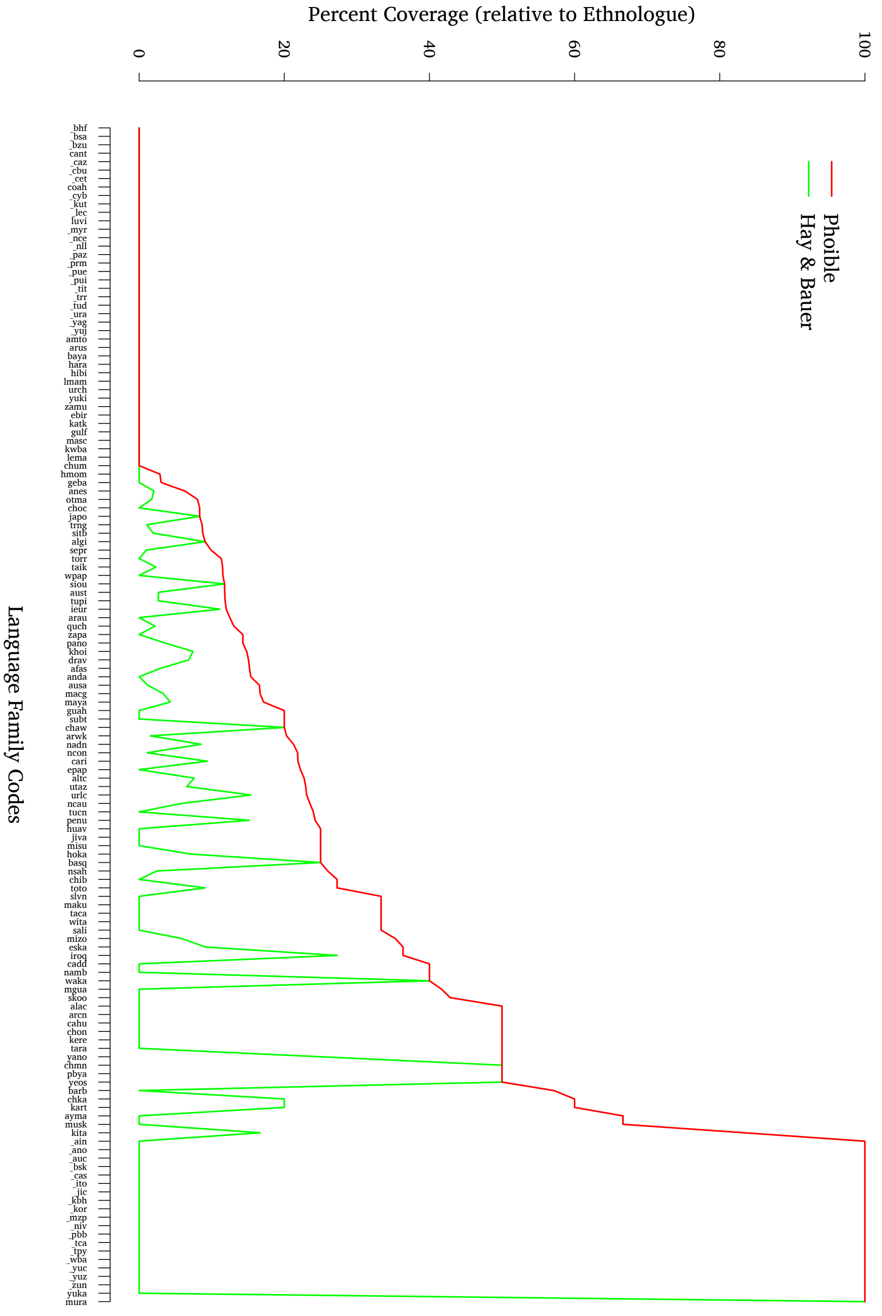
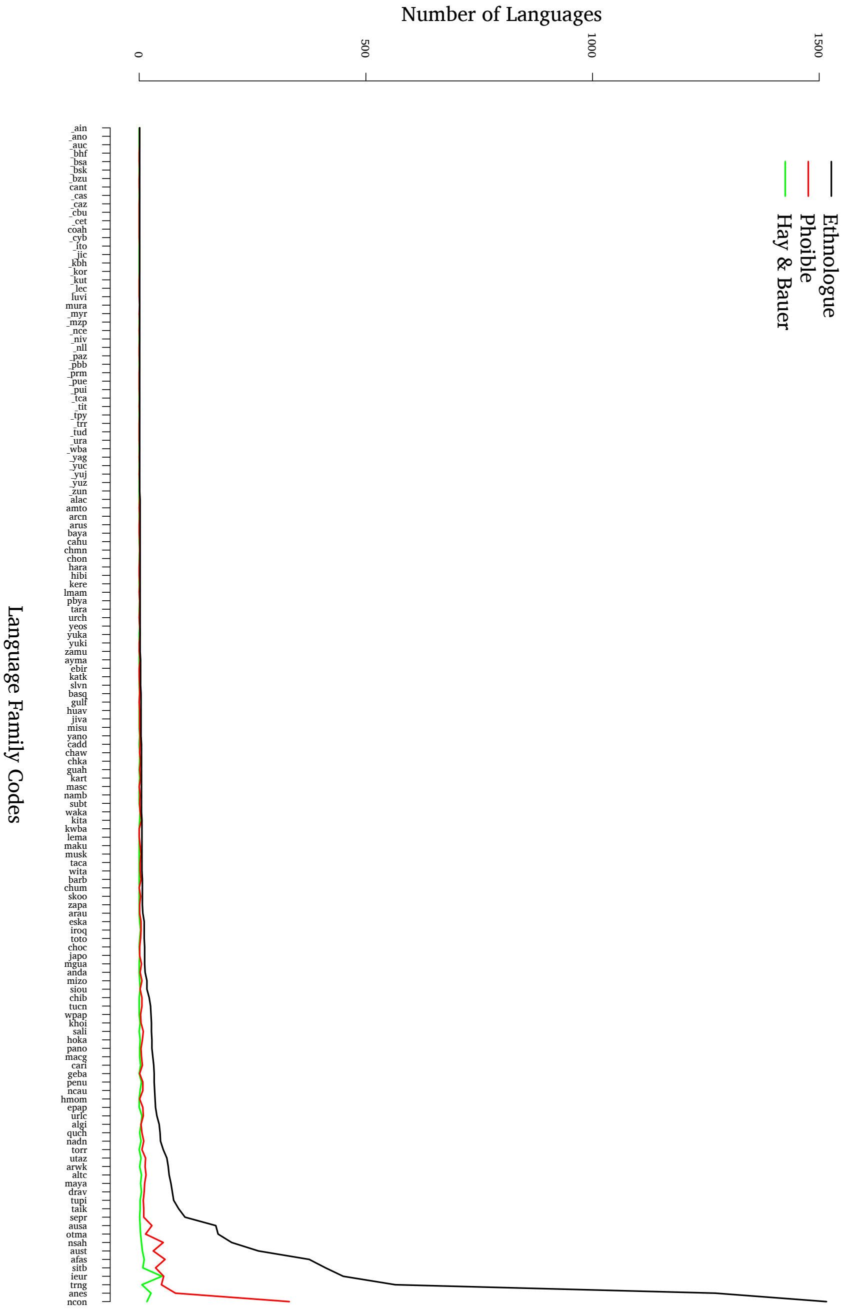


Figure 7.6: Languages per language family in Ethnologue, PHOIBLE and Hay & Bauer 2007



Population figures for my study are taken from the Ethnologue 16 (Lewis, 2009).[19] The measurement of the number of speakers varies over several orders of magnitude (from 1 speaker to 840,000,000) and the use of raw population figures would contain several extreme outliers (e.g. Mandarin, Hindi, Spanish, etc.). I decided to log-transform both the independent (population) and dependent (phoneme counts) variables because it makes the residuals (the error terms) more closely approximate a normal distribution. In linear mixed models it is fine if both the independent and dependent variables are skewed, as is the case with both speaker population and phoneme inventory counts. What is important in linear mixed models (and also in simple linear regression) is that the residuals be normally distributed. A nice side effect of log-transforming both variables is that it becomes easy to interpret the slope, which becomes simply % of change (for example if the slope is 0.5, then there is a 0.5% change in y for every 1% change in x).

I first tried to reproduce Hay & Bauer's results using the PHOIBLE data set with their statistical methods. The results were similar. This was not unexpected since Hay & Bauer also retested their findings with Pericliev's data set (a subset of the UPSID_{451} inventories with Ethnologue population figures). The correlation that they find is "highly significant (Spearman's rho = .21, $p < 0.0001$)", thus providing "strong evidence that the observed correlation is not an artifact of our sampling procedure" (Hay and Bauer, 2007, 397). In fact, I believe it is their method that produces the positive correlation. Spearman rank coefficients for my analysis of the PHOIBLE data ranged from 0.22 to 0.32 with statistically significant correlations between speaker population and full phoneme inventories (Figure 7.7), total consonants (Figure 7.8) and total vowels (Figure 7.9). The correlations are also statistically significant ($p < 0.0001$) for: obstruents (Spearman's rho = 0.2903), sonorants (0.1722), monophthongs (0.234) and non-monophthong vowel qualities (0.2658). As in Hay & Bauer's study, sonorants show the weakest effect in the PHOIBLE data set.

[19] See Section 7.5 for remarks on modern day population figures.

Figure 7.7: LOWESS scatterplot of languages plotted by log(population) and phoneme inventory size

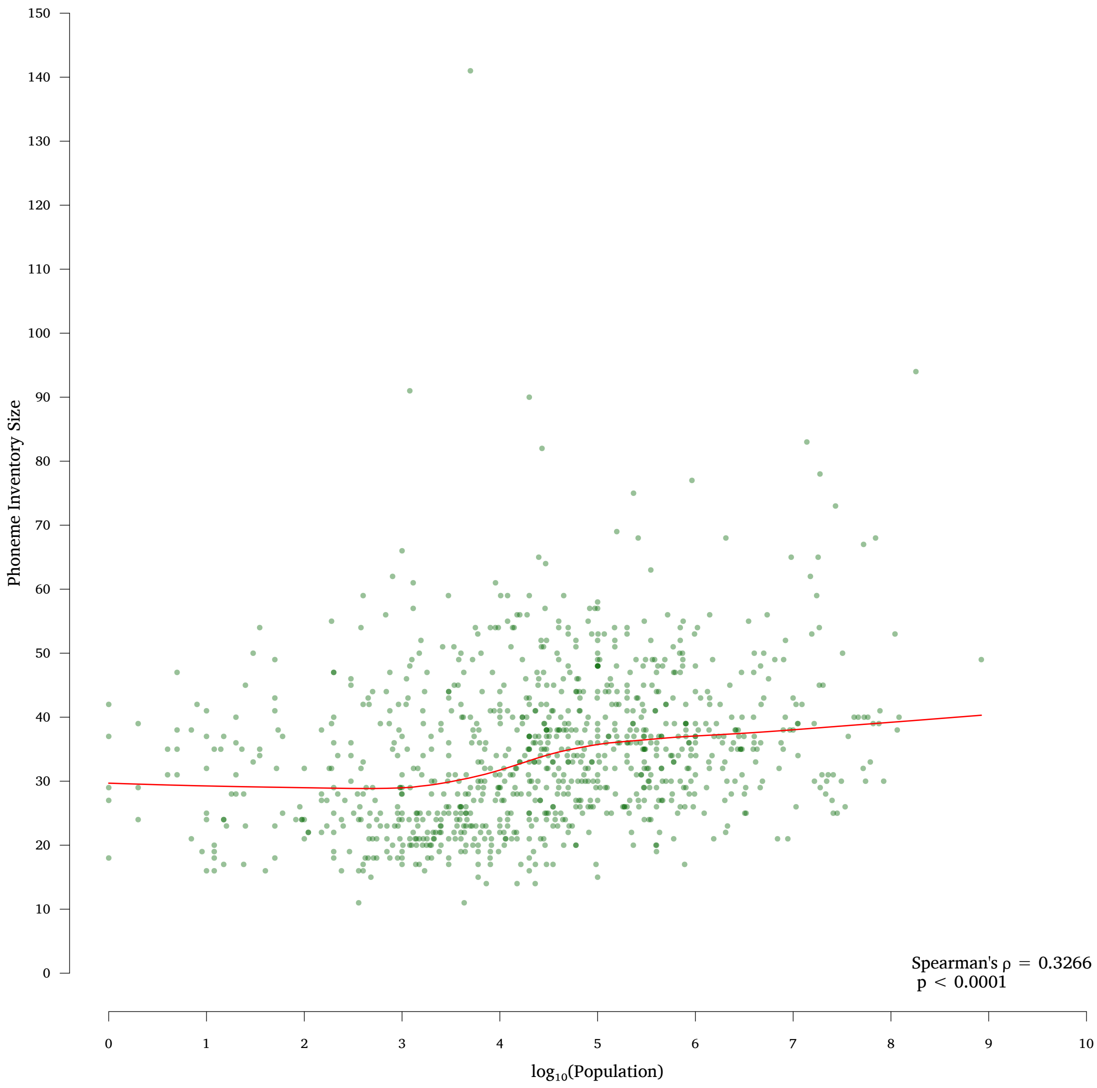


Figure 7.8: LOWESS scatterplot of languages plotted by log(population) and consonant inventory size

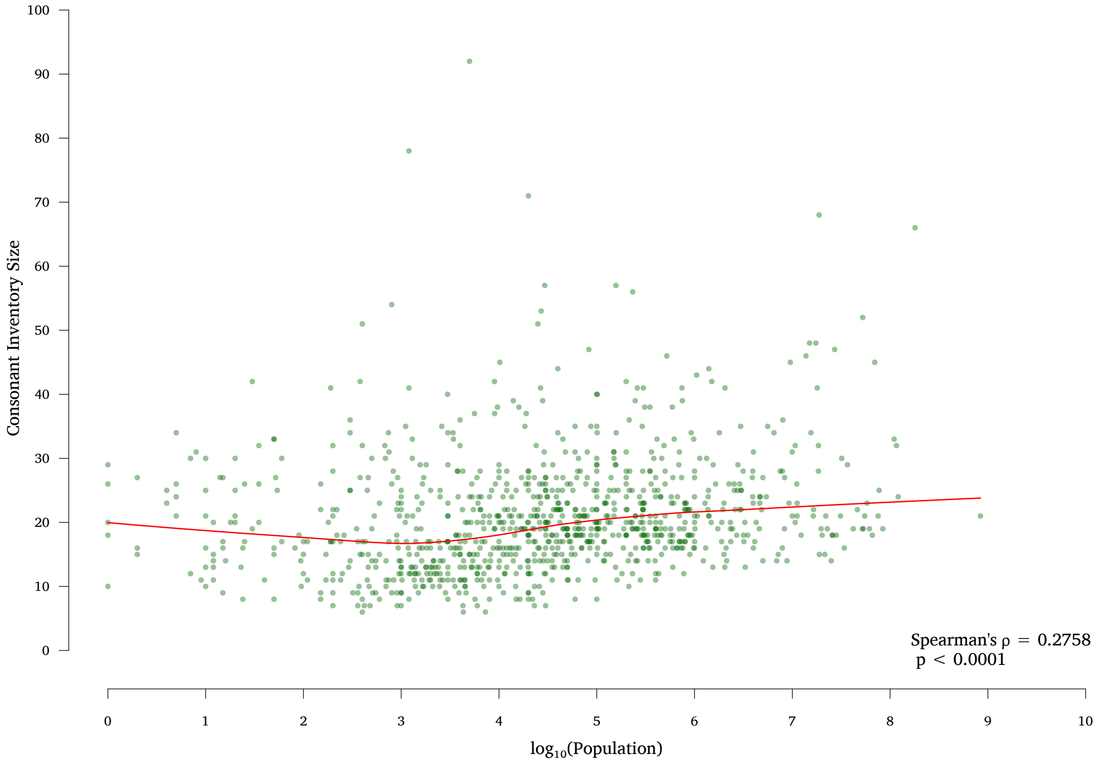
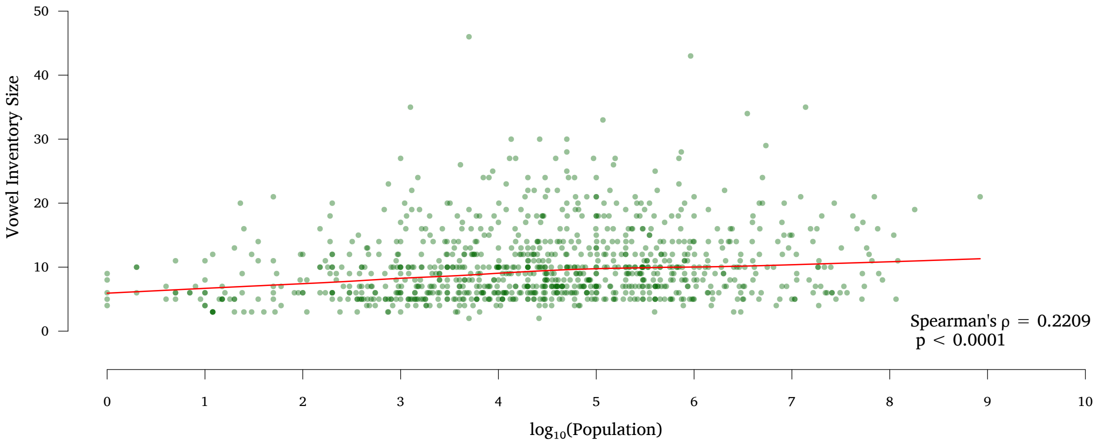


Figure 7.9: LOWESS scatterplot of languages plotted by log(population) and vowel inventory size



Hay & Bauer's use of a simple LOWESS fit and Spearman's rho, however, is not the most appropriate method for phoneme inventory data. A major problem with Hay & Bauer's study has to do with the independence of observations; data points within a language family are more likely to have similar inventories due to shared descent than data points drawn from different families. This is known as a data nesting problem. The phoneme inventory data are hierarchically nested, i.e. languages are nested within genera and genera are nested within a language stock. Additionally, it is more difficult to estimate effect size using a LOWESS model because data points are fit to a curve rather than a straight line.

Instead, I use hierarchical linear modeling (HLM), also known as a mixed effects model or a multilevel model (Raudenbush and Bryk, 2002; Gelman and Hill, 2007; Snijders and Bosker, 1999). HLM is appropriate for nested data because it allows predictors at multiple hierarchical levels. It also uses Bayesian estimation techniques that account for unequal group sizes, thus yielding more precise estimates of variance for groups with lots of data points and less precise estimates for sparely populated groups. An assumption of HLM is that the dependent variable is normally distributed. However, neither speaker population or phoneme inventory counts show normal distribution; both are right-skewed.[20] A standard approach to address skewing is to log-transform the dependent variable.

For ease of comparison with the Hay & Bauer study, I create a model in which log(population) is the independent variable (also called a fixed effect predictor in the mixed models literature). As group-level predictors (aka random effects), language stocks were used. A null model, a random intercept model and a random-slope model were each run with total phonemes (the dependent variable) as language-level predictors. In the null model, no relationship is assumed and each group is modeled by a different horizontal line. If there is no relationship between population size and phoneme inventory size, the null model is expected to be the best fit (where genealogical information is a decent predictor of inventory size, but adding population information does not add any predictive power). In the random-intercept model, a single slope is fit for all groups. If there is a real, cross-linguistic relationship, then the random-intercept model ought to be the best fit. Thus the

[20] For example, see Figure 5.4 on page 221 which shows a histogram of phoneme inventory sizes in PHOIBLE.

intercept ought to account for the language family differences and the effect of population (the slope) ought to be more or less the same for all language families. Also, the slope ought to be non-zero, otherwise we are back to the null model. And in the random-slope model, both the slope and intercept are allowed to vary across groups. If the random-slope random-intercept model is the best fit, then there is either a relationship that is very complex or other factors at play, or there is no relationship and the random slopes are modeling the noise in the data, which is known as overfitting.

For my method, linear mixed models were fit using the *lmer* function in the *lme4* package of R (Bates et al., 2011). Parameter estimates and deviance measures of the three models predicting total number of phonemes are given in Table 7.1.

Table 7.1: Parameter estimates and deviance measures

	Null Model:			**Random Intercept Model:**			**Random Slope Model:**		
	lmer(pho 1+(1|fam))			lmer(pho logPop+(1|fam))			lmer(pho logPop+(1+logPop|fam))		
Parameter	**Parameter**	**S.E.**	**t-value**	**Parameter**	**S.E.**	**t-value**	**Parameter**	**S.E.**	**t-value**
Fixed Effects									
Intercept	30.791	1.015	**30.34**	31.26852	1.3708	**22.810**	30.4384	1.6145	**18.853**
log(population)	–	–	–	-0.05913	0.1126	**-0.525**	0.0592	0.1281	**0.462**
Random Effects									
Intercept variance (τ_{00})	68.202	8.258		69.311	8.3253		113.71077	10.66353	
Slope variance (τ_{11})	–	–		–	–		0.13460	0.13460	
Error covariance (τ_{01})	–	–		–	–		-3.11656	–	
REML deviance			**7284**			**7272**			**7268**

Correlation of random effects in the random slope model is quite strong (-0.797). This suggests that allowing slopes to vary across language families is not adding substantive predictive power and therefore is effectively a redundant predictor with respect to random intercept. Correlation of fixed effects in both the random slope and random intercept models is also quite strong (-0.779 and -0.668, respectively). This suggests that any slope (whether fixed or random) does not add substantive predictive power over and above varying intercepts by group. Therefore, the best fit is the null model. The null model is the statistical model where I assume that population does not have any effect, so I leave it out and use language families to do the prediction.

In testing significance of the models, Baayen et al. (2008) note that the t-distribution for very large numbers of observations converges on a normal distribution. By looking for t-values greater than 1.96, two-tailed significance for fixed effects ($p < 0.05$) can be informally assessed. Using this metric, the varying intercept across families is a highly significant predictor in all three models. However, log(population) as a fixed effect is not. I obtained more precise p-values for the fixed effect by using the *pvals.fnc* function from the R package *LanguageR* (Baayen, 2010). By using a Markov chain Monte Carlo method, the *pvals.fnc* samples from the posterior distribution of fixed-effect parameters. The results of the simulation confirm the assessment based on t-values. Namely, the varying intercept by language family is highly significant ($p < 0.0001$) and log(population) is not significant ($p = 0.60$) as a fixed effect predictor. Figure 7.10 is a plot of the average population size of languages within a language family by the average phoneme inventory size of those languages. Each language family is plotted with its 4-letter Multitree language family code, or when it represents an isolate, its ISO 639-3 code prefixed with an underscore. Figure 7.11 is a plot of log(population) by phoneme inventory count per language. Each language is plotted by its 3-letter ISO 639-3 code. Although there is a correlation in both plots, the R-squared and effect sizes are small: per each increase in population of one order of magnitude, the model predicts an increase of only 0.6 phonemes for family averages or 0.7 phonemes for the individual languages. Figure 7.12 shows a trellis plots (lattice graphic) for the 16 families best represented in PHOIBLE. The lattice graphic confirms the results of the *lmer* function and clearly shows that there is no consistent relationship within language families (some group-level trends are correlated increasing, some are correlated decreasing, and some are not correlated at all).

Figure 7.10: Language families plotted by the average population of their languages (log-transformed) by the average phoneme inventory size of their languages

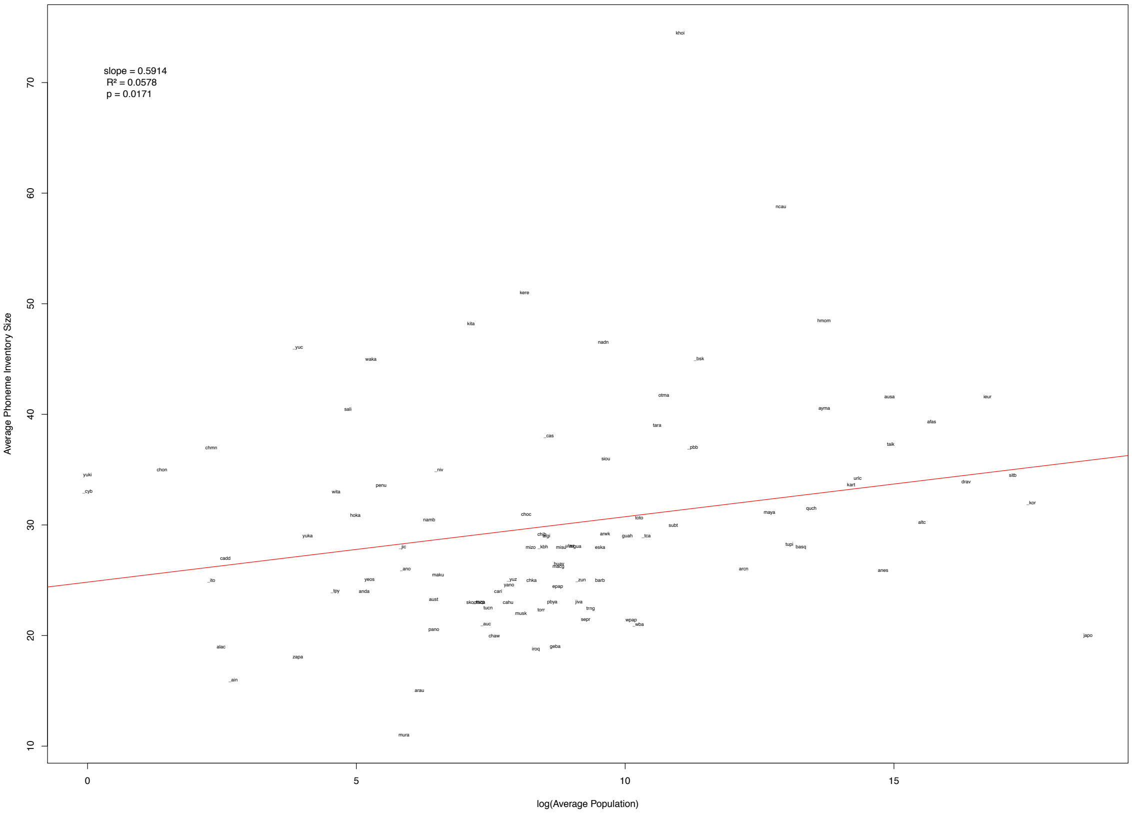


Figure 7.11: Languages plotted by the log(population) of speakers by phoneme inventory size

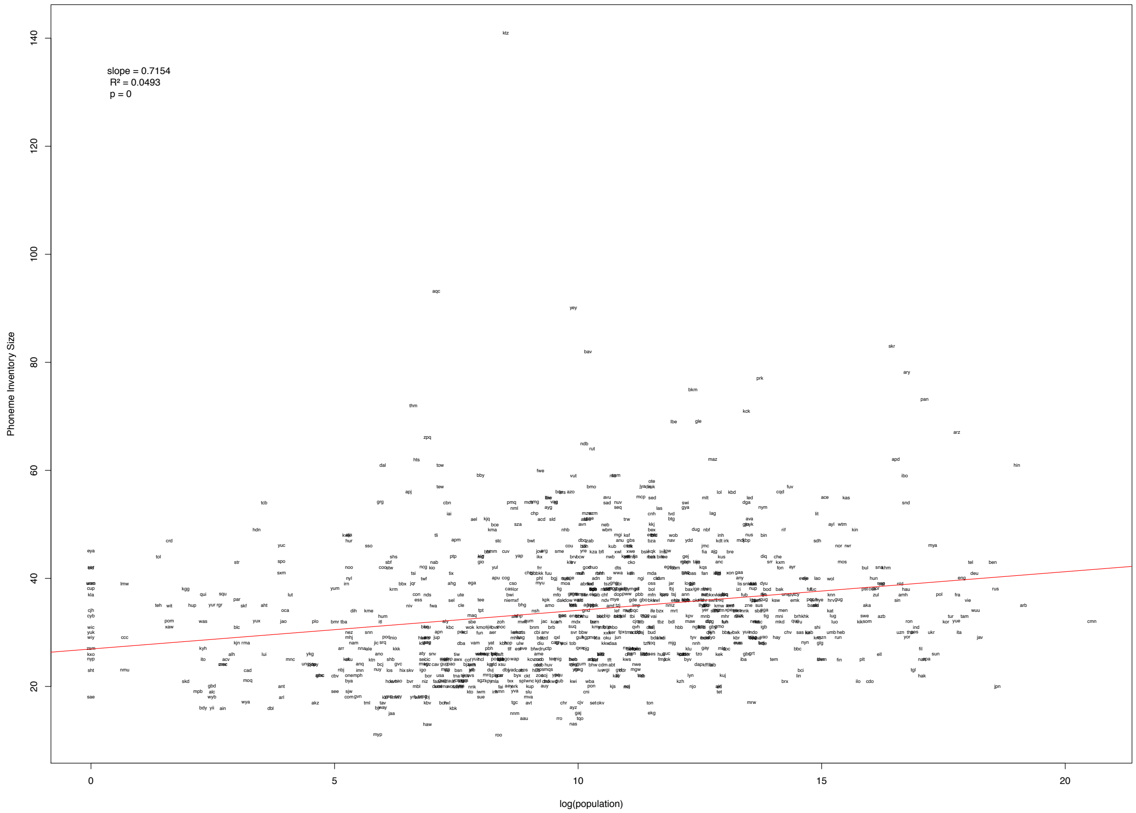
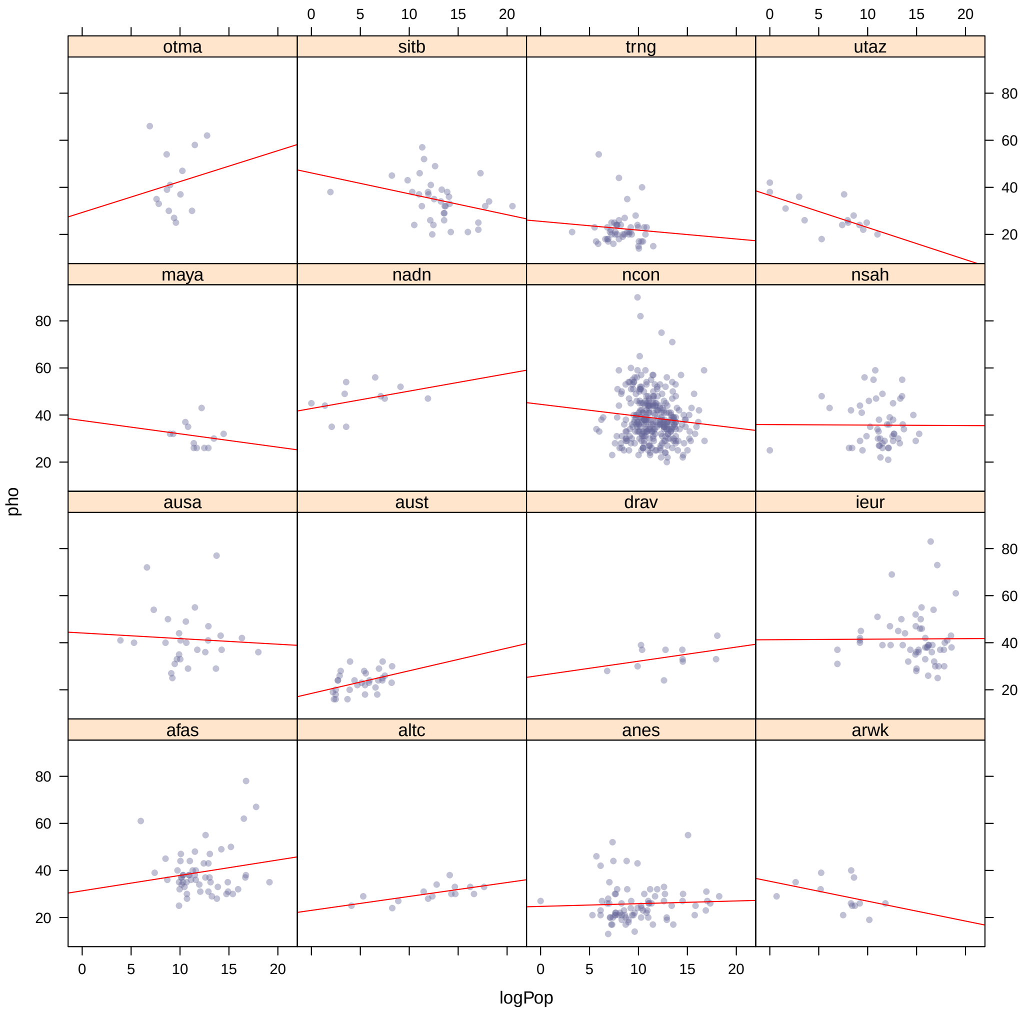


Figure 7.12: Trellis plot of family-level fitted lines from the mixed model predicting total phonemes for the 16 largest families in PHOIBLE



In summary, the results of my analysis contradict the findings reported in Hay and Bauer 2007 and show no correlation between speaker population and phonemic inventory size if language family is accounted for by using HLM to address the nested data problem.

7.5 Discussion

Hay & Bauer were meticulous in their analysis and conservative in their interpretations. However, their sample was too small and too biased to yield reliable results because it limited their choice of statistical method. Their sample of 216 languages includes coverage for 46 language families. Although eight families were represented by six or more data points (6, 6, 7, 8, 11, 17, 26, 50), the majority of families (38) included five or fewer languages and most of these contain only one or two. These radically unequal group sizes are problematic for statistical techniques like ANOVAs because they violate the assumption of homogeneity of variance (Stevens, 2009, chap. 6). The data are also not amenable for mixed models because many groups contain only one data point.

Data points within a given language family are more likely to have similar inventories than data points drawn from different language families. I believe the correlation found in Hay & Bauer is due to the LOWESS fit and Spearman's rho, which are not the most appropriate choices for these assessing data because of the assumption of independence of observations. Additionally, although reducing each language family to one data point may be in general a good method for dealing with unequal group sizes, it may not be an ideal method for dealing with skewed samples, such as their sample, in which some families are well-represented and others are absent from the data sample. Also, unlike Hay & Bauer's study, my method does not require that a threshold be met for lumping languages into language families or into one "other" group to preserve the needed degrees of freedom in the statistical model. In addition, the PHOIBLE data set is less skewed, much larger, and a more representative sample of the world's languages than what has been used in other studies of population size and phoneme inventory size.

For studies using typological data sets and quantitative methods, there are several methodological considerations. One is the data set. Many recent studies using statistical methods rely solely on data from WALS (Haspelmath et al., 2008). Although WALS is a great resource, undertaking quantitative methods using the chapters related to phonological systems is problematic. The chapters on consonant inventories (Maddieson, 2008a, 2011a), vowel quality inventories (Maddieson, 2008c, 2011d) and tone (Maddieson, 2008b, 2011b) provide broad groupings (e.g. small, average, large) of inventory sizes and not actual phoneme counts.[21] For example, Atkinson (2011) combines the features from these three chapters to obtain an estimate of the size of phoneme inventories. Not only are these three categories erroneously weighted equally in Atkinson's study (the number of consonants in languages typically is much higher than vowels or tone), the WALS vowel counts only include the number of vowel qualities; thus ignoring other ways in which languages phonemically distinguish vowels (e.g. vowel length, nasalization, diphthongs). Alternatively, the UPSID_{451} data is publicly available and was used in Pericliev 2004. However, UPSID_{451} does not contain tone in its inventories. Like differing analyses of non-quality vowel distinctions, the description of tone is subject to differences in opinion by language documenters and their descriptions of vowel (or tone) systems may differ widely (cf. Maddieson 2011d,b). To address non-quality vowel distinctions, Hay & Bauer go as far as to divide monophthongs into two categories, basic and all; the former display a greater consistency across analyses. On the other hand, the authors make no reference to tone in their study. I've tried to address these issues in the construction of the PHOIBLE data set by providing non-quality vowel distinctions and tone when they are described in the original resources from which inventories were extracted. These phonemes can also be located in the data set and removed. A last criticism that has to do with data samples is the reproducibility of results. Although it is current practice to list languages by name in linguistic studies, for ease of reproducibility it would be better to also list language names with their ISO 639-3 identifiers. For example, in trying to reproduce Hay & Bauer's study with their sample, one is faced with language names belonging to macrolanguages or sub-genera (e.g. Berber, Malagasy, Malay, etc.) and it is therefore not clear to which particular language the figures (phoneme inventory and population size) belong.

[21] These values are based on data that were collected, so that an average consonant inventory, for example, is categorized as inventories that are ± three consonants above and below the modal consonant inventory size in the sample (22).

Another methodological consideration involves sampling typological data sets to characterize the distribution of linguistic phenomena.[22] Hay & Bauer suggest that an ideal approach might be to randomly sample phoneme inventory counts and population figures from an exhaustive language index, such as the Ethnologue. For statistical evaluation, a random sample is indeed ideal. However, in the case of phoneme inventories, it is not possible to draw a random sample from the entire population of languages. Not all languages are adequately documented and many are not documented at all. A language is also not a clearly demarcated object. Furthermore, true random sampling is not possible because the current state of the world's languages represents actual languages and not necessarily all possible variations of human languages (cf. Cysouw 2005). The studies mentioned in Section 7.2 all drew from different language samples. Trudgill's hypotheses are based on convenience samples, i.e. data from languages that he presumably collected without regard for genealogical or areal stratification. In their rebuttals of Trudgill 2004a, Bakker (2004) uses a convenience sample and Pericliev (2004) uses data from UPSID_{451}, which was constructed with a quota sample aimed at creating a genealogically diverse and representative sample of the world's languages (Maddieson, 1984; Maddieson and Precoda, 1990). Hay and Bauer (2007) drew languages from Bauer 2007, which is also a convenience sample. Wichmann et al. (2008) use a 2140 language sample from WALS (note the problems with the phonological data in WALS, mentioned above).[23] Each of these samples can be criticized in some regard. Convenience samples are chosen with no restrictions on inclusion from data that are readily available. They are typical of exploratory investigations that do not take genealogical or areal stratification into account, which leads to bias. Hay & Bauer's data set has the problem of overrepresenting certain language families and underrepresenting others. Pericliev's use of the UPSID_{451} data set is another example of a methodological challenge of avoiding bias. The UPSID_{451} data aims for a genealogically balanced sample by including one language from each small language family. However, UPSID_{451} fails to capture typological diversity within these groups. My study can also be criticized for not

[22] See Section 2.3.2 for a discussion on sampling.

[23] See Hammarström 2009 for a discussion about the genealogical skew of languages in WALS and problems of making sound statistical inferences based on its distribution of typological features.

taking genealogical or areal stratification into account. However, it was my aim to reproduce Hay & Bauer's study and to use as much data as was available to test the correlation between population size and phoneme inventory size. Thus I did not stratify the data, which involves a sampling methodology that attempts to reduce the language family-level bias due to unequal representation at the family-by-family (or region-by-region, etc.) level. Instead, I chose to control for the influence of language family. I did not assume that all languages were independent, but accounted for the fact that genealogically related families are more likely to have similar inventory sizes. By controlling (and not stratifying) for language family, my method allows me to use more data and to look at within-family trends, which are potentially informative.

Yet another methodological consideration is the genealogical classification of languages, which are prone to ongoing scientific debate. Hay & Bauer use the classification from the original grammars from which they took their data. However, if their data sample is reclassified using the Ethnologue's genealogical classification, then the families that meet Hay & Bauer's seven language minimum cut-off for their linear regression model criterion change, i.e. the group containing Altaic, Austronesian, Indo-European, Niger-Congo and Penutian changes to include Afro-Asiatic, Australian, and Sino-Tibetan; and Altaic and Penutian are thrown out, since both would be reduced to only five representative languages, and therefore would not be included as family predictors. Note that even genealogically stratified samples may change drastically depending on the genealogical classification used (Rijkhoff and Bakker, 1998).

The interpretation of results is another methodological consideration to keep in mind. In a recent article that discusses general statistical models, van der Laan and Rose (2010) state, "We know that for large enough sample sizes, every study—including ones in which the null hypothesis of no effect is true—will declare a statistically significant effect." The standard criteria to determine statistical significance seems to be easier to attain as data samples become increasingly larger, if one uses the same test and criteria for significance. This is due to the larger number of observations that allow one to estimate the variance with greater and greater precision. The problem becomes one of the interpretation of significance; standard criteria such as "$p < 0.05$" or "$p < 0.01$" are not always enough depending on the data and the methods used to estimate significance. Therefore, it is important to calculate effect size as part of statistical interpretations. For example, if a statistically significant non-zero correlation exists, how non-zero is it? Discussion of effect size is often absent from studies that claim statistical significance, such as Hay and Bauer (2007). For example, if for each tenfold increase in speaker population there is an increase of 0.3 phonemes – is this finding interesting? The difference between the smallest and largest speaker populations (over 20 orders of magnitude) would be a difference of only six phonemes, which is within the range of variability within each magnitude.

Finally, there is the question of why (roughly) current population figures are applicable to studies on population size and phoneme inventory size. Early human communities were small, likely ranging from a few hundred up to a thousand in exceptional cases. The existence of large speaker populations is a relatively recent phenomenon that only arose in the context of agriculture long after the peopling of most of the world (cf. Mithen 2003). This means that any correlation between population size and phoneme inventory size is an effect that arose only after human settlement of the world was finished and that any correlation is a product of recent population growth. However, the gain or loss of phonemes in a language seems to be a much slower process than the rate of population change. Also, speaker populations can change dramatically for non-biological reasons, e.g. in the case of cultural expansion leading to bilingualism where the next generation grows up speaking a different language than their parents.

7.6 Conclusion

In this chapter, I have discussed the equivocal results of studies regarding the correlation between population size and phoneme inventory size. I have argued against the findings of Hay and Bauer (2007), who use a LOWESS statistical model with significance assessed Spearman's rho on a set of 216 languages and find a positive correlation between population size and phoneme inventory size. My study addresses the shortcomings of Hay and Bauer 2007 by using a much larger data set with wider and deeper genealogical coverage and a hierarchical linear model to control for the genealogical relatedness of languages. I show that there is no correlation between population size and phoneme inventory size, once language family is accounted for. My work may also cast serious doubts on the results of studies that assume a positive correlation between population size and phoneme inventory size. For example, Atkinson (2011) proposes that a single language origin in Africa is supported by an out-of-Africa serial founder effect in which average phoneme inventory size decreases as one moves away from Africa. This analysis crucially depends on a correlation between population size and phoneme inventory size.

Atkinson argues that this correlation is significant with the WALS data[24] and that it is also significant when restricting speaker populations to 5000 or less, roughly in line with modern hunter-gatherers (the assumption being that pre-historic groups would have been about this size).[25] However, when using the UPSID$_{451}$ data with actual segment counts (and compensating for its lack of tone), the correlation between speaker populations (of 5000 or less) and phoneme inventory size is shown to be not significant (p = 0.64, r = 0.04) and only reaches significance when larger populations of over 100k speakers are included (Cysouw et al., 2012). Again, these studies reach different conclusions regarding a correlation between population size and phoneme inventory size.

There is no direct access to evidence regarding population sizes of prehistoric speaker communities, but what we do know is that larger speaker populations are a relatively recent phenomenon (Mithen, 2003). These factors should be taken into consideration with what is known (or can be inferred) about the rate of language language and sound change (e.g. Johnson 1976; Nettle 1999a; Wichmann and Holman 2009). This is not to say that population size may not have some kind of influence on language structure and that correlations should not be investigated; we should ask if it makes sense to use current population figures when testing correlations such as population size versus phoneme inventory size in light of what we know about population growth and language change.

In this chapter, I have also discussed some of the methodological considerations in undertaking studies using statistical methods with phonological typological data and I have

[24] Note that the WALS data is problematic for this type of analysis because it does not provide specific segment inventory counts, instead only bins of average sizes for consonants, vowels and tone, which were erroneously weighted equally in Atkinson's analysis. See criticisms in Cysouw et al. 2012.

[25] In regard to early speaker population sizes, see also Richard Sproat's criticisms of Atkinson 2011 at: `http://www.cslu.ogi.edu/~sproatr/newindex/atkinson.html`.

illustrated how one might use PHOIBLE to investigate claims of correlations between non-linguistic factors and the phonological system. In other work I am investigating the claim that there exists a correlation between climate and the phonological system, e.g. languages spoken in warm climates use relatively more high-sonority sounds than those spoken in cold climates (Munroe et al., 1996; Munroe and Silander, 1999; Fought et al., 2004; Ember and Ember, 2007; Munroe et al., 2009).

Chapter 8
CONCLUSION

8.1 Summary

In this work I intended to answer the question of whether more sophisticated, knowledge-based approaches to data modeling, coupled with a larger cross-linguistic data set, could extend previous typological observations and provide novel ways of querying segment inventories to undertake phonological typology. Broadly, this work is concerned with:

- creating a cross-linguistic data set to undertake phonological typology
- modeling this data set in ways that facilitate testing typological observations by aligning the data models to questions that typologists wish to ask
- instantiating technological infrastructure that is conducive to data sharing, extensibility and reproducibility of results
- using the data set and data models in this work to validate and extend previous typological observations

In Chapter 2 in Section 2.3, I raise the linguistic and technological challenges involved in creating a useful cross-linguistic typological data set. Issues of what constitutes adequate descriptive categories for linguistic phenomena (Sherman and Vihman, 1972, 173) and whether data stemming from many different linguists' analyses can be typologized (cf. Newmeyer 2007; Haspelmath 2010) are discussed in Section 2.3.1. An overview of the issues of statistical sampling is given in Section 2.3.2. The challenges involved in doing typology with segment inventories are raised in Section 2.3.3 and standardization of linguistic segments and unique language name identifiers are discussed in Sections 2.3.4 & 2.3.5. Lastly in Section 2.3.6, I bring up the thorny and yet-to-be resolved issue of documenting metadata and data provenance.[1]

In Chapter 3, I introduce several data models and explain the approaches that I've taken in encoding the PHOIBLE data set in these data models. In general it is important that data are easily interpretable (Bird and Simons, 2003; Abney and Bird, 2010); a simple machine readable storage model is a practical way to make data available to a large audience. Thus, flat file tables are one format in which the PHOIBLE data set is made available. The tables are convenient as an input format for statistical packages and programming scripts, as I show in Chapters 5 & 7, in which I investigate various properties of segment inventories and a reported correlation between segment inventory size and population size. In Chapter 3, I discuss PHOIBLE's relational database model and its RDF graph model. I also describe aspects of knowledge representation and I show how constructed an RDF/OWL "knowledge base" that allows researchers to manipulate aspects of the PHOIBLE data set without changing its underlying data. The functionality of this knowledge base is illustrated in Chapter 6, in which I use it to query across segment inventories at the feature level to investigate proposed descriptive universals of phonological systems.

In Chapter 4, I provide an overview of PHOIBLE and I describe the extract, transform and load processes that I used to bring the segment inventories from the Stanford Phonology Archive (SPA; Crothers et al. 1979), the UCLA Phonological Segment Inventory Database (UPSID; Maddieson 1984, Maddieson and Precoda 1990) and the Systèmes alphabátiques des langues africaines (AA; Hartell 1993, Chanard 2006) together with hundreds of inventories extracted from grammars and phonological descriptions for this work into one large interoperable data set. Lastly, I discuss the genealogical coverage of PHOIBLE.

In Chapter 5, I revisit some of the typological facts of segment inventories as postulated in other work with previous segment inventory databases. I evaluate these claims against the inventories currently in PHOIBLE and I implement a statistical sampling technique to account for effects of genealogical skewing. I also investigate segment type frequencies cross-linguistically and the balance between consonants and vowels in inventories. Lastly, I

[1] However, see Section 8.4 below.

revisit Crothers's (1978) observation that the vowels /i, a, u/ occur in most languages and I show using multi-dimensional scaling how vowel systems tend to expand after cardinal vowels.

In Chapter 6, I show that distinctive feature sets have poor typological coverage when compared to the numerous segment types found in the combined PHOIBLE segment inventories. I then describe how I expanded the Hayes 2009 feature set to address its typological representation deficiencies and I implement a computational approach to assign distinctive feature vectors to previous undefined segment types. I use the PHOIBLE RDF/OWL knowledge base of segment inventories and distinctive features to investigate the descriptive universals put forth by Hyman (2008) and I show that although nearly all of these universals still hold on the broader sample of languages in PHOIBLE, the proposed universal "all languages have coronals" does not (cf. Blevins 2009).

Finally in Chapter 7, I present a case study that uses the PHOIBLE data set to revisit the claim that there exists a correlation between population size and phoneme inventory size, as speculated in Haudricourt 1961 and Trudgill 1997, 2002, and empirically tested and reported in Hay and Bauer 2007. Using a much larger sample than Hay & Bauer's, which affords a more nuanced statistical approach using a hierarchical mixed-effects linear model that accounts for the non-independence of data points, I show that no correlation between population size and phoneme inventory size exists when genealogical factors are taken into account. The case study shows how one might use PHOIBLE to investigate one of the many reported correlations between linguistic and non-linguistic factors.[2]

In this final chapter I discuss the contributions of my work to the field in Section 8.2. In Section 8.3, I address the issues of linking lexicons to segment inventories, and in Section 8.4, I describe avenues for future research.

8.2 Contributions

This work contributes a large phonological typology data set to the field and makes these data openly available in different formats for researchers to use. These data are far from

[2] See also Section 8.4.4, below.

perfect, but they provide a new and richer perspective on phonological systems of the world's languages. Coupled with additional linguistic and non-linguistic data, this data set provides a rich resource for undertaking phonological typology and it contains data pertinent to statistical sampling. My aim has been to model these data in formats that are extensible and interoperable, so that PHOIBLE can continue to grow and be integrated with new sources of data, such as lexicons, corpora, and non-linguistic data points like climate data and socio-economic variables like gross domestic product (GDP), etc.

In this work I have raised and addressed several challenges pertinent to linguistics and the technological implementation of linguistic data, including:

- encoding linguistic segments in Unicode IPA for standardization and segment interoperability, including:
  - defining the full set of IPA characters in Unicode
  - defining diacritic ordering of IPA segments
  - raising awareness of issues in Unicode and IPA (e.g. keyboard <g> versus Unicode voiced velar stop <ɡ>) and making tests to catch such errors
  - parsing and implementing Unicode normalization forms for multi-character sequences to align their logical and visual orders
- providing the Hayes 2009 distinctive feature set in Unicode and extending its incomplete IPA coverage as "Hayes Prime" that maps all unique Unicode characters to a vector of distinctive features; thus providing the basis for all segments types in PHOIBLE to receive a feature vector
- devising methods to automatically assign feature vectors to all segment types in inventories in PHOIBLE to achieve full typological coverage
- modeling PHOIBLE's data set in data structures that facilitate testing typological observations

- attaining structural interoperability of segments, segment inventories and distinctive features by modeling them in the RDF and OWL data models

- providing a feature geometry based on Hayes Prime and encoded in OWL

I also brought up:

- issues of data provenance, particularly in the area of data reuse and reinterpretation

- issues of genealogical sampling

I have developed technological architecture that allows users to:

- query segment inventories at the level of segments and distinctive features

- query segment inventories by various linguistic and non-linguistic variables, e.g. segment class (i.e. consonant, vowel, tone, diphthong, etc.), language family or genus, geographical region, country or geo-coordinate, population, etc.

- access the data in various formats, including flat file tables, a relational database and an RDF graph model

- add information to the data set by using Linked Data[3]

- manipulate the "surface" data set without changing its underlying contents by using OWL logic constructions and constraints on the RDF segments and features graphs

- test for correlations between linguistic and non-linguistic factors

- extract sample sets that adhere to genealogical and/or geographical constraints

[3] See Section 8.4.6, below.

Using the technological infrastructure and the data instantiated with it, including the segment inventories from three databases and the hundreds of additional inventories extracted from source documents, I revisit some of the typological facts put forth about segments and segment inventories in the world's languages. I show that:

- in general segment frequencies and the mean size of inventories remain close to the figures put forth in Maddieson 1984 and subsequent work using UPSID
- after taking into account genealogical skewing, segment types frequently found in most languages tend not to be far off from their frequency in the combined PHOIBLE data set, which is not genealogically balanced
- as segment inventories have been added to PHOIBLE, the number of new distinct segment types continues to increase at a rate that is not asymptotic
- there is a weak correlation between the number of consonants and vowels in segment inventories in PHOIBLE
- there is no correlation between the number of consonants or the number of vowels and tones in languages
- Crothers's (1978) observation that vowel systems typically have /i, a, u/ holds and I show with multidimensional scaling that vowel systems tend to expand beyond cardinal vowels by first adding a lengthened series of vowels, then a series of nasalized vowels, and then diphthongs

By building a system that allows researchers to query segment inventories at the level of distinctive features, I show that:

- distinctive feature systems have poor typological representation of segment inventories
- distinctive feature vectors can be automatically generated for some segment types, however, some "complex" segment types that are undefined by a distinctive feature

set must be assigned by hand because feature assignment can be ambiguous, e.g. the features of [p] and [f] do not map straightforwardly to the feature set of [pf]

- with one exception, descriptive universals in phonological systems as stipulated in Hyman 2008 continue to hold on a much larger and broader data set than UPSID$_{451}$

Lastly, I have fulfilled my aims to:

- create a cross-linguistic data set to undertake phonological typology
- provide novel access to phonological inventories at the feature level
- provide researchers with a tool to undertake phonological typology in ways and with data that were not previously available
- create a typological tool that is extensible and that can interoperate with other sources of linguistic and non-linguistic information
- publish data in open formats
- create avenues for future research

Next, I will describe the next step in integrating lexical information with segment inventory data, before I describe several paths for future research.

8.3 Where are the lexicons?

When PHOIBLE was envisioned, our plan included linking segment inventories to lexicons with associated audio recordings.[4] Due to the many challenges of creating an interoperable data set for segment inventories, as discussed and addressed in Chapters 2, 3 & 4, our

[4] Adding sound files is a long-term goal that would allow us, along with various software, to do forced alignment of annotations and to extract formant data from audio recordings. At this time, however, software such as the *Forced Alignment and Vowel Extraction Program Suite* (Rosenfelder et al., 2011) is English specific. Indeed most such software is still limited to majority languages. By connecting inventories, lexicons, audio recordings and their formant information, one could search for all recordings that contain a segment (or feature) and compare these “same” sounds cross-linguistically.

initial idea to combine inventories, lexicons and audio files proved too ambitious for this work alone. Nevertheless, I have been developing infrastructure to connect lexicons with segment inventories and distinctive features.

The lexicon data type poses similar challenges in creating interoperable data as did the segment inventories. For example, authors of lexicons use a variety of writing systems that range from their own idiosyncratic transcriptions to already well-established practical or longstanding orthographies. Just as segments in inventories in this work were mapped to IPA, which acts as an interlingual pivot to attain interoperability across the transcriptions systems that encode segment inventories differently, graphemes in each orthographic system must also be identified and standardized if interoperability with segment inventories is to be achieved. In most cases this is more than simply mapping a grapheme to an IPA segment because graphemes must first be identified in context (e.g. is the sequence <sh> one sound or two?) and strings must be parsed, which may include taking orthographic rules into account (e.g. <n> between vowels is /n/ and <n> after a vowel but before a consonant is a nasalized vowel /ṽ/). In this section I describe the challenges of parsing orthographic systems and how we resolve the link between orthographies and segment inventories with what Michael Cysouw and I call *orthography profiles*.

I will start by defining the possible input. By **lexicon** I mean a work about words or groups of related words that might be encoded in a wordlist, dictionary or bilingual dictionary. A wordlist is minimally a list of words in a language. For example, the Swadesh wordlist is a list of 100 words in English, the **concepts** of which are said to be common across languages, including such things as: MAN, WOMAN, SUN, MOON, STAR, etc. (Swadesh, 1971). A wordlist becomes more useful when it includes mappings between concepts and word **counterparts**, i.e. translational equivalents (cf. Haspelmath and Tadmor 2009; Poornima and Good 2010), in one or more target languages. The term counterpart differs from the notions of *definition* or *translation* because the counterpart's core function is to refer to language-independent concepts (Poornima and Good, 2010). For example, the word "man" in English is ambiguous between "male" and "human", but the concepts MAN and HUMAN are represented by the German counterparts "Mann" and "Mensch" (each of which has various other meanings in German). There are many works that use concept wordlists, whether the

Swadesh wordlist or another comparative vocabulary wordlist, to gather counterparts from various languages and to align them on concepts to undertake cross-linguistic comparison for tasks like language comparison and genealogical classification.[5]

A **dictionary** is a work that lists words of a language and defines those words using another language. For example Banfield (1914), in his *Dictionary of the Nupe Language*, defines Nupe words using English. He also provides additional information about the Nupe entries, which is common practice for lexicographers, e.g. part of speech information, multiple meanings and examples. Instead of providing definitions, a **bilingual dictionary** (or translation dictionary) translates words and phrases from one language to another, where the nuisances of pragmatics may be employed, e.g. English "cool" can be translated into German as "kühl", "geil", "krass", "cool", or a host of other words, depending on the context.

In my experience, each lexicon must be individually parsed so that its structure is identified and its contents can be extracted.[6] To extract data for analysis, a lexicon-by-lexicon approach is required before any additional linking of lexical data to segment inventory data can be undertaken. As with extracting segment inventories from phonological descriptions, each lexicon is idiosyncratic in its orthography and thus requires lexicon-specific approaches to mapping orthography to phonology.

There are a variety of formats (e.g. PDF, Word, Excel, Access, MDF for Toolbox, OpenOffice) and a variety of standards for encoding lexicons, e.g. Lexicon Interchange FormaT (LIFT),[7] Lexical Markup Framework (LMF),[8] Text Encoding Initiative (TEI)[9] and lemon.[10] Each format and each encoding standard presents its own set of challenges for

[5] One example with thousands of concepts and over a dozen languages is the Dogon comparative lexicon (Heath et al., 2012). See: `http://dogonlanguages.org/`.

[6] If a lexicon exists only in printed form, it must first be digitized before any parsing can be undertaken. If a lexicon is already in a digital format, there may still be the problem of extracting textual content losslessly, e.g. extraction of the original text from a variety of PDF formats, as encoded by different software vendors, can be notoriously difficult.

[7] `http://code.google.com/p/lift-standard/`

[8] `http://www.lexicalmarkupframework.org/`

[9] `http://www.tei-c.org/`

[10] `http://monnetproject.deri.ie/lemonsource/`

extracting and encoding data. A large-scale project that typifies the process of creating an interoperable model for lexicons is the Lexicon Enhancement via the GOLD Ontology (LEGO) project.[11] The aim of LEGO is to create a "datanet" of interoperable lexicons by tackling the issues of extracting lexical data from various formats and encoding those lexicons into LIFT, an XML format for storing lexical information for dictionary creation. Additionally, the morphosyntactic information with regard to lexical items (e.g. part of speech information) in the various wordlists are mapped to the General Ontology for Linguistic Description (GOLD), which allows searching across the numerous wordlists at the morphosyntactic level (e.g. "give me all nouns that have the morphosyntactic feature gender") to attain semantic interoperability. The goal is to develop enhanced search functionality across once disparate lexicons and to demonstrate the value of abiding by technological standards.

The LEGO vision is admirable and linguists welcome the ability to search across lexicons via an ontology that defines morphosyntactic categories (ILIT, 2012). However, the lexicons were originally encoded in heterogeneous transcription systems or practical orthographies, so searching across the lexicons at the phonological level is not (entirely) possible.[12] Each lexicon faces the same challenges identifying segments and mapping them to an interlingual pivot, as does each description of a phonological inventory for PHOIBLE. For orthographies, identifying graphemes can be even more challenging than identifying phones and phonemes in phonetic transcription because although transcriptions may not adhere strictly to IPA, they tend to have straightforward mappings between sounds and symbols. On the other hand, orthographies can introduce orthographic rules, which add an additional challenges in identifying graphemes in words, as mentioned above. Thus for resources not in IPA or IPA-like transcriptions, graphemes must first be manually identified, whether they are encoded as singletons or multi-character sequences. The identification of graphemes and the formulation of orthographic rules are used to create an **orthography profile**. An orthography profile is a description of the units and rules that are needed to adequately

[11] `http://lego.linguistlist.org/`

[12] Some phonemic/phonetic/graphemic segments may indeed be cross-linguistically queryable, e.g. <p> is more likely to reflect the same element across various lexicons than, say, <y>.

model a writing system for a language variety as described in a particular document. An orthography profile states the Unicode code points, characters, graphemes and orthographic rules used to write a language. Note the different levels of technological and linguistic elements that interact in Table 8.1 for the hypothetical lexical form <tsʰṍ̰shi>.

Table 8.1: Different levels of technological and linguistic elements

1. code points	(10)	t	s	ʰ	o	˜	´	˷	s	h	i
2. characters	(6)	t	sʰ		ṍ̰				s	h	i
3. graphemes	(4)	tsʰ			ṍ̰				sh		i

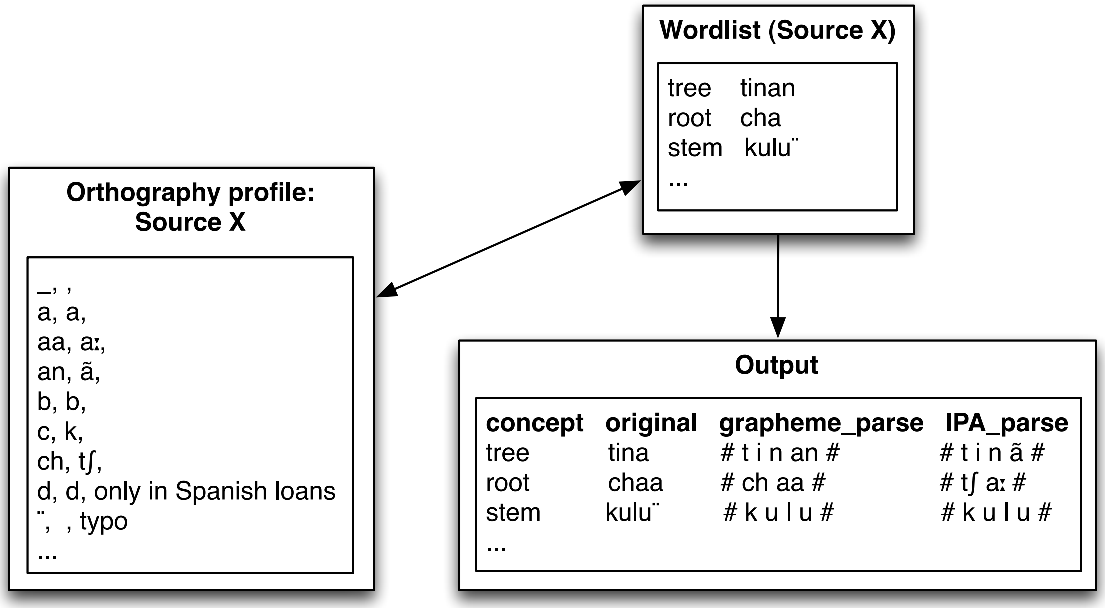
By splitting on Unicode character points, the string <tsʰṍ̰shi> is tokenized into ten characters. Next, in the second row in Table 8.1, the code points have been logically normalized and visually organized into characters in the Unicode Standard.[13] Lastly, in the third row of Table 8.1, an orthography profile is needed to parse sequences of Unicode grapheme clusters into language-specific graphemes as specified in the target language's writing system. For example, our hypothetical orthography profile would specify that the sequence of characters <t> and <sʰ> form a single grapheme <tsʰ>, and that <s> and <h> form <sh>.

Once the graphemes in a particular document are identified and specified in an orthography profile, parsing lexicons is straightforward. An example is given in Figure 8.1. The lexical data is read in, graphemes in the orthography profile are loaded into a *trie* data structure, and then each word is parsed into graphemes based on a greedy match. The output is a white-space grapheme delimited format that uses "#" for word boundaries and between words in multi-word phrases. For example, <tsʰṍ̰shi> would be graphemically parsed and

[13] Note that the character <ʰ> resides in the "Spacing Modifier Letters" Unicode block. Spacing Modifier Letters are intended to form a unit with (typically) their preceding letter, which they modify. These characters differ from diacritic markers because they are treated as free-standing, spacing characters. For example, when parsing strings that contain characters from the Spacing Modifier Letters block and using a Python regular expression parser to match Unicode graphemes ("\X"), <ʰ> (and other Spacing Modifier Letters) are not parsed as graphemes (like <ṍ̰>), but as stand-alone characters (e.g. <s>, <ʰ>). In this example in Table 8.1, I have combined <s> and <ʰ> because Unicode intends them to form as a unit.

output as <# ts^{h} ṍ̰ sh i #>. These graphemes can be converted to phonemes by simply using the second column of the orthography profile (a comma-separated values file) as a look up table. The third column in the orthography profile is used for notes.

Figure 8.1: Orthography profile example



Wordlist (Source X)

tree tinan
root cha
stem kulu¨
...

Orthography profile: Source X

_, ,
a, a,
aa, aː,
an, ã,
b, b,
c, k,
ch, tʃ,
d, d, only in Spanish loans
¨, , typo
...

Output

concept	original	grapheme_parse	IPA_parse
tree	tina	# t i n an #	# t i n ã #
root	chaa	# ch aa #	# tʃ aː #
stem	kulu¨	# k u l u #	# k u l u #
...			

An orthography profile must also specify orthographic rules, if they exist. Our current approach is to write orthographic rules as regular expressions that match and replace graphemes or sequences of graphemes in matching contexts. We also list them in the orthography profile and apply the rules after the initial graphemic parse has been made.[14] For example, a writing system encodings nasalization of vowels with an <n> following the vowel that it nasalizes, e.g. <an> is a nasalized /ã/. However, when the sequence vowel+n is followed by another vowel, <n> is in fact an /n/. We can specify the regular expression, in Python, for a five-vowel system: "([a|e|i|o|u])(n)(\s)([a|e|i|o|u]), \1 \2 \4". This would take as input a form such as < # t an a # > and rewrite it as < # t a n a # >.[15]

[14] The rules could also be applied before graphemic parsing; the application order chosen is arbitrary.

[15] For outliers, forms may have to be specified at the lexical level, i.e. in some cases it may be easiest to simply list exceptions at the word level.

The graphemes in the orthography profile can then be mapped to IPA representations, as shown in the orthography profile in Figure 8.1, so that there exits an interlingual pivot between the graphemic units of a language and its phonemes. Once graphemes are mapped to phonemes, cross-linguistic queries can be made at the phoneme or grapheme levels (Moran, 2009). In cases of shallow orthographies, this mapping is not particularly problematic. In fact for languages with shallow orthographies, orthographic segments and properties can act as a proxy for phonological segments and phonological analysis can be undertaken (cf. Zuraw 2006). Deep orthographies, like English and French, are problematic and this approach does not answer the problem of mapping graphemes-to-phonemes and vice versa.

Take any linguist's wordlist or dictionary of a lesser-studied language, and one will likely encounter an idiosyncratic orthography, influenced by a number of factors such as: 1) learnability – the orthography of the resource may be influenced by other writing system(s) known by the intended audience or by neighboring languages; 2) theory – the linguist's theoretical training; 3) limitations – depending on when the work was undertaken, technological limitations such as typewriters vs computers and legacy fonts vs Unicode. Also, many orthographies have histories and are often the product of bible translation projects.

Orthography profiles are probably not practical for long-established orthographies like English and French, which have lost much of their phonetic transparency.[16] On the other hand, if we focus on the writing and transcription systems used in lesser-described and endangered languages, orthography profiles are useful for describing writing systems and to transpose them into some form of phonetic transcription. Of course IPA and other transcription systems are essentially just orthographies that have more transparent grapheme-phone correspondences than most systems. Sound-based normalization is practical for undertaking comparative analysis of languages with different writing systems. Orthography profiles also allow us to describe and compare different writing systems at the linguistic and technological levels. And it is a mechanism for specifying additional information such as marginal graphemes (e.g. <j> in Dutch) or additional information that can be useful for linguistic

[16] Note that English and French have large pronunciation lexicons already available, with pronunciations in ARPABET or some similar phonetic alphabet, e.g. the CMU Pronouncing Dictionary at: `http://www.speech.cs.cmu.edu/`.

analysis, such as which graphemes are consonants or vowels.

In summary, an orthography profile lists the graphemes in a particular description of language data, e.g. a wordlist, dictionary or corpus. Building on the knowledge that can be extracted from that description by tokenizing words by the graphemes made explicit in the orthography profile, it is straightforward to undertake other analyses of the data. For example, various ngram models of the data can be extracted with a few lines of code. A unigram model with counts, frequencies and positive log probability provides a fair amount of information about a given data source (Goldsmith and Riggle, 2012). Essentially, the orthography profile provides the description that allows this information to be calculated based on the mapping of sequences of characters into graphemic units. Bigram models are also straightforwardly extracted. In the case of bigrams, mutual information can be captured and used in various other statistical analyses, such as quantitative language comparison, inferring phylogenetic trees, etc. In the next section I discuss some further applications that leverage parsing lexical data at the segment and distinctive feature levels.

8.4 Future work

I conclude this work by briefly describing in this section some avenues for future research.

8.4.1 Information theoretic approaches to phonology

The first avenue builds on the integration of segment inventories, distinctive features and lexicons explored in the previous section. One position taken in regard to phonemes is that analyzing them outside of their context is artificial (Hume and Mailhot, 2011). The reasoning is that communication is encoded in the speech stream and since phonemes are abstractions of contrastive sounds that are used to represent the speech stream, then they should be analyzed within their environments. Thus some areas to investigate are the transitions between phonemes and the relations and transitions between distinctive features of segments within and across words, including long distance relations. One tool to investigate these transitions is information theory.

Since the conception of distinctive features (Trubetzkoy, 1939; Jakobson, 1949; Jakobson et al., 1952; Jakobson and Halle, 1956), information theory has had a significant influence on phonological theory (Hume and Mailhot, 2011). Information theoretic approaches, such as entropy and probability, lend themselves naturally as quantitative measures for many phonological concepts; see for example Hume and Mailhot 2010, Mukherjee et al. 2010, Hume et al. 2011 and Goldsmith and Riggle 2012.[17] For example, distinctive features are not equally informative. Entropy, as measured as the transitions between features in words, is useful for calculating the efficiency and predictiveness of certain features. An information theoretic approach is thus measuring the amount of information encoded in distinctive features within their transitions between words. The current approach in the application of information theoretic concepts to phonological processes is to formulate a hypothesis, e.g. "the effects of vowel harmony in a language like Finnish should result in a decrease in entropy if we condition the probability of a vowel on the vowel that precedes" (Goldsmith and Riggle, 2012, 892), identify a language or set of languages, do the necessary parsing and pre-processing of the data, then apply information theoretic concepts to the data and evaluate the results. The combination of segment inventories, distinctive features and lexicons provides an ideal resource to explore many phonological processes via information theoretic concepts.

8.4.2 Complexity

Another avenue of research is to use PHOIBLE to investigate the issue of measuring and comparing language complexity in phonological systems. An assumed truism in linguistics is that if a language's structure simplifies in one place, it is likely to complicate in another (Hockett, 1955). Thus the complexity of different linguistic subsystems may vary within a given language, these differences balance out cross-linguistically so that all languages are equally complex. The difficulty of course is how to measure complexity. In bioinformatics, "linguistic complexity" is loosely defined as the measure of variations in a string, or sequence, of genome (Kinser, 2009, 241). In both biology and linguistics, a sequence is an ordered collection drawn from a fixed set of characters that constitute the basic unit of replication, e.g. in biology proteins are encoded in an alphabet of 20 letters and in linguis-

[17] For an overview of basic notions of information theory and its relevance to phonology, see Goldsmith 1995.

tics words are encoded with sounds. Whereas biological sequences are very long and have a relatively small alphabet, linguistic sequences are short and are formed from a relatively large set of sounds. Additionally, the alphabet used in biology remains stable; there are mutations in DNA, etc., but the alphabet of sounds in languages are constantly changing due factors beyond genealogical descent, such as societal influences and areal proximity to other language varieties, which cause sound change. In general there are two common measurements for complexity of a linguistic subsystem: absolute complexity (as measured by the number of parts of a system) and relative complexity (the cost or difficulty of using that system). In ongoing work, we are using PHOIBLE to do a cross-linguistic comparison of complexity measures in phonological systems. For absolute complexity measures, these include per language: total number of segments in a language, the ratio of consonants vs vowels, and the frequency of sounds vs their cross-linguistic frequency.[18] Acquiring a phonology is also a process of acquiring contrasts and not inventories, per se. Therefore phoneme inventories may be better understood in terms of contrastive features and phonological contexts (Kabak, 2004). This notion aligns with the idea of relative complexity. Thus we can evaluate the economy and distinctiveness of languages' phonological systems by drawing on principles of information theory, such as Shannon entropy (Shannon, 1948). By modeling segment inventories via their distinctive features, we can evaluate their complexity by calculating their entropy over their feature space and by using dimensionality reduction to determine the number of phonetic dimensions minimally needed to describe a given inventory. Once a complexity value for each method for each phoneme inventory is calculated, we can evaluate if these measures correlate with each other, and whether they correlate with other variables, such as genealogical lineage, geographic area and population size, as encoded in PHOIBLE or elsewhere.

[18] Frequency is often related to the notion of markedness (or rarity). Some researchers have reportedly found a link between complexity and rarity. For example, see: Edmonds 1999, Harris 2008 and Sinnemaeki 2011.

8.4.3 Feature-based principles in phonological inventories

There is much evidence that points towards segments, features and sound patterns as emergent probabilistic properties that rise from factors of language usage, including articulatory and perceptual biases, and self-organizing and feature-based principles that appear to govern the structure of phonological inventories (see Blevins 2004, Mielke 2008 and Mohanan et al. 2009 and references therein). Investigating feature-based principles is another avenue of future research that can be investigated with the segment inventories and distinctive features in the PHOIBLE data set.

Building on previous work, including de Groot 1931, Martinet 1955, Martinet 1968, Clements 2003a and Clements 2003b, Clements (2009) presents a detailed description of the effects of features on the typology of segment inventories in terms of five principles: Feature Bounding, Feature Economy, Marked Feature Avoidance, Robustness, and Phonological Enhancement. Feature bounding[19] and feature economy[20] are rather distinct properties from non-feature-based alternatives to phonological theory (Mielke, 2009) and are both directly testable with the given PHOIBLE knowledge base. Additionally, using PHOIBLE these phonetic-feature based principles can be investigated in coordination with other typological variables, such as genealogical and geographic factors.

8.4.4 Correlation studies

As discussed in Chapter 7, there are numerous studies that associate ecological or demographic parameters with changes in linguistic systems. These studies include, but are not limited to: the degree of ecological risk shapes linguistic diversity in West Africa (Nettle, 1996); languages spoken in warm climates tend to use more high-sonority sounds than languages spoken in cold climates (Munroe et al., 1996; Munroe and Silander, 1999; Fought et al., 2004; Ember and Ember, 2007; Munroe et al., 2009); degree of baby-holding is more predictive of the percentage of CV syllables in words than climate or literacy (Ember and

[19] The feature bounding principles states that features set an upper bound on both the number of sounds and the number of phonemic contrasts that may appear in a language (Clements, 2009, 24-25).

[20] Feature economy is the tendency of a segment inventory to maximize feature combinations in the segment inventory (Clements, 2009, 27).

Ember, 1999); languages spoken by small numbers of speakers have disproportionately small or large phonemic inventories (Trudgill, 2004a); there exists a robust correlation between population size and phoneme inventory size (Hay and Bauer, 2007); languages with smaller groups of speakers have more complex inflectional morphology than larger groups of speakers (Lupyan and Dale, 2010); there is a negative correlation between phoneme inventory size of languages and their geographic distance from West Africa (Atkinson, 2011). For studies that claim there is a correlation between a linguistic or non-linguistic parameter[21] and the phonological system of languages, PHOIBLE is a useful resource for revisiting claims of correlation, as I've shown in Chapter 7. I would like to revisit these studies and retest claims made in them.

8.4.5 Tackling provenance

As discussed in Section 2.3.6, linguistic records are data that are ripe for addressing issues of data provenance. The phonemic analysis of a given segment inventory can be the work of a scholar who has consulted multiple descriptions of a particular language. The resulting segment inventory is often then reanalyzed by a subsequent scholar. Ideally the PHOIBLE data set would then contain not only metadata for the original descriptions, but also the trail of reinterpretations of the segment inventory.

Very recently, the World Wide Web Consortium (W3C) Provenance Working Group[22] was formed and it set itself the goal of identifying the issues of data provenance on the Web. Their aim is to publish recommendations that define a language for data provenance information interchange. So far the group has produced a working draft and a preliminary data model for specifying and encoding provenance on the Web and "for building representations of the entities, people and processes involved in producing a piece of data or thing in the world".[23] As this working draft matures into a W3C standard, the bibliographic

[21] The parameters are not limited to what is currently in the PHOIBLE data set because its extensible model allows additional data sets to be added to the system. What is needed is a mapping between some parameter and an ISO 639-3 language code.

[22] http://www.w3.org/2011/prov/wiki/Main_Page

[23] http://www.w3.org/TR/prov-primer/

data from PHOIBLE may be incorporated into their "PROV"(enance) model and thus will provide provenance metadata records that are intended to be compatible and interoperable with existing Semantic Web standards, including RDF. This model would allow users to track data provenance, such as linguists' different interpretations of the same phonological description, the reuse and modification of the same source and additional modifications by users to the data set.[24]

8.4.6 Linked Open Data

A final avenue for future research that I will discuss is the path towards what is currently called a *cyberinfrastructure* for linguistics, i.e. the next generation of technological infrastructure for computational methods and linguistic research.[25] The primary purpose of cyberinfrastructure is to ensure access to data (Bender and Langendoen, 2010, 11). One way to do so is to publish data on the Web in an open and accessible format. This process can be quite straightforward if you follow the recommendations in Bird and Simons 2003 and more recently in Abney and Bird 2010 for publishing linguistic data in a simple storage model. For example, the flat file tables from PHOIBLE are published online in a simple delimiter separated format. The data are straightforwardly interpretable and the tables can be read in as input and their contents can also be easily parsed to extract desired data. Simply putting data on the Web in a simple storage format, however, does not necessarily ensure access to the data. If the data are not published with an explicit license, then users cannot know the state of the copyright permissions of the data.[26] Furthermore, a simple storage format does not mean that the data can be harmonized with other linguistic data sets without processing them in some way to make them comparable with other storage formats, i.e. make them structurally interoperable. On the other hand, publishing linguistic data as Linked (Open) Data is one avenue towards technological infrastructure for sharing linguistic data.

[24] Note that some provenance information would simply not be available, such as information about the history of certain resources before they got to PHOIBLE.

[25] For more information, see the Cyberling blog: `http://blog.cyberling.org/`.

[26] See discussion of Creative Commons licenses: `http://creativecommons.org/licenses/`.

Although originally developed as a data model for representing metadata, RDF has evolved into a generic data format for knowledge representation. It has become part of the foundation of the Semantic Web, or "Web of data" (Berners-Lee et al., 2001). The aim of the Semantic Web is to create a common framework for sharing and reusing data, on the Web, which are designed to be interpretable by machines and humans. RDF is a mature technology and it has a large and active community of developers that have provided it with a rich infrastructure of tools, including APIs, query languages and sub-languages like the Web Ontology Language (OWL), which can be used to create a reserved vocabulary and logic constraints for RDF data to attain semantic interoperability between resources. RDF is one component of Linked Data.[27]

Linked Data is a W3C initiative that aims to connect data sets across the Web by interlinking them and using standard Web technologies like URIs, HTTP and content negotiation that serve to share information and to deliver it in either a machine-readable or human interpretable format. Linked Data practices describe methods for publishing structured data to leverage these Semantic Web technologies for data federation and querying of distributed resources. There is currently a so-called 5-star rating system for publishing Linked Open Data.[28] The first star is achieved by simply publishing data, in any format, on the Web under an open license. The second star is reached if the data are also available as machine-readable structured data, e.g. a dictionary in electronic accessible text instead of a PDF scan of a print dictionary. If the data are available in a non-proprietary format, e.g. a plain or Unicode text in table form instead of an Excel spreadsheet, they acquire three stars. If the data have attained three stars and additionally use open standards from W3C, e.g. RDF and SPARQL, to identify things with URIs, then the data set is rated as four-star. Finally, if the data set has reached four stars and also links to other people's data, then it is considered a five-star Linked Data resource, which means the data set: uses URIs as names for things; uses HTTP and URIs so that users can look up those names; returns useful information to humans and bots via its URIs; and contains links to other

[27] http://linkeddata.org/

[28] http://www.w3.org/DesignIssues/LinkedData

URIs in other data sets, thus making it Linked Data.

The concept of Linked Data is closely coupled with the idea of *openness*. In fact, part of the push for Linked Data originates in the desire for government transparency and accountability, and for kick starting new data-based economies by making data easily accessible and interpretable. The Open Knowledge Foundation (OKFN)[29] defines "openness" as: "A piece of content or data is open if anyone is free to use, reuse, and redistribute it – subject only, at most, to the requirement to attribute and/or share-alike."[30] The movement for Linked Open Data in linguistics is spear-headed by the Open Working Group in Linguistics (OWLG).[31]

The OWLG provides a platform for sharing experiences and technology, for promoting the publication of linguistic data as Linked Open Data and for maintaining an index of open linguistic data sources and tools that link existing resources in the form of a Linked Linguistics Open Data cloud (LLOD).[32] As I have shown in this work, RDF is a suitable representation format for modeling a typological database. The graph model that underlies RDF is also being used as the underlying abstract data structure for other linguistic data types, such as linguistic markup (Farrar and Langendoen, 2003) and annotated corpora and linguistic annotations (Ide and Suderman, 2007). Implementing these resources in RDF and then creating Linked Data is straightforward due to the shared underlying data structure. Additionally, there already exists many standards for semantic interoperability, which are prime for conversion to RDF, OWL and Linked Data.[33]

If linguistic resources are published in accordance with the set of principles put forth by the Linked Open Data initiative, a web of linguistic data makes it possible for linguists to

[29] `http://okfn.org/`

[30] `http://opendefinition.org/`

[31] `http://linguistics.okfn.org/`

[32] `http://linguistics.okfn.org/resources/llod/`

[33] These standards include, but are not limited to: Unicode for encoding characters; IPA for phonetic segments; ToBI for prosody and intonation (Silverman et al., 1992); the Leipzig Glossing Rules for interlinear glossed text (Comrie et al., 2003); ISOCat for describing morphosyntactic categories; the Text Encoding Initiative (TEI) for encoding literary and linguistic texts; the Dublin Core Metadata Initiative (DCMI) and Open Language Archive Community (OLAC) for metadata categories (Bird and Simons, 2003); ISO 639-3 for unique language name identifiers, etc.

not only share data, but also to follow links between existing resource to find and access new data. In this work I instantiated a typological database in RDF and I demonstrated how the RDF graph model is a flexible structure that reduces the challenges of attaining syntactic interoperability with other data sets, when they are also modeled in RDF and each data set uses some of the same URIs. Attaining syntactic interoperability lays the groundwork for achieving semantic interoperability, i.e. when resources share, reuse or link the same vocabularies so that information from one resource can be resolved against information from another resource. The PHOIBLE RDF/OWL data set is now being improved into five-star Linked Data and being added to the LLOD.[34] Publishing linguistic resources as Linked Data helps to overcome the challenges of syntactic and semantic interoperability. This is one path towards the next generation of technological infrastructure and open data sharing in linguistics.

[34] For more information, go to: `http://phoible.org/`.

BIBLIOGRAPHY

Abdalla, A. I. (1973). *Kadugli Language and Language Usage*, volume 3 of *Salambi Prize Series*. Khartoum University Press, Khartoum.

Abdel-Massih, E. T. (1973). *An Introduction to Moroccan Arabic.* Center for Near Eastern and North African Studies, University of Michigan, Ann Arbor, MI.

Abdulla, J. J. and McCarus, E. N. (1967). *Kurdish Basic Course.* University of Michigan Press, Ann Arbor.

Abega, P. (1970). *Grammaire Ewondo.* Département des langues africaines et linguistique, Université Fédérale du Cameroun, Yaoundé.

Abessolo Eto, R. (1990). Esquisse phonologique du bongo. Master's thesis, Universite de Yaounde.

Abney, S. and Bird, S. (2010). The Human Language Project: Building a Universal Corpus of the World's Languages. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 88–97.

Abraham, R. C. (1934). *The Principles of Hausa Vol. 1. Government Printer.* Nigeria Political Service, Kaduna.

Abraham, R. C. (1959a). *Hausa Literature and the Hausa Sound System.* University of London Press, London.

Abraham, R. C. (1959b). *The Language of the Hausa People.* University of London Press, London.

Abramson, A. S. (1962). *The Vowels and Tones of Standard Thai: Acoustical Measurements and Experiments*, volume 28 of *International Journal of American Linguistics.* Indiana University, Bloomington.

Adams, K. and Lauck, L. (1975). A Tentative Phonemic Statement of Patep. In Loving, R., editor, *Workers in Papua New Guinea Languages: Phonologies of Five Austronesian Languages*, volume 13. Summer Institute of Linguistics.

Agard, F. B. (1958). *Structural Sketch of Rumanian*, volume 26 of *Supplement to Language Monographs*. Linguistic Society of America, Baltimore.

Agnew, A. and Pike, E. G. (1957). Phonemes of Ocaina (Huitoto). *International Journal of American Linguistics*, 23:24–27.

Ahua, M. B. (2004). *Conditions linguistiques pour une orthographe de l'agni: une analyse contrastive des dialectes sanvi et djuablin*. PhD thesis, Universitäat Osnabrück.

Akeriweh (2000). A Step Towards the Standardisation of Kànswéynséy (A Grassfield Bantu Language). Master's thesis, Universite de Yaounde I.

Akumbu, P. W. (1999). Nominal Phonological Processes in Babanki. Master's thesis, University of Yaounde I.

Allan, E. J. (1974). Likpe. In Kropp-Dakubu, M. E., editor, *West African Langauge Data Sheets*, volume 2, pages 89–94. West African Linguistics Society.

Allan, E. J. (1976a). Dizi. In Bender, M. L., editor, *The Non-Semitic Languages of Ethiopia*, pages 377–392. African Studies Center, Michigan State University, East Lansing.

Allan, E. J. (1976b). Kullo. In Bender, M. L., editor, *The Non-Semitic Languages of Ethiopia*, pages 324–350. African Studies Center, Michigan State University, East Lansing.

Allen, C. (1973). Sele. In Kropp-Dakubu, M. E., editor, *West African Langauge Data Sheets*, volume 2, pages 208–215. West African Linguistics Society.

Allen, J. and Beason, M. (1975). Petats Phonemes and Orthography. In Loving, R., editor, *Workers in Papua New Guinea Languages: Phonologies of Five Austronesian Languages*, volume 13. Summer Institute of Linguistics.

Allen, W. S. (1950). Notes on the Phonetics of an Eastern Armenian Speaker. *Transactions of the Philological Society*, pages 180–206. Reprint by Hertford 1951.

Allin, T. R. (1976). *A Grammar of Resígaro.* Summer Institute of Linguistics, Horsleys Green, United Kingdom.

Alvarez, J. J. V. (2002). Morfologia del verbo de la lengua chol de Tila, Chiapas. Master's thesis, Instituto Nacional Indigenista.

Ambrazas, V., Vajtkavichjute, V., Valjatskene, A., Morkunas, K., Sabaljauskas, A., and Ul'vidas, K. (1966). Litovskij jazyk. In Vinogradov, V. V., editor, *Jazyki narodov SSSR. Volume 1: Indoevropejskie jazyki*, pages 500–527. Nauka, Leningrad / Moscow.

Anceaux, J. C. (1965). *The Nimboran Language: Phonology and Morphology.* Martinus Nijhoff, The Hague.

Andersen, T. (1987a). An Outline of Lulubo Phonology. *Studies in African Linguistics*, 18(1):39–65.

Andersen, T. (1987b). The Phonemic System of Agar Dinka. *Journal of African Languages and Linguistics*, 9(1):1–27.

Andersen, T. (1987c). The Phonemic System of Dinka. *Journal of African Languages and Linguistics*, 9(1):1–27.

Andersen, T. (1992). Aspects of Mabaan Tonology. *Journal of African Languages and Linguistics*, 13(2):183–204.

Andersen, T. (2004). Jumjum Phonology. *Studies in African Linguistics*, 33(2):133–162.

Andersen, T. (2006). Kurmuk Phonology. *Studies in African Linguistics*, 35(2):29–90.

Anderson, D. (1962). *Conversational Ticuna.* Instituto Lingüístico de Verano, Yarinacocha, Peru.

Anderson, D. (2003). Using the Unicode Standard for Linguistic Data: Preliminary Guidelines. In *Proceedings E-MELD Conference 2003: Digitizing and Annotating Texts and Field Recordings.*

Anderson, G. D. S. (2011). The Velar Nasal. In Dryer, M. S. and Haspelmath, M., editors, *The World Atlas of Language Structures Online*. Max Planck Digital Library, Munich.

Anderson, L. (1959). Ticuna Vowels: With Special Regard to the System of Five Tonemes. *Serie Linguistica Especial do Museu Nacional, Rio de Janeiro*, 1:79–119.

Anderson, R. D. (1963). *A Grammar of Laz*. PhD thesis, University of Texas.

Anderson, S. (1972). On Nasalization in Sundanese. *Linguistic Inquiry*, 3(3):253–268.

Anderson, S. C. (1980). The Noun Class System of Amo. In Hyman, L. M., editor, *Noun Classes in the Grassfields Bantu Borderland*, volume 8 of *Southern California Occasional Papers in Linguistics*, pages 155–178. University of Southern California, Los Angeles.

Anderson, S. R. (1978). Syllables, Segments, and the Northwest Caucasian Languages. In Bell, A. and Hooper, J. B., editors, *Syllables and Segments*, pages 47–58. North Holland Publishing Co., Amsterdam.

Anderson, S. R. (1985). *Phonology in the Twentieth Century: Theories of Rules and Theories of Representations*. University of Chicago Press, Chicago, IL.

Anderson, T. (1986). Tone Splitting and Vowel Quality: Evidence From Lugbara. *Studies in African Linguistics*, 17:55–68.

Anderson, V. B. (2000). *Giving Weight to Phonetic Principles: The Case of Place of Articulation in Western Arrernte*. PhD thesis, The University of California at Los Angeles.

Anderton, A. (1989). The Sounds of Cacua, Based on Data Collected by the Summer Institute of Linguistics.

Andreev, I. A. (1966). Chuvashskij jazyk. In Baskakov, N. A., editor, *Jazyki narodov SSSR. Volume 2: Tjurkskie jazyki*, pages 43–65. Nauka, Moscow.

Andrews, K. R. (1994). *Shawnee Grammar*. PhD thesis, University of South Carolina.

Andrzejewsky, B. W. (1955). The Problem of Vowel Representation in the Isaaq Dialect of Somali. *Bulletin of the School of Oriental and African Studies*, 17:567–580.

Andrzejewsky, B. W. (1956). Accentual Patterns in Verbal Forms in the Isaaq Dialect of Somali. *Bulletin of the School of Oriental and African Studies*, 18:103–129.

Andvik, E. E. (1999). *Tshangla Grammar*. PhD thesis, University of Oregon.

Anonby, E. J. (2006). Illustrations of the IPA: Mambay. *Journal of the International Phonetic Association*, 36(2):221–233.

Anonymous (1927). *Savara*. Tea Districts Labour Association.

Anonymous (1982). Da-gon-er Yu Jianzhi. A Brief Guide to the Da-gon-er (Dagur) Language. In Bao, S.-q., editor, *No Booktitle*. Minzu Chubanshe, Beijing.

Ansre, G. (1961). The Tonal Structure of Ewe. Master's thesis, Hartford Seminary Foundation, Hartford.

Antunes, G. (2004). *A Grammar of Sabanê: A Nambikwaran Language*. PhD thesis, Vrije Universiteit.

Aoki, H. (1966). Nez Perce Vowel Harmony and Proto-Sahaptian Vowels. *Language*, 42:759–767.

Aoki, H. (1970a). A Note on Glottalized Consonants. *Phonetica*, 21(2):65–74.

Aoki, H. (1970b). *Nez Perce Grammar*, volume 62 of *Publications in Linguistics*. University of California Press, Berkeley. Reprinted 1973, California Library Reprint series.

Applegate, J. R. (1958). *An Outline of the Structure of Shilha*. American Council of Learned Societies, New York.

Arensen, J. E. (1982). *Murle Grammar*, volume 2 of *Occasional Papers in the Study of Sudanese Languages*. Summer Institute of Linguistics and University of Juba, Juba, Sudan.

Armendáriz, R. G. F. (2005). *A Grammar of River Warihío*. PhD thesis, Rice University.

Armstrong, L. E. (1964). The Phonetic Structure of Somali. *Mitteilungen des Seminars für Orientalische Sprachen Berlin*, 37(3):116–161.

Armstrong, R. G. (1968). Yala (Ikom) a Terraced Level Language With Three Tones. *Journal of West African languages*, 5(1):37–52.

Arnott, D. W. (1968a). Fula (Nigeria). In Kropp-Dakubu, M. E., editor, *West African Langauge Data Sheets*, volume 1, pages 233–244. West African Linguistics Society.

Arnott, D. W. (1968b). Tiv. In Kropp-Dakubu, M. E., editor, *West African Langauge Data Sheets*, volume 2, pages 241–248. West African Linguistics Society.

Aronson, H. I. (1968). *Bulgarian Inflectional Morphophonology.* Mouton, The Hague.

Arroyo, V. M. (1972). *Lenguas Indigenas Costarricenses.* Editorial Universitaria Centroamericana, San José, Costa Rica.

Asal, B. (1969). *The Sedik Language of Formosa.* Cercle Linguistique de Kanazawa.

Aschmann, H. P. (1946). Totonaco Phonemes. *International Journal of American Linguistics*, 12:34–43.

Asobo, I. S. (1989). The Noun Class System of Kɔlɛ. Master's thesis, University of Yaounde.

Asyik, A. G. (1987). *A Contextual Grammar of Acehnese Sentences.* PhD thesis, University of Michigan.

Atkinson, Q. D. (2011). Phonemic Diversity Supports a Serial Founder Effect Model of Language Expansion From Africa. *Science*, 332:346–359.

Atta, S. E. (1993). The Phonology of Lukundu (Bakundu). Master's thesis, University of Yaounde I.

Augustitis, D. (1964). *Das lithauische Phonationssystem*, volume 12 of *Slavistiche Beiträge.* Otto Sagner, München.

Austerlitz, R. (1956). Gilyak Nursery Words. *Word*, 12:260–265.

Austerlitz, R. (1967). The Distributional Identification of Finnish Morphophonemes. *Language*, 43:20–33.

Austin, P. (1981). *A Grammar of Diyari, South Australia.* Cambridge University Press, Cambridge.

Austin, W. (1962). The Phonemics and Morphophonemics of Manchu. In Poppe, N., editor, *American Studies in Altaic Linguistics*, volume 13 of *Uralic and Altaic Series*, pages 15–22. Indiana University Press, Bloomington.

Avrorin, V. A. (1968). Nanajski jazyk. In Skorik, P. J., editor, *Jazyki narodov SSSR. Volume 5: Mongol'skie, tunguso-man'chzhurskie i paleoaziatskie jazyki*, pages 129–148. Nauka, Leningrad and Moscow.

Awah, V. N. (1997). WH-Movement in Mungaka: A Generative Approach. Master's thesis, University of Yaounde I.

Awobuluyi, A. O. (1971). The Phonology of Yerwa Kanuri. *Research Notes of the Department of Linguistics and Nigerian Languages*, 4(1):1–21.

Baader, F., Calvanese, D., McGuinness, D., Nardi, D., and Patel-Schneider, P. (2003). *The Description Logic Handbook: Theory, Implementation, and Applications.* Cambridge, UK: Cambridge University Press.

Baader, F., Horrocks, I., and Sattler, U. (2008). Description Logics. In van Harmelen, F., Lifschitz, V., and Porter, B., editors, *Foundations of Artificial Intelligence: Handbook of Knowledge Representation*, volume 3. Elsevier.

Baader, F. and Nutt, W. (2003). Basic Description Logics. In *The Description Logic Handbook: Theory, Implementation, and Applications.* Cambridge University Press.

Baader, F. and Sattler, U. (2001). An Overview of Tableau Algorithms for Description Logics. *Studia Logica*, 69(1):5–40.

Baayen, H. (2010). languageR: Data Sets and Functions With "Analyzing Linguistic Data: A Practical Introduction to Statistics" (Version 1.0) [R Package]. Retrieved from http://cran.r- project.org/web/packages/languageR/index.html.

Baayen, R. H., Davidson, D. J., and Bates, D. M. (2008). Mixed-effects Modeling With Crossed Random Effects for Subjects and Items. *Journal of Memory and Language*, 59(4):390–412.

Bacelar, L. N. (2004). *Gramática Da Língua Kanoê*. PhD thesis, Katholieke Universiteit Nijmegen.

Badie, M. J. (1995). *Contribution a une etude morphosyntaxique du n'cam*. PhD thesis, University of Paris VII.

Baird, L. (2002). Illustrations of the IPA: Kéo. *Journal of the International Phonetic Association*, 32(1):93–97.

Baker, B. J. (1999). *Word Structure in Ngalakgan*. PhD thesis, University of Sydney.

Bakker, D. (2011). Language Sampling. In Song, J. J., editor, *Handbook of Linguistic Typology*. Oxford University Press, Oxford, UK.

Bakker, P. (2004). Phoneme Inventories, Language Contact, and Grammatical Complexity: A Critique of Trudgill. *Linguistic Typology*, 8(3):368–375.

Ballantyne, K. G. (2005). *Textual Structure and Discourse Prominence in Yapese Narrative*. PhD thesis, University of Hawai'i.

Ballard, W. L. (1975). Aspects of Yuchi Morphonology. In Crawford, J. M., editor, *Studies in Southeastern Indian Languages*, pages 164–187. University of Georgia Press, Athens, Georgia.

Baltaxe, C. A. M. (1978). *Foundation of Distinctive Feature Theory*. University Park Press.

Bambose, A. (1982). Issues in the Analysis of Serial Verb Constructions. *Journal of West African Languages*, 12(2):3–21.

Bamgbose, A. (1966). *A Grammar of Yoruba*. Cambridge University Press, Cambridge.

Bamgbose, A. (1967). Notes on the Phonology of Mbe. *Journal of West African Languages*, 4(1):5–11.

Bandhu, C. M., Dahal, B. M., Holzhausen, A., and Hale, A. (1971). *Nepali Segmental Phonology*. Summer Institute of Linguistics, Tribhuvan University, Kirtipur.

Banfield, A. W. (1914). *Dictionary of the Nupe Language*. The Niger Press, Shonga, N. Nigeria, W. Africa.

Bangha, G. F. (2003). The Mmen Noun Phrase. Master's thesis, University of Yaounde I.

Banhidi, Z., Jokay, Z., and Szabo, D. (1965). *Lehrbuch der ungarischen Sprache*. Publishing House for Textbooks, Budapest.

Barabási, L. and Albert, R. (1999). Emergence of Scaling in Random Networks. *Science*, 286:509–512.

Barker, M. A. R. (1964). *Klamath Grammar*, volume 32 of *University of California Publications in Linguistics*. University of California Press, Berkeley.

Barrett, E. R. (1999). *A Grammar of Sipakapense Maya*. PhD thesis, The University of Texas at Austin.

Bateman, J. A. (1995). On the Relationship Between Ontology Construction and Natural Language: A Socio-semiotic View. *International Journal of Human-Computer Studies*, 43:929–944.

Bates, D., Maechler, M., and Bolker, B. (2011). lme4: Linear Mixed-effects Models Using S4 Classes (Version 0.999375-38) [R Package]. Retrieved from http://lme4.r-forge.r-project.org/.

Bauer, L. (2007). *The Linguistics Student's Handbook*. Edinburgh University Press, Edinburgh.

Bauer, L., Warren, P., Bardsley, D., Kennedy, M., and Major, G. (2007). Illustrations of the IPA: English, New Zealand. *Journal of the International Phonetic Association*, 37(1):97–102.

Bauernschmidt, A. (1965). Amuzgo Syllable Dynamics. *Language*, 41:471–483.

Baumbach, E. J. (1997a). Languages of the Eastern Caprivi. In Haacke, W. H. and Elderkin, E. E., editors, *Namibian Languages: Reports and Papers*, volume 4 of *Namibian African Studies*. Rüdiger Köppe.

Baumbach, E. J. (1997b). Languages of the Eastern Caprivi. In Haacke, W. H. and Elderkin, E. E., editors, *Namibian Languages: Reports and Papers*, volume 4 of *Namibian African Studies*. Rüdiger Köppe.

Baumbach, E. J. (1997c). Languages of the Eastern Caprivi. In Haacke, W. H. and Elderkin, E. E., editors, *Namibian Languages: Reports and Papers*, volume 4 of *Namibian African Studies*. Rüdiger Köppe.

Baumbach, E. J. (1997d). Languages of the Eastern Caprivi. In Haacke, W. H. and Elderkin, E. E., editors, *Namibian Languages: Reports and Papers*, volume 4 of *Namibian African Studies*. Rüdiger Köppe.

Beach, D. M. (1938). *The Phonetics of the Hottentot Language*. William Heffer & Sons, Cambridge.

Beam de Azcona, R. G. (2004). *A Coatlan-Loxicha Zapotec Grammar*. PhD thesis, University of California, Berkeley.

Bearth, T. and Zemp, H. (1967). The Phonology of Dan (Santa). *Journal of African Languages*, 6:9–29.

Beasley, D. and Pike, K. L. (1957). Notes on Huambisa Phonemics. *Lingua Posnaniensis*, 6:1–8.

Beaton, A. C. (1968). *A Grammar of the Fur Language*, volume 1 of *Linguistic Monograph Series*. Sudan Research Unit, University of Khartoum, Khartoum.

Beaumont, C. H. (1979). *The Tigak Language of New Ireland*, volume 58 of *Pacific Linguistics, Series B*. Australian National University, Canberra.

Beckett, D. (2004). RDF/XML Syntax Specification (Revised). Technical report, W3C.

Beckman, M. E. and Venditti, J. J. (2010). Tone and Intonation. In Hardcastle, W. J., Laver, J., and Gibbon, F. E., editors, *The Handbook of Phonetic Sciences, Second edition*. Blackwell Publishing.

Bee, D. (1965a). *Usarufa: A Descriptive Grammar*. PhD thesis, Indiana University.

Bee, D. (1965b). Usarufa Distinctive Features and Phonemes. *Linguistic Circle of Canberra Publications A*, 6:39–68.

Beeler, M. S. (1970). Sibilant Harmony in Chumash. *International Journal of American Linguistics*, 36(1):14–17.

Begné II, L. P. (1979). *The Phonology of Bikele, a Cameroonian Language*. PhD thesis, Illinois Institute of Technology.

Bell, A. (1978). Language Samples. In Greenberg, J. H., editor, *Universals of Human Language Volume 1: Method and Theory*, pages 123–156. Stanford: Stanford University Press.

Bell, A. M. (1867). *Visible Speech: The Science of Universal Alphabetics*. London: Simpkin, Marshal.

Bell, H. (1968). The Tone System of Mahas Nubian. *Journal of African Languages*, 7(1):26–32.

Bell, H. (1971). The Phonology of Nobiin Nubian. *African Language Review*, 9:115–139.

Benaissa, T. (1979). Fonologia del Saliba. In E., C. M. e. a., editor, *Sistemas fonológicos de idiomas colombianos 4*, volume 4, pages 89–98. Ministerio de Gobierno and Instituto Lingüístico de Verano, Bogotá.

Bender, E. and Bender, Z. S. (1946). The Phonemes of North Carolina Cherokee. *International Journal of American Linguistics*, 12(1):14–21.

Bender, E. M. and Langendoen, D. T. (2010). Computational Linguistics in Support of Linguistic Theory. *Linguistic Issues in Language Technology (LiLT)*, 3(2):1–31.

Bender, M. L. (1968). Analysis of a Barya Word List. *Anthropological Linguistics*, 10(9):1–24.

Bendor-Samuel, D. (1966). *Hierarchical Structures in Guajajara.* Summer Institute of Linguistics.

Bendor-Samuel, J. (1961). *The Verbal Piece in Jebero*, volume 4 of *Linguistic Circle of New York Monograph.* Linguistic Circle of New York, New York.

Bendor-Samuel, J., Olsen, E. J., and White, A. R. (1989). Dogon. In Bendor-Samuel, J., editor, *The Niger-Congo Languages*, pages 169–177. University Press of America and Summer Institute of Linguistics, Lanham.

Bendor-Samuel, J. T. and Hartell, R. L., editors (1989). *The Niger-Congo Languages: A Classification and Description of Africa's Largest Language Family.* University Press of America, Lanham, MD.

Bergman, R., Gray, I., and Gray, C. (1969). *Collected Field Reports on the Phonology of Tampulma.* Institute of African Studies, University of Ghana, Legon.

Bergsland, K. (1956). Some Problems of Aleut Phonology. In Halle, M., editor, *For Roman Jakobson*, pages 38–43. Mouton, The Hague.

Bergsland, K. (1959). *Aleut Dialects of Atka and Attu*, volume 49 of *Transactions of the American Philosophical Society.* American Philosophical Society, Philadelphia.

Berners-Lee, T., Hendler, J., and Lassila, O. (2001). The Semantic Web. *Scientific American*, 284(5):34–43.

Berry, J. (1951a). *The Pronunciation of Ewe.* Heffer, Cambridge.

Berry, J. (1951b). *The Pronunciation of Ga.* Heffer, Cambridge.

Berthiaume, S. C. (2003). *A Phonological Grammar of Northern Pame.* PhD thesis, The University of Texas at Arlington.

Beukema, R. W. (1975). *A Grammatical Sketch of Chimborazo Quichua*. PhD thesis, Yale University.

Bhat, D. N. S. (1967). *Descriptive Analysis of Tulu*, volume 15 of *Building Centenary and Silver Jubilee Series*. Deccan College Postgraduate and Research Institute, Pune.

Bhat, D. N. S. (1968). *Boro Vocabulary*. Deccan College, Pune.

Bhat, R. (1987). *A Descriptive Study of Kashmiri*. Amar Prakashan, Delhi.

Bhattacharjya, D. (2001). *The Genesis and Development of Nagamese: Its Social History and Linguistic Structure*. PhD thesis, City University of New York.

Bhattacharya, P. C. (1977). *A Descriptive Analysis of the Boro Language*. Gauhati University.

Bickel, B. (2008). A Refined Sampling Procedure for Genealogical Control. *Sprachtypologie und Universalienforschung*, 61:221–233.

Bickel, B. (2010). Capturing Particulars and Universals in Clause Linkage: A Multivariate Analysis. In Bril, I., editor, *Clause-Hierarchy and Clause-Linking: The Syntax and Pragmatics Interface*, pages 51–101. Benjamins, Amsterdam.

Bickel, B. (In Press). Distributional Biases in Language Families. In Bickel, B., Grenoble, L. A., Peterson, D. A., and Timberlake, A., editors, *Language Typology and Historical Contingency: A Festschrift to Honor Johanna Nichols*. Benjamins, Amsterdam. Online: http://www.uni-leipzig.de/ bickel/research/papers/stability.fsjn.2011bickelrevised.pdf.

Bickmore, L. (2007). *Cilungu Phonology*. CSLI.

Bickoe, D. H. (2000). Viatlité et morphologie verbale du fulfulde fuunaangere. Master's thesis, University of Yaounde I.

Bidwell, C. (1968). The Stress Patterns of the Noun in Bulgarian. *Studies in Linguistics*, 20:41–47.

Biligiri, H. S. (1965). *Kharia: Phonology, Grammar and Vocabulary*, volume 3 of *Building Centenary and Silver Jubilee Series*. Deccan College Postgraduate and Research Institute, Poona.

Bills, G. D., Vallejo, C. B., and Troike, R. C. (1969). *An Introduction to Spoken Bolivian Quecha*. University of Texas Press for Institute of Latin American Studies, Austin.

Bird, C., Hutchinson, J., and Kante, M. (1977). *An Kan Bamanakan Kalan: Beginning Bambara*. Indiana University Linguistics Club, Bloomington.

Bird, S. and Simons, G. F. (2003). Seven Dimensions of Portability for Language Documentation and Description. *Language*, 79(3):557–582.

Birk, D. B. W. (1975). *The Phonology of Malakmalak*, volume 39 of *Pacific Linguistics, Series A*. Australian National University, Canberra.

Black, K. and Black, K. (1971). *The Moro Language Grammar and Dictionary*, volume 6 of *Linguistic Monograph Series*. Sudan Research Unit, Faculty of Arts, Khartoum.

Blackings, M. and Fabb, N. (2003a). *A Grammar of Ma'di*. Number 32 in Mouton Grammar Library. Mouton de Gruyter.

Blackings, M. and Fabb, N. (2003b). *A Grammar of Ma'di*, volume 32 of *Mouton Grammar Library*. Mouton de Gruyter, Berlin.

Blake, B. J. (1979). *A Kalkatungu Grammar*, volume 57 of *Pacific Linguistics, Series B*. Australian National University, Canberra.

Blench, R. (2005a). A Dictionary of Ibani, an Ijoid Language of the Niger Delta. Online: `http://www.rogerblench.info/`.

Blench, R. (2005b). Kirieni Okueingbolu Diri: Okrika Dictionary. Online: `http://www.rogerblench.info/`.

Blench, R. (2006a). A Dictionary of Mada, a Plateau Language of Central Nigeria. Online: `http://www.rogerblench.info/`.

Blench, R. (2006b). A Dictionary of the Jili (Migili) Language of Central Nigeria. Online: http://www.rogerblench.info/.

Blench, R. (2006c). A dictionary of ekpeye, an igboid language of southern nigeria. Online: http://www.rogerblench.info/.

Blench, R. and Hepburn, I. D. (2006). A Dictionary of Eggon. Draft Manuscript January 2, 2006.

Blevins, J. (2004). *Evolutionary Phonology: The Emergence of Sound Patterns*. Cambridge University Press.

Blevins, J. (2009). Another Universal Bites the Dust: Northwest Mekeo Lacks Coronal Phonemes. *Oceanic Linguistics*, 48(1):264–273.

Blight, R. C. and Pike, E. V. (1976). The Phonology of Tenango Otomi. *International Journal of American Linguistics*, 42:51–57.

Bloch, B. (1948). A Set of Postulates for Phonemic Analysis. *Language*, 24(1):3–46.

Bloch, B. (1950). Studies in Colloquial Japanese. Part 4: Phonemics. *Language*, 26:86–125.

Block, K. L. (1994). Discourse Grammar of First Person Narrative in Plang. Master's thesis, University of Texas at Arlington.

Blood, D. L. (1967). Phonological Units in Cham. *Anthropological Linguistics*, 9:15–32.

Bloomfield, L. (1917). *Tagalog Texts With Grammatical Analysis*, volume 3 of *University of Illinois Studies in Language and Literature*. University of Illinois, Urbana, Illinois.

Bloomfield, L. (1926). A Set of Postulates for the Science of Language. *Language*, 2(3):153–164.

Bloomfield, L. (1927). On Some Rules of Pānini. *Journal of the American Oriental Society*, 47:61–70.

Bloomfield, L. (1957). *Eastern Ojibwa: Grammatical Sketch, Texts and Word List*. University of Michigan Press, Ann Arbor.

Bluhme, H. (1970). The Phoneme System and Its Distribution in Roro. In Wurm, S. A. and Laycock, D. C., editors, *Pacific Linguistic Studies in Honor of Arthur Capell*, volume 13 of *Pacific Linguistics, Series C*, pages 867–877. Australian National University, Canberra.

Boas, F. (1911). Kwakiutl. In Boas, F., editor, *Handbook of American Indian Languages 1*, volume 40 of *Smithsonian Institution Bureau of American Ethnology Bulletin*, pages 423–558. Government Printing Office, Washington, D.C.

Boas, F. (1947). Kwakiutl Grammar, With a Glossary of the Suffixes. *Transactions of the American Philosophical Society*, 37:203–377.

Boas, F. and Deloria, E. (1941). *Dakota Grammar*, volume 23 of *Memoirs of the National Academy of Sciences*. U.S. Government Printing Office, Washington, D.C.

Bohtlingk, O. (1964). *Über Die Sprache Der Jakuten*, volume 35 of *Indiana University Publications, Uralic and Altaic Series*. Indiana University Press, Bloomington.

Bolima, F. A. (1998). The Tonological Outline of Ngishe. Master's thesis, University of Yaounde I.

Bolton, R. A. (1990). A Preliminary Description of Nuaulu Phonology and Grammar. Master's thesis, The University of Texas at Arlington.

Bommelyn, L. M. (1997). *The Prolegomena to the Tolowa Athabaskan Grammar*. PhD thesis, University of Oregon.

Bond, O. and Veselinova, L. (2011). Sampling Language Isolates. In *Paper Presented at the Association for Linguistic Typology 9, Hong Kong, Jul 21–24*.

Borg, A. (1973). The Segmental Phonemes of Maltese. *Linguistics*, 109:5–11.

Borg, I. and Groenen, P. J. F. (2005). *Modern Multidimensional Scaling: Theory and Applications*. Springer Verlag, Mannheim, Germany.

Borges, J. L. (1935). *Historia universal de la infamia*. Editorial Tor.

Borgman, D. M. (1990). Sanuma. In Derbyshire, D. C. and Pullum, G. K., editors, *Handbook of Amazonian Languages*, volume 2. Mouton de Gruyter.

Borgman, D. M. and Cue, S. L. (1963). Sentence and Clause Types in Central Waica (Shiriana). *International Journal of American Linguistics*, 29:222–229.

Borman, M. B. (1962). Cofan Phonemes. In Elson, B. F., editor, *Studies in Ecuadorian Indian languages 1*, volume 7 of *Linguistic Series*, pages 45–59. Summer Institute of Linguistics of the University of Oklahoma, Norman.

Bosteon, K. (2009). Shanjo and Fwe as Part of Bantu Botatwe: A Diachronic Phonological Approach. In Ojo and Moshi, editors, *Selected Proceedings of the 39th Annual Conference on African Linguistics*, pages 110–130. Cascadilla Proceedings Project.

Bostrom, P. (1998). Nominalizations and Relative Clauses in Tatuyo: A Prototype Approach. Master's thesis, The University of Texas at Arlington.

Bothorel, A. (1982). *Étude Phonétique et Phonologique du Breton Parlé à Argol (Finistere-Sud).* Atelier National Reproduction des Thèses, Université Lille III, Lille.

Botne, R. (2003). Lega (Beya Dialects) (D25). In Nurse, D. and Philippson, G., editors, *The Bantu languages*, pages 422–449. Routledge.

Bouny, P. (1977). Inventaire phonetique d'un parler Kotoko: le Mandagué de Mara. In Caprile, J.-P., editor, *Etudes Phonologiques Tschadiennes*, pages 59–77. Société d'Études linguistiques et anthropologiques de France, Paris.

Bouquiaux, L. (1970). *La Langue Birom (Nigeria Septentrional) Phonologie, Morphologie, Syntaxe.* Bibliothèque de la faculté de philosophie et lettres de l'université de Liége, Paris.

Bowden, J. (1997a). *Taba (Makian Dalam): Description of an Austronesian Language from Eastern Indonesia.* PhD thesis, University of Melbourne.

Bowden, J. (1997b). *Taba (Makian Dalam): Description of an Austronesian Language From Eastern Indonesia.* PhD thesis, University of Melbourne.

Bowden, J. and Hajek, J. (1996). Illustrations of the IPA: Taba. *Journal of the International Phonetic Association*, 26(1):55–57.

Boyd, V. L. (1997). A Phonology and Grammar of Mbódɔ̀mɔ̀. Master's thesis, The University of Texas at Arlington.

Boyeldieu, P. (1985). *La Langue Lua ('Niellim'). Groupe Boua - Moyen-Chari, Tchad.* Cambridge University Press, Cambridge.

Boyeldieu, P. (1987). *Les langues fer ('kara') et yulu du nord centrafricain: Esquisses Descriptives et Lexiques.* Paul Geuthner, Paris.

Boyeldieu, P. (2000). *La langue bagiro (République Centrafricaine).* Number 4 in Research in African Studies. Peter Lang GmbH.

Bradley, D. (1975). Nahsi and Proto-Burmese-Lolo. *Linguistics of the Tibeto-Burman Area*, 2(1):93–150.

Branks, T. and Branks, J. (1973). *Fonologia del Guambiano*, volume 2 of *Sistemas fonologicos de idiomas colombianos.* Summer Institute of Linguistcs, Loma Linda, Colombia.

Brauner, S. and Ashiwaju, M. G. (1965). *Lehrbuch der Hausa-Sprache*, volume X of *Lehrbücher für das Studium der Orientalischen und Afrikanischen Sprachen.* Max Hueben Verlag, Munich.

Bray, T., Paoli, J., and Sperberg-McQueen, C. M. (1998). Extensible Markup Language (XML) 1.0. Online: http://www.w3.org/TR/1998/REC-xml-19980210.

Bright, J. O. (1964). The Phonology of Smith River Athapaskan (Tolowa). *International Journal of American Linguistics*, 30(2):101–107.

Bright, W. (1957). *The Karok Language*, volume 13 of *University of California Publications in Linguistics.* University of California Press, Berkeley.

Bright, W. (1965). Luiseño Phonemics. *International Journal of American Linguistics*, 31(4):342–345.

Bright, W. (1968). *A Luiseño Dictionary.* University of California Press, Berkeley.

Broadbent, S. M. (1964). *The Southern Sierra Miwok Language*, volume 38 of *University of California Publications in Linguistics.* University of California Press, Berkeley.

Broadbent, S. M. and Pitkin, H. (1964). A Comparison of Miwok and Wintu. In Bright, W., editor, *Studies in Californian Linguistics*, pages 19–45. University of California Press, Berkeley and Los Angeles.

Broadwell, G. A. (2006). *Choctaw.* University of Nebraska Press.

Brockway, E. (1963). The Phonemes of North Puebla Nahuatl. *Anthropological Linguistics*, 5(3):14–18.

Bromley, H. M. (1961). *The Phonology of Lower Grand Valley Dani: A Comparative Structural Study of Skewed Phonemic Patterns*, volume 34 of *Verhandelingen van het Koninklijk Instituut voor Taal, Land en Volkenkunde.* Martinus Nijhoff, The Hague.

Brothers, C. (1905). *Aids to the Pronunciation of Irish.* M. H. Gill and Son, Dublin.

Browman, C. P. and Goldstein, L. M. (1986). Towards an Articulatory Phonology. *Phonology Yearbook*, 3:219–252.

Browman, C. P. and Goldstein, L. M. (1989). Articulatory Gestures as Phonological Units. *Phonology*, 6:201–251.

Browman, C. P. and Goldstein, L. M. (1992). Articulatory Phonology: An Overview. *Phonetica*, 49(3-4):155–180.

Brown, A. R. (1914). Notes on the Languages of the Andaman Islands. *Anthropos*, 9(1/2):36–52.

Brown, H. A. (1973). The Eleman Language Family. In Franklin, K., editor, *The Linguistic Situation in the Gulf District and Adjacent Areas (Papua New Guinea)*, volume 26 of *Pacific Linguistics, Series C*, pages 279–375. Australian National University, Canberra.

Brown, R. (1988). Waris Case System and Verb Classification. *Language and Linguistics in Melanesia*, 19:37–80.

Bruce, G. (1989). Report From the IPA Working Group on Suprasegmental Categories. In *Working Papers, Lund University, Department of Linguistics*, volume 35, pages 15–40. Lund University.

Bruce, L. (1984). *The Alamblak Language of Papua New Guinea (East Sepik)*, volume 81 of *Pacific Linguistics, Series C*. Australian National University, Canberra.

Bryant, M. G. (1999). Aspects of Tirmaga Grammar. Master's thesis, The University of Texas at Arlington.

Bubrix, D. V. (1949a). *Grammar of the Komi Literary Language*. Leningrad State University, Leningrad.

Bubrix, D. V. (1949b). *Grammatika literaturnogo komi jazyka*. Leningrad State University, Leningrad.

Burgess, E. and Ham, P. (1968). Multilevel Conditioning of Phoneme Variants in Apinaye. *Linguistics*, 41:5–18.

Burke, J. F. (1970). *The Irish of Tourmakeady, Co. Mayo*. The Dublin Institute for Advanced Studies, Dublin.

Burling, R. (1961). *A Garo Grammar*, volume 25 of *Deccan College Monograph Series*. Deccan College, Poona.

Burquest, D. A. (1971). *A Preliminary Study of Angas Phonology*. Institute of Linguistics, Zaria.

Butcher, A. and Tabain, M. (2004). On the Back of the Tongue: Dorsal Sounds in Australian Languages. *Phonetica*, 61:22–52.

Byarushengo, E. R. (1977). Preliminaries. In Byarushengo, E. R., Duranti, A., and Hyman, L. M., editors, *Haya Grammatical Structure: Phonology, Grammar, Discourse*, number 6 in Southern California Occasional Papers in Linguistics. Department of Linguistics: University of Southern California.

Cahill, M. C. (1999). *Aspects of the Morphology and Phonology of Kɔnni*. PhD thesis, The Ohio State University.

Callaghan, C. A. (1963). *A Grammar of the Lake Miwok Language*. PhD thesis, University of California at Berkley.

Calvanese, D., De Giacomo, G., Lenzerini, M., and Nardi, D. (2001). Reasoning in Expressive Description Logics. In *Handbook of Automated Reasoning*. Elsevier Science Publishers (North-Holland).

Camara, J. M. (1972). *The Portuguese Language (Translated by Anthony J. Naro*. University of Chicago Press, Chicago.

Cândido, G. V. (2004). *Descrição Morfossintática da Língua Shanenawa (Pano)*. PhD thesis, Instituto de Estudos da Linguagem.

Capell, A. (1967). Sound Systems in Australia. *Phonetica*, 16(2):85–110.

Capell, A. and Hinch, H. E. (1970). *Maung Grammar*, volume 98 of *Janua Linguarum, Series Practica*. Mouton de Gruyter, The Hague.

Capo, Haunkpati, B. C. (1991). *A Comparative Phonology of Gbe*. Foris, Garome, Bénin.

Carbonell, J. F. and Llisterri, J. (1992). Illustrations of the IPA: Catalan. *Journal of the International Phonetic Association*, 22(1–2):53–56.

Cardona, G. R. (1981). Profilo Fonologico Del Somalo. In Cardona, G. R. and Agostini, F., editors, *Fonologia e lessico*, volume 1 of *Studi Somali*, pages 3–26. Dipartimento per la Cooperazione allo Sviluppo; Comitato Tecnico Linguistico per l'Universita Nazionale Somala, Ministero degli Affari Esteri, Rome.

Cardoso, J. and Sheth, A. P., editors (2006). *Semantic Web Services, Processes and Applications*. Springer.

Carlson, B. F. (1972). *A Grammar of Spokan: A Salish Language of Eastern Washington*. PhD thesis, University of Hawaii.

Carlson, R. (1993). A Sketch of Jɔ: A Mande Language With a Feminine Pronoun. *Mandenkan*, 25:1–109.

Carnochan, J. (1948). A Study on the Phonology of an Igbo Speaker. *Bulletin of the School of Oriental and African Studies*, 12(2):416–27.

Carrington, J. F. (1977). Esquisse morphophonoloqie de la language Iikile. *Africana Linguistica*, 7:65–88.

Casimir, K. K. (1988). *Lexikon der Tagbana-Sprache*. PhD thesis, Universität Bielefeld.

Castren, M. A. (1966). *Grammatik der samojedischen Sprachen*, volume 53 of *Indiana University Publications, Uralic and Altaic Series*. Indiana University, Bloomington.

Cathcart, M. (1979). *Fonologia del Cacua. Sistemos fonologicos de idiomas colombianos*. Summer Institute of Linguistics, Loma Linda, Colombia.

Caudmont, J. (1954). Fonologia del Guambiano. *Revista Colombiana de Antropologia*, 5:189–206.

Caughley, R. C. (1982). *The Syntax and Morphology of the Verb in Chepang*. Number 84 in Pacific Linguistics. Australian National University.

Cauty, A. (1974a). Los sistemas fonologicos y silabicos de la lengua Panare. *Revista Colombiana de Antropologia*, 17:251–254.

Cauty, A. (1974b). Un criterio de decision sobre la presencia de la oclusiva glotal en el idioma Panare. *Revista Colombiana de Antropologia*, 17:255–258.

Cauty, A. (1978). An Approach to the Phonological and Syllabic Systems of the Panare Language. *Amerindia*, 3:85–103.

Central Intelligence Agency (2010). The World Factbook. Technical report, Central Intelligence Agency.

Cha, J.-K. (1995). Narrative Discourse Structure in Lhasa Tibetan. Master's thesis, The University of Texas at Arlington.

Chafe, W. L. (1967). *Seneca Morphology and Dictionary*. Smithsonian Press, Washington, D.C.

Chafe, W. L. (1976). *The Caddoan, Iroquoian, and Siouan Languages*. Mouton, The Hague / Paris.

Chai, N. M. (1971). *A Grammar of Aklan*. PhD thesis, University of Pennsylvania.

Chan, M. K. M. (1980). Zhong-shan Phonology. Master's thesis, University of British Columbia, Vancouver.

Chanard, C. (2006). Systèmes Alphabétiques Des Langues Africaines. Online: http://sumale.vjf.cnrs.fr/phono/.

Chao, Y.-R. (1947). *Cantonese Primer*. Harvard University Press, Cambridge, Massachusetts.

Chao, Y.-R. (1951). Taishan yuliao. *Bulletin of the Institute of History and Philology*, 23:25–76.

Chao, Y.-R. (1968). *A Grammar of Spoken Chinese*. University of California Press, Berkeley.

Chao, Y.-R. (1970). The Changchow Dialect. *Journal of the American Oriental Society*, 90:45–59.

Charachidzé, G. (1981). *Grammaire de la langue avar (langue du Caucase Nord-Est)*, volume 38 of *Document de linguistique quantitative*. Farvard, Paris.

Charney, J. O. (1993). *A Grammar of Comanche*. Studies in the Anthropology of North American Indians. University of Nebraska Press.

Chayen, M. J. (1973). *The Phonetics of Modern Hebrew*, volume 163 of *Janua Linguarum, Series Practica*. Mouton, The Hague.

Chelliah, S. L. (1992). *A Study of Manipuri Grammar.* PhD thesis, University of Texas at Austin.

Chen, P. P.-S. (1976). The Entity-Relationship Model–Toward a Unified View of Data. *ACM TODS*, 1(1):9–36.

Cheng, C.-C. (1973a). *A Synchronic Phonology of Mandarin Chinese.* Mouton, The Hague.

Cheng, T. M. (1973b). The Phonology of Taishan. *Journal of Chinese Linguistics*, 1:256–322.

Cho, S.-B. (1967). *A Phonological Study of Korean*, volume 2 of *Acta Universitatis Upsaliensis, Studia Uralica et Altaica Upsaliensia.* Almqvist and Wiksells, Uppsala.

Chomsky, N. (1964). *Current Issues in Linguistic Theory.* The Hague: Mouton.

Chomsky, N. and Halle, M. (1965). Some Controversial Questions in Phonological Theory. *Journal of Linguistics*, 1:97–138.

Chomsky, N. and Halle, M. (1968). *The Sound Pattern of English.* Harper & Row, New York, NY.

Clairis, C. (1977). Première approche du qawasqar: identification et phonologie. *La Linguistique*, 13:145–152.

Clark, L. E. (1995). *Vocabulario Popoluca De Sayula.* Serie de vocabularios y diccionarios indeígenas. Summer Institute of Linguistics, 104 edition.

Clayre, I. F. (1973). The Phonemes of Sa'ban: A Language of Highland Borneo. *Linguistics*, 100:26–46.

Cleal, A. M. (1973a). Gechode. In Kropp-Dakubu, M. E., editor, *West African Langauge Data Sheets*, volume 1, pages 253–259. West African Linguistics Society.

Cleal, A. M. (1973b). Genyanga. In Kropp-Dakubu, M. E., editor, *West African Langauge Data Sheets*, volume 1, pages 261–267. West African Linguistics Society.

Cleal, A. M. (1973c). Krache. In Kropp-Dakubu, M. E., editor, *West African Langauge Data Sheets*, volume 1, pages 366–373. West African Linguistics Society.

Clements, G. N. (1985). The Geometry of Phonological Features. *Phonology Yearbook*, 2:225–252.

Clements, G. N. (2003a). Feature Economy as a Phonological Universal. In *Proceedings of the 15th International Congress of Phonetic Sciences*.

Clements, G. N. (2003b). Feature Economy in Sound Systems. *Phonology*, 20:287–333.

Clements, G. N. (2009). The Role of Features in Phonological Inventories. In Raimy, E. and Cairns, C. E., editors, *Contemporary Views on Architecture and Representations in Phonology*, pages 19–68. MIT Press.

Clements, G. N. and Hume, E. V. (1995). The Internal Organization of Speech Sounds. In Goldsmith, J. A., editor, *The Handbook of Phonological Theory*, pages 245–306. Blackwell.

Clements, G. N., Michaud, A., and Patin, C. (2010). Do We Need Tone Features? In Goldsmith, J. A., Hume, E., and Wetzels, L., editors, *Tones and Features: Phonetic and Phonological Perspectives*. De Gruyter Mouton.

Coate, H. H. J. and Elkin, A. P. (1974). *Ngarinjin – English Dictionary*, volume 16 of *Oceania Linguistic Monograph*. University of Sydney, Sydney.

Coates, W. A. and de Silva, M. W. S. (1960). The Segmental Phonemes of Sinhalese. *University of Ceylon Review*, 18:163–175.

Codd, E. F. (1970). A Relational Model of Data for Large Shared Data Banks. *Communications of the ACM*, 13(6):377–387.

Coelho, G. M. (2003). *A Grammar of Betta Kurumba*. PhD thesis, The University of Texas at Austin.

Cohen, D. and Zafrani, H. (1968). *Grammaire de l'Hebreu Vivant*. Presses Universitaives de France, Paris.

Collier, K. and Collier, M. (1975). A Tentative Phonemic Statement of the Apoze Dialect, Kela Language. In Loving, R., editor, *Workers in Papua New Guinea Languages:*

Phonologies of Five Austronesian Languages, volume 13, pages 129–162. Summer Institute of Linguistics. No squib made due to pdf permissions.

Comrie, B., Haspelmath, M., and Bickel, B. (2003). The Leipzig Glossing Rules. Online: `http://www.eva.mpg.de/lingua/resources/glossing-rules.php`.

Connell, B., Ahoua, F., and Gibbon, D. (2002). Illustrations of the IPA: Ega. *Journal of the International Phonetic Association*, 32(1):99–104.

Conrad, R. J. (1978). Some Muhiang Grammatical Notes. In Loving, R., editor, *Workpapers in Papua New Guinea Linguistics*, volume 25. Summer Institute of Linguistics.

Cook, T. L. (1969a). Some Tentative Notes on the KoHumono Language. *Research Notes of the Deptartment of Linguistics and Nigerian Languages, University of Ibadan*, 2(3):1–49.

Cook, T. L. (1969b). *The Pronunciation of Efik for Speakers of English.* Indiana University Press, Bloomington.

Cook, W. H. (1979). *A Grammar of North Carolina Cherokee.* PhD thesis, Yale University, New Haven.

Cooke, J. R., Hudspith, J. E., and Morris, J. A. (1976). Phlong (Pwo Karen of Hot District, Chiang Mai). In Smalley, W. A., editor, *Phonemes and Orthography: Language Planning in Ten Minority Languages of Thailand*, volume 43 of *Pacific Linguistics, Series C*, pages 187–220. Australian National University, Canberra.

Costenoble, H. (1935). *Die Chamoro Sprache.* Nijhoff, The Hague.

Cottle, M. and Cottle, S. (1958). *The Significant Sounds of Ivatan*, volume 3 of *In Studies in Philippine linguistics; Oceania Linguistic Monographs.* University of Sydney, Sydney.

Coulmas, F. (1999). *The Blackwell Encyclopedia of Writing Systems.* Blackwell Publishers.

Coulmas, F. (2003). *Writing Systems: An Introduction to Their Analysis.* Cambridge University Press, Cambridge, UK.

Counts, D. R. (1969). *A Grammar of Kaliai Kove*. Oceanic Linguistics Special Publication, No. 6. University of Hawaii Press, Honolulu.

Cowan, H. K. J. (1965). *Grammar of the Sentani Language. With Specimen Texts and Vocabulary*, volume 47 of *Verhandelingen van het Koninklijk Instituut voot Taal-, Land- en Volkenkunde*. Martinus Nijhoff, The Hague.

Coward, D. F. (1990). An Introduction to the Grammar of Selaru. Master's thesis, The University of Texas at Arlington.

Cox, F. and Palethorpe, S. (2007). Illustrations of the IPA: English, Australian. *Journal of the International Phonetic Association*, 37(3):341–350.

Craig, C. G. (1977). *The Structure of Jacaltec*. University of Texas Press, Austin.

Crawford, J. C. (1963). *Totontepec Mixe Phonotagmemics*. Summer Institute of Linguistics, University of Oklahoma, Norman.

Crawford, J. M. (1973). Yuchi Phonology. *International Journal of American Linguistics*, 39(3):173–179.

Crazzolara, J. P. (1960). *A Study of the Logbara (Ma'di) Language*. Oxford University Press, London.

Creider, C. A. and Creider, J. T. (1989). *A Grammar of Nandi*, volume 4 of *Nilo-Saharan Linguistic Analyses and Documentation*. Helmut Buske Verlag, Hamburg.

Croft, W. (1990). *Typology and Universals*. Cambridge University Press.

Crothers, J. (1978). Typology and Universals of Vowel Systems in Phonology. In Greenberg, J. H., Ferguson, C. A., and Moravcsik, E. A., editors, *Universals of Human Language Volume 2: Phonology*, pages 93–152. Stanford University Press.

Crothers, J. H., Lorentz, J. P., Sherman, D. A., and Vihman, M. M. (1979). Handbook of Phonological Data From a Sample of the World's Languages: A Report of the Stanford Phonology Archive.

Crowell, T. H. (1979). *A Grammar of Bororo.* PhD thesis, Cornell University.

Crumrine, L. S. (1961). *The Phonology of Arizona Yaqui.* Number 5 in Anthropological Papers of the University of Arizona. University of Arizona, Tucson.

Cubar, E. H. and Cubar, N. I. (1994). *Writing Filipino Grammar: Traditions & Trends.* New Day.

Cunningham, M. C. (1969). *A Description of the Yugumbir Dialect of Bandjalang*, volume 1 of *University of Queensland Faculty of Arts Papers.* University of Queensland Press, Brisbane.

Curnow, T. J. (1997). *A Grammar of Awa Pit (Cuaiquer): An Indigenous Language of South-western Colombia.* PhD thesis, The Austrailian National University.

Cysouw, M. (2003). Against Implicational Universals. *Linguis*, 7(1):89–100.

Cysouw, M. (2005). Quantitative Methods in Typology. In Altmann, G., Köhler, R., and Piotrowski, R. G., editors, *Quantitative Linguistics: An International Handbook*, pages 554–578. Berlin: Walter de Gruyter.

Cysouw, M. (2010). On the Probability Distribution of Typological Frequencies. In Ebert, C., Jäger, G., and Michaelis, J., editors, *The Mathematics of Language*, pages 29–35. Berlin: Springer.

Cysouw, M., Dediu, D., and Moran, S. (2012). Still No Evidence for an Ancient Language Expansion From Africa. *Science*, 335:657–b.

da Silva Tavares, P. (2005). *A Grammar of Wayana.* PhD thesis, Rice University.

Dahl, O. C. (1952). Étude de phonologie et de phonétique malgache. *Norsk Tidsskrift for Sprogvidenskap*, 16:148–200.

Dalby, D. (1966). Lexical Analysis in Temne With an Illustrative Wordlist. *Journal of West African Languages*, 3(2):5–26.

Daniel, R. and Shaw, K. A. (1977). Samo Phonemes. In Loving, R., editor, *Workpapers in Papua New Guinea Languages*, volume 19, pages 97–135. Summer Institute of Linguistics.

Daniels, P. T. (1990). Fundamentals of Grammatology. *Journal of the American Oriental Society*, 110(4):727–731.

Daniels, P. T. (1996). The Study of Writing Systems. In Daniels, P. T. and Bright, W., editors, *The World's Writing Systems*. Oxford University Press, New York, NY.

Daniels, P. T. and Bright, W. (1996). *The World's Writing Systems*. Oxford University Press, New York, NY.

Dankovičová, J. (1997). Illustrations of the IPA: Czech. *Journal of the International Phonetic Association*, 27(1-2):77–80.

Das, A. R. (1977). *A Study of the Nicobarese Language*. Anthropological Survey of India, Government of India, Calcutta.

Dasgupta, D. and Sharma, S. R. (1982). *A Hand Book of the Onge Language*. Linguistic Series. Anthropological Survey of India, 5 edition.

Davidson, M. (2002). *Studies in Southern Wakashan (Nootkan) Grammar*. PhD thesis, State University of New York at Buffalo.

Davis, D. R. (1969). The Distinctive Features of Wantoat Phonemes. *Linguistics*, 47:5–17.

Davis, K. (2003). *A Grammar of the Hoava Language: Western Solomons*. Pacific Linguistics.

Davis, M. M. (1974). The Dialects of the Roro Language of Papua: A Preliminary Survey. *Kivung*, 7:3–22.

Day, C. (1973). *The Jacaltec Language*, volume 12 of *Indiana University Publications, Language Science Monographs*. Indiana University Press, Bloomington.

Dayley, J. P. (1985). *Tzutujil Grammar*, volume 107 of *University of California Publications in Linguistics*. University of California Press.

De Armond, R. C. (1975). Some Rules of Brahui Conjugation. In Schiffman, H. G. and Eastman, C. M., editors, *Dravidian Phonological Systems*, pages 242–299. University of Washington Press, Seattle.

de Groot, A. W. (1931). Phonologie und Phonetik als funktionswissenschaften. *Travaux du Cercle Linguistique de Prague*, 4:114–147.

de Oliveira Borges E Souza, P. (2004). Estudos de aspectos de lingua kaiabi (Tupi). Master's thesis, Universidade Estadual de Campinas.

de Voogt, A. (2009). A Sketch of Afitti Phonology. *Studies in African Linguistics*, 38(1):35–52.

Décsy, G. (1966). *Yurak Chrestomathy*, volume 50 of *Indiana University Publications, Uralic and Altaic Series*. Indiana University Press, Bloomington.

Dediu, D. and Ladd, D. R. (2007). Linguistic Tone Is Related to the Population Frequency of the Adaptive Haplogroups of Two Brain Size Genes, ASPM and Microcephalin. *PNAS*, 104(26):10944–10949.

Dell, F. (1981). *La langue bai: Phonologie et Lexique*, volume 2 of *Etudes Linguistiques*. Editions de l'Ecole des Hautes Etudes en Sciences Sociales, Paris.

Dempwolff, O. (1916). *Die Sandawe: Linguistisches und ethnographisches Material aus Deutsch-Ostafrika*, volume 34 of *Abhandlungen des Hamburgischen Kolonialinstituts*. L. Friederichsen & Co., Hamburg.

Demuth, K. (1992). Acquisition of Sesotho. In Slobin, D., editor, *The Cross-Linguistic Study of Language Acquisition*, volume 3, pages 557–638. Lawrence Erlbaum Associates.

Derbyshire, D. C. (1985). *Hixkaryana and Linguistic Typology*. Summer Institute of Linguistics, Dallas.

Desheriev, J. D. (1953). *Batsbijskij jazyk: fonetika, morfologija, sintaksis, leksika*. Izdatel'stvo Akademii nauk SSSR, Moskva.

Deutscher, G. (2009). Overall Complexity – a Wild Goose Chase? In Sampson, G., Gil, D., and Trudgill, P., editors, *Language Complexity as an Evolving Variable*, pages 243–251. Oxford University Press.

Di Luzio, A. (1972). Preliminary Description of the Amo Language. *Afrika und Übersee*, 56:3–61.

Diagana, O. M. (1995). *La langue soninkée: Morphosyntax et sens.* Editions L'Harmattan, Paris.

Dickinson, C. (2002). *Complex Predicates in Tsafiki.* PhD thesis, University of Oregon.

Diffloth, G. (1980). The Wa Languages. *Linguistics of the Tibeto-Burman Area*, 5(2):1–182.

Diffloth, G. (1984). *The Dvaravati Old Mon Language and Nyah Kur.* Chulalongkorn University, Bangkok.

Dillon, J. A. (1994). A Grammatical Description of Tatana'. Master's thesis, The University of Texas at Arlington.

Dirks, S. (1953). Campa (Arawak) Phonemes. *International Journal of American Linguistics*, 19:302–304.

Disner, S. (1983). *Vowel Quality: The Relation Between Universal and Language-specific Factors.* PhD thesis, UCLA.

Dixon, R. M. W. (1966a). Mbabaram: A Dying Australian Language. *Bulletin of the School of Oriental and African Languages (London)*, 29:97–121.

Dixon, R. M. W. (1966b). *Mbarabam Phonology.* Stephen Austin and Sons, Ltd., Hertford.

Dixon, R. M. W. (1972). *The Dyirbal Language of North Queensland*, volume 9 of *Cambridge Studies in Linguistics.* Cambridge University Press, Cambridge.

Dixon, R. M. W. (1977). *A Grammar of Yidin*, volume 19 of *Cambridge Studies in Linguistics.* Cambridge University Press, Cambridge.

Dixon, R. M. W. (1988). *A Grammar of Boumaa Fijian.* University of Chicago Press, Chicago.

Dixon, R. M. W. (1997). *The Rise and Fall of Languages.* Cambridge University Press, Cambridge, UK.

Dixon, R. M. W. (2009a). *Basic Linguistic Theory: Grammatical Topics*, volume 2. Oxford University Press.

Dixon, R. M. W. (2009b). *Basic Linguistic Theory: Methodology*, volume 1. Oxford University Press.

Djawanai, S. (1983). *Ngadha Text Tradition: The Collective Mind of the Ngadha People, Flores*, volume 55 of *Pacific Linguistics Series D.* Department of Linguistics, Research School of Pacific Studies, Australian National University.

Djiafeua, P. (1989). Esquisse phonologique du mpumpuŋ. Master's thesis, Universite de Yaounde.

Doak, I. G. (1997). *Coeur D'Alene Grammatical Relations.* PhD thesis, The University of Texas at Austin.

Doble, M. (1962). Essays on Kapauku Grammar. *Nieuw Guinea Studien*, 6:152–155.

Doble, M. (1987). A Description of Some Features of Ekari Language Structure. *Oceanic Linguistics*, 26:55–113.

Doke, C. M. (1926). *The Phonetics of the Zulu Language.* Bantu Studies. Wiwatersrand University Press, Johannesburg.

Doke, C. M. (1961). *Textbook of Zulu Grammar.* Longmans, Cape Town. 6th edition (1st edition 1927).

Dol, P. H. (1999). *A Grammar of Maybrat: A Language of the Bird's Head, Irian Jaya, Indonesia.* PhD thesis, Universiteit Leiden.

Dolphyne, F. A. (1971). Classification of Akan Verb Stems. In Houis, M., editor, *In Actes du 8e Congrès de la Société Linguistique d'Afrique Occidentale*. Annales de l'Université d'Abidjan, Série H, 1, Abidjan.

Dolphyne, F. A. (1988a). *The Akan (Twi-Fante) Language: Its Sound Systems and Tonal Structure*. Ghana Universities Press, Accra, Ghana.

Dolphyne, F. A. (1988b). *The Akan (Twi-Fante) Language: Its Sound Systems and Tonal Structure*. Ghana Universities Press, Accra.

Donaldson, T. (1980). *Ngiyambaa: The Language of the Wangaaybuwan*, volume 29 of *Cambridge Studies in Linguistics*. Cambridge University Press, Cambridge.

Donohue, M. (1994). Illustrations of the IPA: Tukang Besi. *Journal of the International Phonetic Association*, 24(1):39–41.

Donohue, M. (2004). A Grammar of the Skou Language of New Guinea.

Donohue, M. and Roque, L. S. (2002). *I'saka: A Sketch Grammar of a Language of North-Central New Guinea*. Canberra: Pacific Linguistics.

Donwa, S. O. (1982). *The Sound System of Isoko*. PhD thesis, University of Ibadan.

Douglas, W. H. (1955). Phonology of the Australian Aboriginal Language Spoken at Ooldea, South Australia, 1951-1952. *Oceania*, 25:216–229.

Douglas, W. H. (1964). *An Introduction to the Western Desert Language*, volume 4 of *Oceania Linguistics Monographs*. The University of Sydney, Australia, Sydney.

Dow, F. D. M. (1972). *An Outline of Mandarin Phonetics*. Faculty of Asian Studies, Australian National University, Canberra.

Drame, M. (1981). *Aspects of Mandingo Grammar*. PhD thesis, University of Illinois at Urbana-Champaign.

Dryer, M. S. (1989). Large Linguistic Areas and Language Sampling. *Studies in Language*, 13:257–292.

Dryer, M. S. (1991). SVO Languages and the OV: VO Typology. *Journal of Linguistics*, 27:443–482.

Dryer, M. S. (1992). The Greenbergian Word Order Correlations. *Language*, 68:81–138.

Dryer, M. S. (2000). Counting Genera vs. Counting Languages. *Linguistic Typology*, 4(3):123–145.

Dryer, M. S. (2003). Significant and Non-significant Implicational Universals. *Linguistic Typology*, 7(1):108–127.

Dryer, M. S. (2006). Descriptive Theories, Explanatory Theories, and Basic Linguistic Theory. In Ameka, F. K., Dench, A. C., and Evans, N., editors, *Catching Language: The Standing Challenge of Grammar Writing*. Mouton de Gruyter, Berlin.

Ducos, G. (1974). Pajade. In Kropp-Dakubu, M. E., editor, *West African Langauge Data Sheets*, volume 2, pages 193–202. West African Linguistics Society.

Dul'zon, A. P. (1968). *Ketskij jazyk*. Tomsk University, Tomsk.

Dumestre, G. and Duponchel, L. (1971). Note sur les groupes consonantiques en ebrie et en alladian. *Annales de l'Universite d'Abidjan, serie H*, 3(1):31–46.

Dunham, M. (2005). *Éléments de description du langi: Langue bantu F.33 de tanzanie*. Peeters Publishers.

Dunn, J. A. (1978). *A Practical Dictionary of the Coast Tsimshian Language*, volume 42 of *National Museum of Man Mercury Series, Canadian Ethnology Service*. National Museums of Canada, Ottawa.

Dunn, J. A. (1979). *A Reference Grammar for the Coast Tsimshian Language*, volume 55 of *Canadian Ethnology Service Paper*. National Museums of Canada, Ottawa.

Dunn, M., Greenhill, S. J., Levinson, S. C., and Gray, R. D. (2011). Evolved Structure of Language Shows Lineage-specific Trends in Word-order Universals. *Nature*.

Dunstan, E. (1964). Towards a Phonlogy of Ngwe. *Journal of West African Languages*, 1(1):39–42.

Duponchel, L. (1971). Contacts de cultures et création lexicale en alladian. *Annales de l'Universite d'Abidjan, serie H*, 3(1):47–70.

Durbin, M. and Seijas, H. (1972). The Phonological Structure of the Western Carib Languages of the Sierra De Perija, Venezuela. *Atti del XL Congresso Internazionale degli Americanisti*, 3:69–77.

Dutton, T. E. (1969). *The Peopling of Central Papua*, volume 9 of *Pacific Linguistics, Series B*. Australian National University, Canberra.

Dyen, I. (1971). Malagasy. In Sebeok, T. A., editor, *Linguistics in Oceania*, volume 8 of *Current Trends in Linguistics*, pages 211–239. Mouton, The Hague.

Dzhejranishvili, E. F. (1967). Rutul'skij jazyk. In Bokarev, E. A. and Lomtatidze, K. V., editors, *Jazyki narodov SSSR. Volume 4: Iberijsko-kavkazskie jazyki*, pages 580–590. Akademija Nauk, Moscow.

Eades, D. and Hajek, J. (2006). Illustrations of the IPA: Gayo. *Journal of the International Phonetic Association*, 36(1):107–115.

Ebah Ebude, L. (1990). The Noun Class System of Lefɔ. Master's thesis, University of Yaounde.

Ebert, K. H. (1976). *Sprache Und Tradition Der Kera (Tschad). Volume 2: Lexicon.* Dietrich Reimer, Berlin.

Ebert, K. H. (1979). *Sprache Und Tradition Der Kera (Tschad). Volume 3: Grammatik.* Reimer, Berlin.

Echeverría, M. S. and Contreras, H. (1965). Araucanian Phonemics. *International Journal of American Linguistics*, 31(2):132–135.

Edika, E. S. F. (1990). Esquisse phonologique du bakako. Master's thesis, Universite de Yaounde.

Edmonds, B. (1999). *Syntactic Measures of Complexity.* PhD thesis, University of Manchester.

Edmonson, B. W. (1988). *A Descriptive Grammar of Huastec (Potosino Dialect).* PhD thesis, Tulane University.

Edwards, W. F. (1978). Some Synchronic and Diachronic Aspects of Akawaio Phonology. *Anthropological Linguistics*, 20:77–84.

Efimov, V. A. (1986). *Iazyk Ormuri: v sinkhronnom i istoricheskom osveshchenii.* Nauka, Moscow.

Egerod, S. (1966). A Statement on Atayal Phonology. *Artibus Asiae*, 23:120–130.

Ehrman, M. E. (1972). *Contemporary Cambodian: Grammatical Sketch.* Foreign Service Institute, US Department of State, Washington, D.C.

Einarsson, S. (1949). *Icelandic.* John Hopkins Press, Baltimore.

Einaudi, P. F. (1974). *A Grammar of Biloxi.* PhD thesis, University of Colorado.

Ekambi, A. (1990). Esquisse phonologique du nulibie. Master's thesis, Universite de Yaounde.

Eko, J. E. (1974). Description phonologique du faŋ. Master's thesis, Universite de Yaounde.

Elbert, S. H. and Pukui, M. K. (1979). *Hawaiian Grammar.* University of Hawaii Press, Honolulu.

Elderkin, E. D. (1982). Tanzanian and Ugandan Isolates. In Vossen, R. and Bechhaus-Gerst, M., editors, *Nilotic Studies. Proceedings of the International Symposium on Languages and History of the Nilotic Peoples*, volume 10 of *Kölner Beiträge zur Afrikanistik*, pages 499–521. Universität zu Köln, Institut für Afrikanistik, Köln.

Elderkin, E. D. (2003). Herero (R31). In Nurse, D. and Philippson, G., editors, *The Bantu languages*, pages 581–608. Routledge.

Elfitoury, A. A. (1976). *A Descriptive Grammar of Libyan Arabic.* PhD thesis, Georgetown University.

Elimelech, B. (1976). *A Tonal Grammar of Etsako.* PhD thesis, University of California at Los Angeles.

Ember, C. R. and Ember, M. (2007). Climate, Econiche, and Sexuality: Influences on Sonority in Language. *American Anthropologist*, 109(1):180–185.

Ember, M. and Ember, C. R. (1999). Cross-Language Predictors of Consonant-Vowel Syllables. *American Anthropologist*, 101(4):730–742.

Emeneau, M. B. (1937). Phonetic Observations on the Brahui Language. *Bulletin of the School of Oriental Studies*, 8:981–983.

Emeneau, M. B. (1944). *Kota Texts 1*, volume 2 of *University of California Publications in Linguistics.* University of California Press, Berkeley / Los Angeles.

Emeneau, M. B. (1962). *Brahui and Dravidian Comparative Grammar*, volume 27 of *University of California Publications in Linguistics.* University of California Press, Berkeley.

Endresen, R. T. (1991). Diachronic Aspects of the Phonology of Nizaa. *Journal of African Languages and Linguistics*, 12(2):171–194.

Engstrand, O. (1990). Illustrations of the IPA: Swedish. *Journal of the International Phonetic Association*, 20(1):41–42.

Epps, P. (2005). *A Grammar of Hup.* PhD thesis, University of Virginia.

Evans, N., Besold, J., Stoakes, H., and Lee, A. (2005). *Materials on Golin: Grammar, Texts and Dictionary.* Department of Linguistics and Applied Linguistics, University of Melbourne.

Everest, G. C. (1986). *Database Management.* McGraw-Hill, Inc.

Everett, C. (2006). *Patterns in Karitiana: Articulation, Perception, and Grammar.* PhD thesis, Rice University.

Everett, D. and Kern, B. (1997). *Wari'*. Routledge.

Everett, D. L. (1982). Phonetic Rarities in Piraha. *Journal of the International Phonetic Association*, 12(2):94–9.

Exter, M. (2003). Phonetik und Phonologie des Wogeo. Master's thesis, Institut für Sprachwissenschaft Universität zu Köln.

Facundes, S. d. S. (2000). *The Language of the Apurinã People of Brazil (Maipure/Arawak)*. PhD thesis, State University of New York at Buffalo.

Fagan, J. L. (1988). Javanese Intervocalic Stop Phonemes: The Light/Heavy Distinction. In McGinn, R., editor, *Studies in Austronesian Linguistics*, pages 173–200. Ohio University Center for International Studies, Center for Southeast Asia Studies, Athens, Ohio.

Fallon, P. D. (2006). Consonant Mutation and Reduplication in Blin Singular and Plurals. In et al., M., editor, *Selected Proceedings of the 35th Annual Conference on African Linguistics*, pages 114–124. Cascadilla Proceedings Project.

Fallon, P. D. (2009). The Velar Ejective in Proto-Agaw. In Ojo and Moshi, editors, *Selected Proceedings of the 39th Annual Conference on African Linguistics*, pages 10–22. Cascadilla Proceedings Project.

Faraclas, N. (1984). *A Grammar of Obolo*. Studies in African Grammatical Systems. Indiana University Linguistics Club.

Faraclas, N. G. (1989). *A Grammar of Nigerian Pidgin*. PhD thesis, University of California, Berkeley.

Farnetani, E. (1981). Dai tratti ai parametri: introduzione all'analisi strumentale della lingua somala. In Cardona, G. R. and Agostini, F., editors, *Studi Somali 1: Fonologia e lessico*, pages 27–108. Dipartimento per la Cooperazione allo Sviluppo; Comitato Tecnico Linguistico per l'Universita Nazionale Somala, Ministero degli Affari Esteri, Rome.

Farquhar, B. B. (1974). *A Grammar of Antiguan Creole*. PhD thesis, Cornell University.

Farr, C. J. M., Furoke, B. T., and Farr, J. B. (1996). Tafota Baruga Grammar Notes. Unpublished Manuscript.

Farr, J. and Farr, C. (1974). A Preliminary Korafe Phonology. In Healey, A., editor, *Workpapers in Papua New Guinea Languages: Three Studies in Languages of Eastern Papua*, volume 3, pages 5–39. Summer Institute of Linguistics.

Farrar, S. (2003). *An Ontology for Linguistics on the Semantic Web*. PhD thesis, University of Arizona.

Farrar, S. and Langendoen, D. T. (2003). A Linguistic Ontology for the Semantic Web. *GLOT*, 7(3):97–100.

Farrar, S. and Langendoen, D. T. (2010). An OWL-DL Implementation of GOLD: An Ontology for the Semantic Web. In Witt, A. and Metzing, D., editors, *Linguistic modeling of information and Markup Languages. Contributions to language technology*, number 40 in Text, Speech and Language Technology. Springer, Dordrecht.

Farrar, S. and Lewis, W. D. (2005). The GOLD Community of Practice: An Infrastructure for Linguistic Data on the Web. In *In Proceedings of the EMELD 2005 Workshop on Digital Language Documentation: Linguistic Ontologies and Data Categories for Language Resources*.

Fast, P. W. (1953). Amuesha (Arawak) Phonemes. *International Journal of American Linguistics*, 19(3):191–194.

Fedry, J. (1977). Apercu sur la phonologie et la tonologie de quatre langues du groupe Mubi-Karbo (Guera,Dangaleat-est, Dangaleat-ouest, Bidiyo, Dyongor). In Caprile, J., editor, *Etudes Phonologiques Chadiennes*, pages 87–112. Société d'Études Linguistiques et Anthropologiques de France, Paris.

Feldman, H. (1978). Some Notes on Tongan Phonology. *Oceanic Linguistics*, 17(2):133–139.

Feldman, H. (1986). *A Grammar of Awtuw*. Number 94 in Pacific Linguistics. The Australian National University.

Feldpausch, T. and Feldpausch, B. (1992). *Namia Grammar Essentials.* Data Papers on Papua New Guinea Languages. Summer Institute of Linguistics.

Ferguson, C. A. and Chowdhury, M. (1960). The Phonemes of Bengali. *Language*, 36:22–59.

Fernandez, F. (1968). *A Grammatical Sketch of Remo: A Munda Language.* PhD thesis, The University of North Carolina at Chapel Hill.

Ferrel, R. (1982). *Paiwan Dictionary.* Department of Linguistics and Research School of Pacific Studies, Australian National University, Canberra.

Fiensong S. Chia, A. (1993). Phonology of Bubia. Master's thesis, University of Yaounde I.

Filchenko, A. Y. (2007). *A Grammar of Eastern Khanty.* PhD thesis, Rice University.

Firchow, I. and Firchow, J. (1969a). An Abbreviated Phoneme Inventory. *Anthropological Linguistics*, 11:271–276.

Firchow, I. and Firchow, J. (1969b). An Abbreviated Phoneme Inventory. *Anthropological Linguistics*, 11:271–276.

Firth, J. R. (1957). *Papers in Linguistics 1934–51.* Oxford University Press, Oxford, UK.

Fleck, D. W. (2003). *A Grammar of Matses.* PhD thesis, Rice University.

Fleisch, A. (2000). *Lucazi Grammar: A Morphosemantic Analysis.* Number 15 in Grammatische Analysen Afrikanischer Sprachen. Rüdiger Köppe Verlag.

Fleming, H. C. (1976). Kefa (Gonga) Languages. In Bender, M. L., editor, *The Non-Semitic Languages of Ethiopia*, pages 351–376. African Studies Center, Michigan State University, East Lansing.

Fleming, I. and Dennis, R. K. (1977). Tol (Jicaque) Phonology. *International Journal of American Linguistics*, 43:121–127.

Flemming, E. (2004). Contrast and Perceptual Distinctiveness. In Hayes, B., Kirchner, R., and Steriade, D., editors, *Phonetically Based Phonology*, pages 232–276. Cambridge University Press.

Fointein, J. N. (1986). *Phonology of Esimbi.* PhD thesis, University of Yaounde.

Foreman, V. and Marten, H. (1973). Yessan-Mayo Phonemes. In Healey, A., editor, *Phonologies of Three Languages of Papua New Guinea*, volume 2 of *Workpapers in Papua New Guinea Languages*, pages 79–108. Summer Institute of Linguistics, Ukarumpa.

Forges, G. (1983). *Phonologie et morphologie du kwezo*, volume 113. Musée Royal de l'Afrique Centrale (MRAC), Tervuren.

Foris, D. P. (1993). *A Grammar of Sochiapan Chinantec.* PhD thesis, University of Auckland.

Forku, D. T. (2000). A Sketch of the Phonology of Mamenyan and Standardization Perspectives. Master's thesis, University of Yaounde I.

Fortune, G. (1955). *An Analytical Grammar of Shona.* Stephen Austin and Sons.

Fortune, R. F. (1942). *Arapesh.* Number 19 in Publications of the American Ethnological Society. J.J. Augustin.

Foster, M. L. (1969). *The Tarascan Language*, volume 56 of *University of California Publications in Linguistics.* University of California Press, Berkeley.

Fought, J. G., Munroe, R. L., Fought, C. R., and Good, E. M. (2004). Sonority and Climate in a World Sample of Languages: Findings and Prospects. *Cross-Cultural Research*, 38:27–51.

Fozoh, A. V. M. (2002). The Phonology of Círàmbɔ́. Master's thesis, University of Yaounde I.

Frajzyngier, Z. (2001). *A Grammar of Lele.* CSLI.

Francois, A. (2001). *Contraintes de Structures et Liberte dans l'Organisation du Discours: Une Description de Mwotlap, Langue Oceanienne du Vanuatu.* PhD thesis, Universite Paris-IV Sorbonne.

Frank, W. J. (1999). Nuer Noun Morphology. Master's thesis, State University of New York at Buffalo.

Franklin, K. and Franklin, J. (1962). Kewa 1: Phonological Asymmetry. *Anthropological Linguistics*, 4(7):29–37.

Franks, P. S. (1985). *A Grammar of Ika.* PhD thesis, The University of Pennsylvania.

Frantz, C. I. and Frantz, M. E. (1966). Gadsup Phoneme and Toneme Units. In *Papers in New Guinea Linguistics 5*, volume 7 of *Pacific Linguistics, Series A*, pages 1–11. Australian National University, Canberra.

Freeland, L. S. (1951). *Language of the Sierra Miwok*, volume 6 of *Indiana University Publications in Anthropology and Linguistics, Memoir.* Indiana University.

Freeze, R. A. (1975). *A Fragment of an Early Kekchi Vocabulary*, volume 2 of *University of Missouri Monographs in Anthropology.* University of Missouri, Museum of Anthropology, Columbia Missouri.

Freudenburg, A. and Freudenburg, M. (1974). Boikin Phonemes. In Loving, R., editor, *Phonologies of four Austronesian Languages*, Workpapers in Papua New Guinea Languages, pages 97–128. The Summer Institute of Linguistics.

Frick, Esther, J. (1973). Dghwede. In Kropp-Dakubu, M. E., editor, *West African Langauge Data Sheets*, volume 1, pages 135–142. West African Linguistics Society.

Friedman, V. A. (2002). Macedonian. In *Languages of the World/Materials 117.* LinCom Europa.

Friedrich, P. (1975). *A Phonology of Tarascan*, volume 4 of *Series in social, cultural and linguistic anthropology.* Department of Anthropology, University of Chicago, Chicago.

Frisch, S. (1997). *Similarity and Frequency in Phonology.* PhD thesis, Northwestern University.

Fromkin, V. A. (1977). *The Phonology of Akan Revisited.* Hornbeam Press, Inc.

Fudge, E. (1975). English Word Stress: An Examination of Some Basic Assumptions. In Goyvaerts, D. L. and Pullum, G. K., editors, *Essays on the Sound Pattern of English*, pages 277–323. E. Story-Scientia P.V.B.A., Ghent.

Furby, C. E. (1974). Garawa Phonology. In *Papers in Australian Linguistics 7*, volume 37 of *Pacific Linguistics, Series A*, pages 1–11. Australian National University, Canberra.

Gabas, N. J. (1999). *A Grammar of Karo, Tupi (Brazil)*. PhD thesis, The University of California at Santa Barbara.

Gajendragadkar, S. N. (1970). *East Indian Fisher Folk*, volume 1 of *Dialects in the Bombay Area*. University of Bombay, Bombay.

Gallagher, S. and Baehr, P. (2005). Bariai Grammar Sketch. In van den Berg, R., editor, *Data Papers on Papua New Guinea Languages*, volume 49.

Galloway, B. D. (1977). *A Grammar of Chilliwack Halkomelem*. PhD thesis, University of California, Berkeley.

Galucio, A. V. (2001). *The Morphosyntax of Mekens (Tupi)*. PhD thesis, University of Chicago.

Ganong, T. W. (1998). Features of Baga Morphology, Syntax, and Narrative Discourse. Master's thesis, The University of Texas at Arlington.

Garbell, I. (1965). *The Jewish Neo-Aramaic Dialect of Persian Azerbayan*. Mouton, The Hague.

Gardner, I. (1966). Abua. In Kropp-Dakubu, M. E., editor, *West African Langauge Data Sheets*, volume 1, pages 22–26. West African Linguistics Society.

Garvin, P. L. (1950). Wichita I: Phonemics. *International Journal of American Linguistics*, 16(4):179–184.

Gavel, H. (1929). *Grammaire Basque*. Courrier, Bayonne.

Gedney, W. J. (1965). Yay, a Northern Tai Language in North Vietnam. *Lingua*, 14:180–193.

Gelman, A. and Hill, J. (2007). *Data Analysis Using Regression and Multilevel/hierarchical Models*. Cambridge University Press, New York, NY.

Gerdel, F. (1973). Fonemas del paez. In Waterhouse, V., editor, *Sistemas fonológicas de idiomas colombianos 2*, pages 7–37. Summer Institute of Linguistics, Lomalinda, Colombia.

Gerzenstein, A. (1968). *Fonologia de la lengua Gununa-kena*, volume 5 of *Cuadernos de Linguistica Indigen.* Centro de Estudios Linguisticos, University of Buenos Aires, Buenos Aires.

Gill, H. S. and Gleason, H. A. (1969). *A Reference Grammar of Punjabi.* Number 3 in Hartford Studies in Linguistics. Hartford Seminary Foundation, Hartford.

Gimba, A. M. (2000). *Bole Verb Morphology.* PhD thesis, The University of California at Los Angeles.

Gimson, A. C. (1962). *An Introduction to the Pronunciation of English.* Edward Arnold, London.

Gisele, A. (1994). *Phonologie Structurale de l'Awing.* PhD thesis, Universite de Yaounde I.

Glasgow, D. and Glasgow, K. (1967). The Phonemes of Burera. In *Papers in Australian Linguistics 1*, volume 10 of *Pacific Linguistics, Series A*, pages 1–14. Australian National University, Canberra.

Godfrey, T. J. (1981). *Grammatical Categories for Spatial Reference in the Western Mam Dialect of Tacaná.* PhD thesis, University of Texas at Austin.

Goldsmith, J. A. (1976). *Autosegmental Phonology.* PhD thesis, Massachusetts Institute of Technology.

Goldsmith, J. A. (1990). *Autosegmental and Metrical Phonology.* Blackwell, Oxford, UK.

Goldsmith, J. A. (1995). On Information Theory, Entropy, and Phonology in the 20th Century. *Folia Linguistica*, 34(1–2):1–17.

Goldsmith, J. A. and Laks, B. (To appear). Generative Phonology: Its Origins, Its Principles, and Its Successors. In Waugh, L., Joseph, J. E., and Monville-Burston, M., editors, *The Cambridge History of Linguistics.* Cambridge University Press.

Goldsmith, J. A. and Riggle, J. (2012). Information Theoretic Approaches to Phonological Structure: The Case of Finnish Vowel Harmony. *Natural Language and Linguistic Theory*, 30:859–896.

Golla, V. K. (1970). *Hupa Grammar*. PhD thesis, University of California at Berkeley.

Gómez, G. G. (1990). *The Shiriana Dialect of Yanam (Northern Brazil)*. PhD thesis, Columbia University.

González, H. A. (2005). *A Grammar of Tapiete (Tupi-Guarani)*. PhD thesis, University of Pittsburgh.

Good, P. I. (2006). *Resampling Methods: A Practical Guide to Data Analysis*. Birkhauser, Boston, MA, 3rd edition.

Gordon, M., Munro, P., and Ladefoged, P. (2000). Some Phonetic Structures of Chickasaw. *Anthropological Linguistics*, 42(3):366–400.

Gordon, R. G., editor (2005). *Ethnologue: Languages of the World, Fifteenth Edition*. Summer Institute of Linguistics, Dallas, TX.

Goudswaard, N. E. (2005). *The Begak (Ida'an) Language of Sabah*. LOT.

Gower, J. C. (1966). Some Distance Properties of Latent Root and Vector Methods Used in Multivariate Analysis. *Biometrika*, 53:325–328.

Gowlett, D. (2003). Zone S. In Nurse, D. and Philippson, G., editors, *The Bantu Languages*, pages 609–638. Routledge.

Granadillo, T. (2006). *An Ethnographic Account of Language Documentation Among the Kurripako of Venezuela*. PhD thesis, The University of Arizona.

Green, T. M. (1999). *A Lexicographic Study of Ulwa*. PhD thesis, Massachusetts Institute of Technology.

Greenberg, J. H. (1941). Some Problems in Hausa Phonology. *Language*, 17(4):316–323.

Greenberg, J. H. (1963). Some Universals of Grammar With Particular Reference to the Order of Meaningful Elements. In Greenberg, J. H., editor, *Universals of Language*, pages 73–113. MIT Press.

Greenberg, J. H., Ferguson, C. A., and Moravcsik, E. A., editors (1978). *Universals of Human Language*. Stanford University Press.

Greene, L. A. (1994). *A Grammar of Belizean Creole: Compliations From Two Existing United States Dialects*. PhD thesis, Tulane University.

Gregersen, E. A. (1961). *Luo: A Grammar*. PhD thesis, Yale University, New Haven.

Gregores, E. and Suárez, J. A. (1967). *A Description of Colloquial Guaraní*. Mouton, The Hague.

Greive, J. A. (1973). Kilba. In Kropp-Dakubu, M. E., editor, *West African Langauge Data Sheets*, volume 1, pages 326–334. West African Linguistics Society.

Grignard, S. J. (1924). Oraon-English Dictionary. *Anthropos: Internationale Sammlung Linguistischer Monographien*.

Grinevald, C. G. (1990). A Grammar of Rama. Report to National Science Foundation.

Grjunberg, A. L. (1987). *Ocherk grammatiki afganskogo jazyka (pashto)*. Nauka, Leningrad.

Grondona, V. M. (1998). *A Grammar of Mocovi*. PhD thesis, University of Pittsburgh.

Grubb, D. M. (1977). *A Practical Writing System and Short Dictionary of Kwakw'ala (Kwakiutl)*. National Museum of Man, Ottawa.

Gruber, T. R. (1993). A Translation Approach to Portable Ontology Specifications. *Knowledge Acquisition*, 5:199–220.

Grüninger, M. and Fox, M. (1995). Methodology for the Design and Evaluation of Ontologies. In *Proceedings of the Workshop on Basic Ontological Issues in Knowledge Sharing, IJCAI–95*.

Guarisma, G. (2006). Kpaʔ (A53). In Nurse, D. and Philippson, G., editors, *The Bantu languages*, pages 307–334. Routledge.

Gudschinsky, S. C., Popovich, F. B., and Popovich, H. (1970). Native Reaction and Phonetic Similarity in Maxakali Phonology. *Language*, 46(1):77–88.

Gueche, F. H. C. (2004). Noun Morphology of Befang. Master's thesis, University of Yaounde I.

Gueye, G. (1986). Les correlats articulatoires et acoustiques de la distinction +/- ATR en Ndut. *Travaux de l'Institut de Phonetique de Strasbourg*, 18:137–249.

Guillaume, A. (2004). *A Grammar of Cavineña, an Amazonian Language of Northern Bolivia.* PhD thesis, La Trobe University.

Güldemann, T. (2001). Phonological Regularities of Consonant Systems Across Khoisan Lineages. In *University of Leipzig Papers on Africa, Language and Literatures.* University of Leipzig.

Gulya, J. (1966). *Eastern Ostyak Chrestomathy*, volume 51 of *Indiana University Publications, Uralic and Altaic Series.* Indiana University Press, Bloomington.

Gumperz, J. J. and Bilibiri, H. S. (1957). Notes on the Phonology of Mundari. *Indian Linguistics*, 17:6–15.

Guoqiao, Z. and Yang, Q. (1988). The Sounds of Rongjiang Kam. In Edmondson, J. A. and Solnit, D. B., editors, *Comparative Kadai: Linguistic Studies Beyond Tai*, pages 43–58. Summer Institute of Linguistics and University of Texas at Arlington, Dallas.

Gurubasave Gowda, K. S. (1972). *Ao-Naga Phonetic Reader.* Central Institute of Indian Languages, Mysore.

Gurubasave Gowda, K. S. (1975). *Ao Grammar*, volume 1 of *CIIL grammar series.* Central Institute of Indian Languages, Mysore.

Gusain, L. (2003). *Mewati.* Number 386 in Languages of the World/Materials. Lincom GmbH.

Haan, J. W. (2001). *The Grammar of Adang: A Papuan Language Spoken on the Island of Alor East Nusa Tenggara - Indonesia*. PhD thesis, University of Sydney.

Haas, M. R. (1941). Tunica. In *Handbook of American Indian Languages: Vol IV*, pages 1–143. J. J. Augustin Publisher, New York.

Haas, M. R. (1956). *The Thai System of Writing*, volume 5 of *American Council of Learned Societies Program in Oriental Languages*. American Council of Learned Societies.

Haas, M. R. (1964). *Thai-English Student's Dictionary*. Stanford University Press, Stanford.

Haeseriju, E. V. (1966). *Ensayo de la Gramatica del K'ekchi*. Suquinay, Purulha.

Hagège, C. (1970). *La langue mbum de Nganha (Cameroun)*. Societé pour l'étude des langues africaines, Paris.

Haggard, M. (1978). The Devoicing of Voiced Fricatives. *Journal of Phonetics*, 6:95–102.

Haiman, J. (1980). *Hua: A Papuan Language of the Eastern Highlands of New Guinea*, volume 5 of *Studies in Language Companion Series*. John Benjamins, Amsterdam.

Hajdú, P. (1963). *The Samoyed Peoples and Languages*, volume 14 of *Indiana University Publications, Uralic and Altaic Series*. Indiana University Press, Bloomington.

Hale, A. and Hale, M. (1969). *Newari Phonemic Summary*, volume 5 of *Tibeto-Burman Phonemic Summaries*. Summer Institute of Linguistics, Tribhuvan University, Kirtipur.

Hale, K. (1959). *A Papago Grammar*. PhD thesis, Indiana University, Bloomington.

Hall, R. A. (1938). An Analytical Grammar of the Hungarian Language. *Language*, 14(2):9–113.

Hall, R. A. (1944). *Hungarian Grammar*, volume 21 of *Language Monographs*. Linguistic Society of America, Baltimore.

Halle, M. (1959). *The Sound Pattern of Russian*. Mouton, The Hague.

Halle, M. (1962). Phonology in Generative Grammar. *Word*, 18(1/2):54–72.

Halle, M. (1970). Is Kabardian a Vowel-less Language? *Foundations of Language*, 6:95–103.

Halle, M. (1973). Stress Rules in English: A New Version. *Linguistic inquiry*, 4(4):451–464.

Halle, M. (1992). Phonological Features. *International Encyclopedia of Linguistics*, 3:207–212.

Halpern, A. (1944). Yuma. In Osgood, C., editor, *Viking Fund Publications in Anthropology: Linguistic Structures of Native America*, volume 6. Johnson Reprint Corporation.

Hamel, P. J. (1985). *A Grammar of Loniu*. PhD thesis, University of Kansas.

Hammarberg, R. (1974). Another Look at Finnish Consonant Gradation. *Soviet Finno-Ugric Studies*, 10:171–178.

Hammarström, H. (2009). Sampling and Genealogical Coverage in WALS. *Linguistic Typology*, 13(1):105–19.

Hamouma, H. (1987). *Manuel de grammaire berbere*. Edition Association de Culture Berbere.

Hangin, J. G. (1968). *Basic Course in Mongolian*, volume 73 of *Uralic and Altaic Series*. Indiana University Press, Bloomington.

Hanke, W. (1956). Beobachtungen über den Stamm der Huari (Rio Corumbiara) Brasilien. *Archiv für Völkerkunde*, 11:67–82.

Hardman, M. J. (1966). *Jaqaru: Outline of Phonological and Morphological Structure*, volume XXII of *Janua Linguarum: Series Practica*. Mouton, The Hague.

Hardman, M. J. (1983). *Jaqaru: Compendio de estructura fonologica y morfologica*, volume 5 of *Lengua y sociedad*. Instituto des estudios peruanos, Peru.

Harms, P. L. (1984). Fonologia del Epena Pedee (Saija). *Sistemos Fonologicos de Idiomas Colombianos*, 5:157–201.

Harms, P. L. (1985). Epena Pedee (Saija): Nasalization. In Brend, R. M., editor, *From Phonology to Discourse: Studies in Six Colombian Languages*, volume 9 of *Language Data, Amerindian Series*, pages 13–18. Summer Institute of Linguistics, Dallas.

Harms, R. T. (1964). *Finnish Structural Sketch*. Indiana University Press, Bloomington.

Harms, R. T. (1966). Review Of: Proto-Finnic Final Consonants, by T. Itkonen. *Language*, 42(4):825–831.

Harms, R. T. (1969). Review of Kelkar, A.R. Studies in Hindi-Urdu I. *Language*, 45:913–927.

Harrell, R. (1962). *A Short Reference Grammar of Moroccan Arabic*, volume 1 of *Georgetown Arabic Series*. Georgetown University Press, Washington, D.C.

Harrell, R. (1965). *A Basic Course in Moroccan Arabic*. Number 8 in Georgetown Arabic Series. Georgetown University Press, Washington, D.C.

Harrington, J. P. (1928). *Vocabulary of the Kiowa Language*. Government Printing Office, Washington, D.C.

Harris, A. (2008). On the Explanation of Typologically Unusual Structures. In Good, J., editor, *Linguistic Universals and Language Change*, pages 54–76. Oxford: Oxford University Press.

Harris, A. C. (1991). Mingrelian. In Greppin, J. A. C., editor, *The Kartvelian Languages*, volume 1 of *The Indigenous Languages of the Caucasus*. Caravan Books.

Harris, Herbert Raymond, I. (1981). *A Grammatical Sketch of Comox*. PhD thesis, University of Kansas.

Harris, J. W. (1969). *Spanish Phonology*. MIT Press, Cambridge, Massachusetts.

Harrison, K. D. (2000a). *Topics in the Phonology and Morphology of Tuvan*. PhD thesis, Yale University.

Harrison, K. D. (2000b). *Topics in the Phonology and Morphology of Tuvan*. PhD thesis, Yale University.

Harry, O. G. (2003). Illustrations of the IPA: KalaḄarỊ-ỊjỌ. *Journal of the International Phonetic Association*, 33(1):113–120.

Harry, O. G. (2006). Illustrations of the IPA: Jamaican Creole. *Journal of the International Phonetic Association*, 36(1):125–131.

Hartell, R. L., editor (1993). *Alphabets des langues africaines*. UNESCO and Société Internationale de Linguistique.

Harvey, M. (1986). Ngoni Waray Amungal-Yang: The Waray Language From Adelaide River. Master's thesis, Australian National University, Canberra.

Hashimoto, M. J. (1973). *The Hakka Dialect: A Linguistic Study of Its Phonology, Syntax and Lexicon*. Cambridge University Press, Cambridge.

Haspelmath, M. (2007). Pre-established Categories Don't Exist: Consequences for Language Description and Typology. *Linguistic Typology*, 11:119–132.

Haspelmath, M. (2010). Comparative Concepts and Descriptive Categories in Crosslinguistic Studies. *Language*, 86(3):663–687.

Haspelmath, M., Dryer, M., Gil, D., and Comrie, B., editors (2005). *The World Atlas of Language Structures*. Oxford University Press.

Haspelmath, M., Dryer, M. S., Gil, D., and Comrie, B. (2008). The World Atlas of Language Structures Online. Munich: Max Planck Digital Library. Available online at http://wals.info/.

Haspelmath, M. and Tadmor, U. (2009). The Loanword Typology Project and the World Loanword Database. In Haspelmath, M. and Tadmor, U., editors, *Loanwords in the World's Languages: A Comparative Handbook*, pages 1–3. De Gruyter, Berlin.

Hasselbrink, G. (1965). *Alternative Analyses of the Phonemic System in Central-South Lappish*, volume 49 of *Indiana University Publications, Uralic and Altaic Series*. Indiana University Press, Bloomington.

Haudricourt, A. G. (1961). Richesse en phonèmes et richesse en locuteurs. *L'Homme*, 1:5–10.

Haudricourt, A.-G. (1967). La langue lakkia. *Bulletin de la Société de Linguistique de Paris*, 62:165–182.

Haudricourt, A.-G. (1971). New Caledonia and the Loyalty Islands. *Current Trends in Linguistics*, 8:359–396.

Haugen, E. (1958). The Phonemics of Modern Icelandic. *Language*, 34(1):55–88.

Hay, J. and Bauer, L. (2007). Phoneme Inventory Size and Population Size. *Language*, 83:388–400.

Hayami-Allen, R. (2001). *A Descriptive Study of the Language of Ternate, the Northern Moluccas, Indonesia*. PhD thesis, University of Pittsburgh.

Hayes, B. (2009). *Introductory Phonology*. Blackwell.

Hayes, P. (1977). In Defense of Logic. In *Proceedings of the 5th International Joint Conference on Artificial Intelligence*.

Hayes, P. (1979). The Naive Physics Manifesto. In *Expert Systems in the Micro Electronic Age*. Edinburgh University Press.

Hayes, P. (1985). The Second Naive Physics Manifesto. In Hobbs, J. and Moore, R., editors, *Formal Theories of the Commonsense World*. Ablex Publishing.

He, J. (1981). Gelaoyu Gaikuang (A Brief Description of the Gelao Language). *Minzu Yuwen*, 4:67–76.

He, J. (1983). *Gelaoyu Jianzhi (Brief Guide to Gelao Language)*. Minzu Chubanshe, Beijing.

Head, B. F. (1964). *A Comparison of the Segmental Phonology of Lisbon and Rio De Janeiro*. PhD thesis, University of Texas, Austin.

Healey, A. (1964). *Telefol Phonology*, volume 3 of *Pacific Linguistics, Series B*. Australian National University, Canberra.

Heath, J. (1981). *Basic Materials in Mara: Grammar, Texts and Dictionary.* Pacific Linguistics C-60. Australian National University.

Heath, J. (1999). *A Grammar of Koyraboro Senni: The Songhay of Gao, Mali.* Rudiger Koppe.

Heath, J. (2005a). *Tondi Songway Kiini (Songhay, Mali): Reference grammar and TSK–English–French Dictionary.* CSLI Publications.

Heath, J. (2005b). *Tondi Songway Kiini (Songhay, Mali): Reference Grammar and TSK-English-French Dictionary.* CSLI.

Heath, J. (2008). *Grammar of Jamsay.* Mouton de Gruyter.

Heath, J. and McPherson, L. (n.d.). Tonosyntax in Dogon NPs and Relative Clauses. Submitted.

Heath, J., Moran, S., Prokhorov, K., McPherson, L., and Cansler, B. (2012). Dogon Comparative Lexicon. Online: `http://dogonlanguages.org/`.

Hebeler, J., Fisher, M., Blace, R., and Perez-Lopez, A. (2009). *Semantic Web Programming.* John Wiley & Sons, Inc.

Heine, B. (1975a). Ik: eine ostafrikanische Restsprache. *Afrika und Übersee*, 59:31–56.

Heine, B. (1975b). Tepes Und Nyangi: Zwei Ostafrikanische Restsprachen. *Afrika und Übersee*, 58:263–300.

Henderson, E. J. A. (1965). *Tiddim Chin: A Descriptive Analysis of Two Texts*, volume 15 of *London Oriental Series.* Oxford University Press, London.

Henry, J. (1935). A Kaingang Text. *International Journal of American Linguistics*, 8(4):172–218.

Herault, G. (1971). *L'Aizi: Esquisse phonologique et enquête lexicale.* Institut de Linguistique Appliquee, Unversite d'Abidjan, Abidjan.

Herbert, R. J. and Poppe, N. (1963). *Kirghiz Manual*, volume 33 of *Uralic and Altaic Series*. Indiana University Press, Bloomington.

Herington, J., Kennemur, A., Lovestrand, J., Seay, K., Smith, M., and Zande, C. V. (2009). A Brief Introduction to Tangari Phonology. In *Occasional Papers in Applied Linguistics*, volume 1, pages 1–14. Online: `http://www.gial.edu/academics/opal`.

Herrfurth, H. (1964). *Lehrbuch des modernen Djawanish*. VEB Verlag Enzyklopaedie, Leipzig.

Hervas, D. L. (1784). *Catalogo delle lingue conosciute e notizia della loro affinita', e diversita' opera*. Per Gregorio Biasini all'Insegna di Pallade.

Hettich, B. G. (1997). Ossetian: Revisiting Inflectional Morphology. Master's thesis, University of North Dakota.

Hetzron, R. (1969a). *The Verbal System of Southern Agaw*, volume 12 of *Near Eastern Studies*. University of California Press.

Hetzron, R. (1969b). *The Verbal System of Southern Agaw*. University of California Press, Berkeley / Los Angeles.

Hetzron, R. (1977). *The Gunnan-Gurage Languages*. Instituto Orientale di Napoli.

Heuvel, W. v. d. (2006). *Biak: Description of an Austronesian Language of Papua*. LOT.

Heye, J. and Heye, C. A. (1967). An Outline of Southern Ivatan Phonology. *General Linguistics*, 7(2):105–120.

Hidalgo, C. A. and Hidalgo, A. (1971). *A Tagmemic Grammar of Ivatan*. Linguistic Society of the Philippines, Manila.

Hildebrandt, K. A. (2004). A Grammar and Glossary of the Manange Language. Unpublished Manuscript.

Hill, J. H. (2005). *A Grammar of Cupeño*. Number 136 in Linguistics. University of California Press.

Hill, K. C. (1967). *A Grammar of the Serrano Language*. PhD thesis, University of California, Los Angeles.

Hillenbrand, J. M. (2003). Illustrations of the IPA: English, American: Southern Michigan. *Journal of the International Phonetic Association*, 33(1):121–126.

Ho, D.-a. (1977). The Phonological System of Butonglu: A Paiwan Dialect [Paiwanyu Danlu Fangyan De Yinyun Xitong]. *Bulletin of the Institute of History and Philology. Academia Sinica [BIHP]*, 48(4):595–618.

Hoard, J. E. (1978). Obstruent Voicing in Gitksan: Some Implications for Distinctive Feature Theory. In Cook, E.-D. and Kaye, J., editors, *Linguistic Studies of Native Canada*, pages 111–119. University of British Columbia Press, Vancouver.

Hochstetler, J. L., Durieux, J. A., and Durieux-Boon, E. I. K. (2004). *Sociolinguistic Survey of the Dogon Language Area*. SIL International.

Hockett, C. F. (1955). *A Manual of Phonology*. Indiana University.

Hoddinott, W. and Kofod, F. M. (1988). *The Ngankikurungkurr Language (Daly River Area, Northern Territory)*. PhD thesis, Australian National University.

Hodge, C. T. (1947). An Outline of Hausa Grammar. *Language*, 23(4).

Hodge, C. T. and Umaru, I. (1963). *Hausa Basic Course*. Foreign Service Institute, US Department of State, Washington, D.C.

Hoff, B. J. (1968). *The Carib Language: Phonology, Morphology, Texts and Word Index*, volume 55 of *Verhandelingen van het Koninklijk Instituut voor Taal, Land en Volkenkunde*. Martinus Nijhoff, The Hague.

Hoffmann, C. (1963). *A Grammar of the Margi Language*. Oxford University Press for International African Institute, London.

Hofmann, e. (1990). A Preliminary Phonology of Bana. Master's thesis, University of Victoria.

Höftmann, H. (1971). *The Structure of the Lelemi Language.* Verlag Enzyklopädie, Leipzig.

Hohepa, P. W. (1967). *A Profile-Generative Grammar of Maori.* Indiana University Publications in Anthropology and Linguistics, Memoir 20 of the International Journal of American Linguistics. Waverly Press, Baltimore.

Hoijer, H. (1944). Chiricahua Apache. In Osgood, C., editor, *Linguistic Structures of Native America*, number 6 in Viking Fund Publications in Anthropology. Johnson Reprint Corporation.

Hoijer, H. (1946). Tonkawa. In Osgood, J., editor, *Linguistic Structures of Native America*, pages 289–311. Viking Fund Inc., New York. Reprinted in 1971 by Johnson Reprint Corp., New York.

Hoijer, H. (1949). *An Analytic Dictionary of the Tonkawa Language.* University of California Press, Berkeley / Los Angeles.

Hoijer, H. (1972). *Tonkawa Texts*, volume 73 of *University of California Publications in Linguistics.* University of California Press, Berkeley.

Hoijer, H. and Dozier, E. P. (1949). The Phonemes of Tewa, Santa Clara Dialect. *International Journal of American Linguistics*, 15(3):139–144.

Holmer, N. M. (1949). Goajiro (Arawak) I: Phonology. *International Journal of American Linguistics*, 14:45–56.

Holt, D. (1986). *The Development of the Paya Sound System.* PhD thesis, University of California at Los Angeles.

Holton, G. (2000a). *The Phonology and Morphology of the Tanacross Athabaskan Language.* PhD thesis, University of California, Santa Barbara.

Holton, G. (2000b). *The Phonology and Morphology of the Tanacross Athabaskan Language.* PhD thesis, University of California, Santa Barbara.

Holzknecht, K. G. (1973). The Phonemes of the Adzera Language. In Holzknecht, K. G. and Phillips, D. J., editors, *Papers in New Guinea Linguistics 17*, volume 38 of *Pacific Linguistics, Series A*, pages 1–11. Australian National University, Canberra.

Horne, E. C. (1961). *Beginning Javanese.* Yale University Press, New Haven.

Horrocks, I., Parsia, B., Patel-Schneider, P., and Hendler, J. (2005). Semantic Web Architecture: Stack or Two Towers? In Fages, F. and Soliman, S., editors, *Principles and Practice of Semantic Web Reasoning (PPSWR 2005)*, volume 3703, pages 37–41. Springer.

Horrocks, I., Patel-Schneider, P. F., and van Harmelen, F. (2003). From SHIQ and RDF to OWL: The Making of a Web Ontology Language. *Journal of Web Semantics*, 1(1):7–26.

Horton, A. E. (1949). *A Grammar of Luvale.* Witwatersrand University Press, Johannesburg.

Hostetler, R. and Hostetler, C. (1975). A Tentative Description of Tinputz Phonology. In Loving, R., editor, *Workers in Papua New Guinea Languages: Phonologies of Five Austronesian Languages*, volume 13. Summer Institute of Linguistics.

Householder, F. W., Kazasis, K., and Koutsoudas, A. (1964). *A Reference Grammar of Literary Dhimotiki.* Indiana University Press, Bloomington, Indiana.

Householder Jr, W. and Lofti, M. (1965). *Basic Course in Azerbaijani*, volume 45 of *Indiana University Publications, Uralic and Altaic Series.* Indiana University Press, Bloomington.

Howard, L. (1967). Camsa Phonology. In Waterhouse, V. G., editor, *Phonemic Systems of Colombian Languages*, volume 14 of *Summer Institute of Linguistics Publications in Linguistics and Related Fields*, pages 73–87. Summer Institute of Linguistics of the University of Oklahoma, Norman.

Howard, L. (1972). Fonología del camsá. In Waterhouse, V. G., editor, *Sistemas fonológicos de idiomas colombianos*, volume 1, pages 77–92. Ministerio de Gobierno, Bogota.

Howe, D. (2003). Segmental Phonology. ms.

Hudson, R. A. (1974). A Structural Sketch of Beja. In Arnott, D. W., editor, *African Language Studies*, pages 111–142. School of Oriental and African Studies, London.

Hudson, R. A. (1976). Beja. In Bender, M. L., editor, *The Non-Semitic Languages of Ethiopia*, pages 97–132. Michigan University, African Studies Centre, East Lansing.

Huffman, F. E. (1970a). *Modern Spoken Cambodian*. Yale University Press, New Haven.

Huffman, F. E. (1970b). *The Cambodian System of Writing and Beginning Reading*. Yale University Press, New Haven.

Hughes, E. J. and Leeding, V. J. (1971). The Phonemes of Nunggubuyu. *Papers on the Languages of Australian Aboriginals, Australian Aboriginal Studies*, 38:72–81.

Huisman, R. (1973). Angaataha Verb Morphology. *Linguistics*, 110:43–54.

Huisman, R. and Lloyd, J. (1981). Angaatiha Tone, Stress, and Length. In Healey, P., editor, *Angan Languages are Different*, volume 12 of *Language Data, Asian-Pacific Series*, pages 63–82. Summer Institute of Linguistics, Ukarumpa.

Huisman, R. D., Huisman, R. D., and Lloyd, J. (1981). Angaatiha Syllable Patterns. In Healey, P., editor, *Angan Languages are Different*, volume 12 of *Language Data, Asian-Pacific Series*, pages 51–62. Summer Institute of Linguistics, Ukarumpa.

Hulstaert, G. (1961). *Grammarie du Lɔmɔngɔ*, volume 39. Musée Royal de l'Afrique Centrale (MRAC), Tervuren.

Hume, E., Hall, K. C., Wedel, A., Ussishkin, A., Adda-Decker, M., and Gendrot, C. (2011). Anti-Markedness Patterns in French Epenthesis: An Information Theoretic Approach. In *Proceedings of the Berkeley Linguistic Society*, volume 37.

Hume, E. and Mailhot, F. (2010). The Role of Entropy and Surprisal in Phonologization and Language Change. In Jurafsky, D., Bell, A., Gregory, M., and Raymond, W., editors, *Origins of Sound Patterns: Approaches to Phonologization*, pages 229–254. University Press.

Hume, E. and Mailhot, F. (2011). Distinctive Features and Information Theory. Presentation given at the Laboratoire de Phonétique et Phonologie. Paris. March 14, 2011.

Hunter, G. G. and Pike, E. V. (1969). The Phonology and Tone Sandhi of Molinos Mixtec. *Linguistics*, 47:24–40.

Hurd, C. and Hurd, P. (1966). *Nasioi Language Course*. Summer Institute of Linguistics and Department of Information and Extension Services, Port Moresby, Papua New Guinea.

Hutchison, J. P. (1981). *The Kanuri Language: A Reference Grammar*. African Studies Program, University of Wisconsin, Madison.

Hutchisson, D. and Hutchisson, S. (1975). A Preliminary Phonology of Sursurunga. In Loving, R., editor, *Workers in Papua New Guinea Languages: Phonologies of Five Austronesian Languages*, volume 13, pages 163–202. Summer Institute of Linguistics.

Huttar, G. L. and Kirton, J. F. (1981). Contrasts in Yanywa Consonants. In Gonzalez, A. and Thomas, D., editors, *Linguistics Across Continents: Studies in Honor of Richard S. Pittman*, volume 2 of *Linguistic Society of the Philippines Monograph*, pages 109–116. Summer Institute of Linguistics and Linguistic Society of the Philippines, Manila.

Hyde, V. (1971). *An Introduction to the Luiseño Language*. Malki Museum Press, Morongo Indian Reservation, Banning, California.

Hyman, L. M. (1972). *A Phonological Study of Fe'fe' – Bamileke*, volume 4 of *Studies in African Linguistics, Supplement*. Indiana University, Program in African Languages and Linguistics and the African Studies Program, Bloomington.

Hyman, L. M. (1973). Notes on the History of Southwestern Mande. *Studies in African Linguistics*, 4:183–196.

Hyman, L. M. (1979). Phonology and Noun Structure. In Hyman, L. M., editor, *Aghem Grammatical Structure*, volume 7 of *Southern California Occasional Papers in Linguistics*, pages 1–72. Department of Linguistics, University of Southern California, Los Angeles.

Hyman, L. M. (1981). *Noni Grammatical Structure With Special Reference to Verb Morphology*, volume 9 of *Southern California Occasional Papers in Linguistics*. Department of Linguistics, University of Southern California, Los Angeles.

Hyman, L. M. (2008). Universals in Phonology. *The Linguistic Review*, 25:83–137.

Hyman, L. M. (2010a). Do Tones Have Features? In Goldsmith, J. A., Hume, E., and Wetzels, L., editors, *Tones and Features: Phonetic and Phonological Perspectives*. De Gruyter Mouton.

Hyman, L. M. (2010b). Does Gokana Really Have No Syllables? (Or: What's So Great About Being Universal?). In *UC Berkeley Phonology Lab Annual Report (2010)*.

Hyman, L. M. and Magaji, D. J. (1970). *Essentials of Gwari Grammar*, volume 27 of *Occasional publication*. University of Ibadan-Institute of African Studies, Ibadan.

Ibragimov, G. K. (1978). *Rutul'skij jazyk*. Nauka, Moskva.

Iddah, R. K. (1975). Siwu. In Kropp-Dakubu, M. E., editor, *West African Langauge Data Sheets*, volume 2, pages 208–215. West African Linguistics Society.

Ide, N. and Suderman, K. (2007). GrAF: A Graph-based Format for Linguistic Annotations. In *Proceedings of the Linguistic Annotation Workshop*, pages 1–8, Prague, Czech Republic.

ILIT (2012). LEGO: Lexicon Enhancement via the GOLD Ontology. Ypsilanti, MI: Institute for Language Information and Technology (LINGUIST List), Eastern Michigan University. Online: `http://lego.linguistlist.org/`.

Inmon, W. H. (1992). *Building the Data Warehouse*. John Wiley & Sons, Inc.

International Phonetic Association (2005). International Phonetic Alphabet. Technical report, International Phonetic Association.

Israel, M. (1979). *A Grammar of the Kuvi Language (With Texts and Vocabulary)*. Dravidian Linguistics Association Publication No. 27. Dravidian Linguistics Association.

Jaccard, P. (1901). Distribution de la flore alpine dans le bassin des Dranses et dans quelques régions voisines. *Bulletin de la Société Vaudoise des Sciences Naturelles*, 37:241–272.

Jacob, J. M. (1968). *Introduction to Cambodian.* Oxford University Press, London.

Jacobsen, W. H. (1964). *A Grammar of the Washo Language.* PhD thesis, University of California, Berkeley.

Jacobsen, W. H. J. (1969). Origin of the Nootka Pharyngeals. *International Journal of American Linguistics*, 35(2):125–153.

Jacottet, E. (1896). *Grammaire Louyi.* L'Ecole des Lettres d'Alger.

Jacq, P. (1998). How Many Dialects Are There? In Austin, P. K., editor, *Working Papers in Sasak*, volume 1, pages 67–90.

Jacques, G. (2004). *Phonologie et morphologie du Japhug (rGyalrong).* PhD thesis, Université Paris VII.

Jacquot, A. (1962). Notes sur la phonologie du beembe (Congo). *Journal of African Languages*, 1:232–242.

Jacquot, A. (1981). *Etudes Beembes (Congo).* Travaux et Documents de l'ORSTOM. ORSTOM, Paris.

Jacquot, A., Meeussen, A. E., and Grégoire, H. C. (1976). *Études Bantoues II.* CNRS.

Jakobson, R. (1944). A Note on Aleut Speech Sounds. In Yarmolinsky, A., Jochelson, W., Boas, F., and Jakobson, R., editors, *Aleutian Manuscript Collection.* New York Public Library.

Jakobson, R. (1949). On the Identification of Phonemic Entities. *Travaux du Cercle Linguistique de Copenhague*, 5:205–213.

Jakobson, R., Fant, G., and Halle, M. (1952). *Preliminaries to Speech Analysis.* MIT Press.

Jakobson, R. and Halle, M. (1956). *Fundamentals of Language.* Mouton, The Hague.

Jamieson, A. R. (1976a). Chiquihuitlan Mazatec Phonology. In Merrifield, W. R., editor, *Studies in Otomanguean Phonology*, pages 93–105. Summer Institute of Linguistics and University of Texas, Arlington, Dallas.

Jamieson, A. R. (1976b). Chiquihuitlan Mazatec Tone. In Merrifield, W. R., editor, *Studies in Otomanguean Phonology*, pages 107–136. Summer Institute of Linguistics and University of Texas, Arlington, Dallas.

Janssen, D., Bickel, B., and Zúñiga, F. (2006). Randomization Tests in Language Typology. *Linguistic Typology*, 10:419–440.

Jany, C. (2007). *Chimariko in Areal and Typological Perspective.* PhD thesis, University of California at Santa Barbara.

Jassem, W. (2003). Illustrations of the IPA: Polish. *Journal of the International Phonetic Association*, 33(1):103–107.

Javkin, H. R. (1980). Reviewed Work(s): Universals of Human Language, II: Phonology by Joseph H. Greenberg; Charles A. Ferguson; Edith A. Moravcsik. *Language*, 56:830–834.

Jelaska, Z. and Machata, M. G. (2005). Prototypicality and the Concept Phoneme. *Glossos*, 6:1–13.

Jiahua, Y. (1960). *Hanyu Fangyan Gaiyao [An Outline of the Chinese Dialects].* Wenzi Gaige Chubanshe, Beijing.

Jiang, Z. (1980). Naxiyu Gaikuang (A Brief Description of the Naxi Language). *Minzu Yuwen*, 3:59–73.

Johnson, H. (2000). *A Grammar of San Miguel Chimalapa Zoque.* PhD thesis, The University of Texas at Austin.

Johnson, J. B. (1962). *El idioma yaqui*, volume 10 of *Departamento de Investigaciones Antropologicas, Publicaciones.* Instituto Nacional de Antropologia e Historia, Mexico.

Johnson, L. (1976). A Rate of Change Index for Language. *Language in Society*, 5(2):165–172.

Johnson, O. E. and Levinsohn, S. H. (1990). *Gramatica Secoya*. Number 11 in Cuadernos Etnolinguisticos. Instituto Linguistico de Verano.

Johnstone, T. M. (1975). The Modern South Arabian Languages. *Afroasiatic Linguistics*, 1:93–121.

Jones, A. A. (1995). Mekeo. In Tryon, D. T., editor, *Comparative Austronesian Dictionary: An Introduction to Austronesian Studies, Part 1: Fascicle 2*. Mouton de Gruyter.

Jones, A. A. (1998). *Towards a Lexicogrammar of Mekeo (An Austronesian Language of Western Central Papua)*. Pacific Linguistics, Canberra.

Jones, D. (1967). *The Phoneme: Its Nature and Use*. Cambridge: Heffer.

Jones, D. and Ward, D. (1969). *The Phonetics of Russian*. Cambridge University Press, Cambridge.

Jones, L. K. (1986). Yawa Phonology. In *Papers in New Guinea Linguistics 25*, volume 74 of *Pacific Linguistics, Series A*, pages 1–30. Australian National University, Canberra.

Jones Jr, R. B. (1961). *Karen Linguistic Studies: Description, Comparison, and Texts*, volume 25 of *University of California Publications in Linguistics*. University of California Press, Berkeley.

Jorden, E. H. (1963). *Beginning Japanese, Part 1*, volume 5 of *Yale Linguistic Series*. Yale University Press, New Haven.

Josephs, L. S. (1975). *Palauan Reference Grammar*. Pali Language Texts: Micronesia. The University Press of Hawaii.

Judy, R. A. and Judy, J. (1962). *Fonemas del movima, con atención especial a la serie glotal*, volume 5 of *Notas Lingüísticas de Bolivia*. Instituto Lingüístico de Verano, Cochabamba.

Junusaliev, B. M. (1966). Kirgizskij jazyk. In Vinogradov, V. V., editor, *Jazyki Narodov SSSR. Volume 2: Tjurkskie Jazyki*, Jazyki Narodov SSSR, pages 482–504. Nauka, Moscow and Leningrad.

Jurafsky, D. and Martin, J. H. (2009). *Speech and Language Processing: An Introduction to Natural Language Processing, Speech Recognition, and Computational Linguistics.* Prentice-Hall, 2nd edition.

Justeson, J. S. and Stephens, L. D. (1984). On the Relationship Between the Numbers of Vowels and Consonants in Phonological Systems. *Linguistics*, 22:531–545.

Kabak, B. (2004). Acquiring Phonology Is Not Acquiring Inventories but Contrasts: The Loss of Turkic and Korean Primary Long Vowels. *Linguistic Typology*, 8(3):351–368.

Kagaya, R. and Olomi, R. (2006). *A Kiw'oso Vocabulary*, volume 14 of *Bantu Vocabulary Series*. Institute for the Study of Languages and Cultures of Asia and Africa (ILCAA), Tokyo University of Foreign Studies, Tokyo.

Kaisse, E. (1975). A Mor- or Less-pheme: A Case of Delexicalization in Modern Greek. Paper delivered at the 1975 Winter Meeting of the LSA, San Francisco.

Kaisse, E. (1976). Stress Melodies and a Fast Speech Rule in Modern Greek. *NELS VI*, pages 165–175.

Kalman, B. (1972). Hungarian Historical Phonology. In Benkö, L. and Imre, S., editors, *The Hungarian Language*. Mouton, The Hague.

Kari, J. (1979). *Athabaskan Verb Theme Categories: Ahtna*, volume 2 of *Alaska Native Language Center Research Paper*. Alaska Native Language Center, Fairbanks, Alaska.

Kari, J. and Buck, M. (1975). *Ahtna Noun Dictionary.* Alaska Native Language Center, Center for Northern Educational Research, Fairbanks, Alaska.

Karlgren, B. (1926). *Etudes sur la phonologie chinoise.* Number 15 in Archives d'Etudes Orientales. K. W. Appelberg/E. J. Brill.

Kaschube, D. V. (1967). *Structural Elements of the Language of the Crow Indians of Montana.* Series in Anthropology. University of Colorado Press.

Katz, H. (1975a). *Generative Phonologie und phonologische Sprachbunde des Ostjakischen und Samojedischen*, volume 1 of *Münchener Universitäts-Schriften, Finnisch-ugrische Bibliothek*. Universität München, München.

Katz, H. (1975b). *Selcupica 1: Materialien von Tym*, volume 1 of *Veroeffentlichungen des Finnisch-Ugrischen Seminars an der Universität München, Serie C*. Universität München, München.

Kaufman, T. (1971). *Tzeltal Phonology and Morphology*. University of California Press, Berkeley / Los Angeles.

Kawachi, K. (2007). *A Grammar of Sidaama (Sidamo), a Cushitic Language of Ethiopia*. PhD thesis, University at Buffalo, the State University of New York.

Kawasha, B. K. (2003). *Lunda Grammar: A Morphosyntactic and Semantic Analysis*. PhD thesis, University of Oregon.

Kaye, J. D. (1989). *Phonology: A Cognitive View*. Hillsfale, NJ: Lawrence Erlbaum.

Keane, E. (2004). Illustrations of the IPA: Tamil. *Journal of the International Phonetic Association*, 34(1):111–116.

Keesing, R. M. (1985). *Kwaio Grammar*, volume 88 of *Pacific Linguistics, Series B*. Australian National University, Canberra.

Kelkar, A. R. (1968). *Studies in Hindi-Urdu. Volume 1: Introduction and Word Phonology*. Postgraduate and Research Institute, Deccan College, Poona.

Kelkar, A. R. and Trisal, P. N. (1964). Kashmiri Word Phonology: A First Sketch. *Anthropological Linguistics*, 6(1):13–22.

Keller, C. E. (1976). *A Grammatical Sketch of Brao, a Mon-Khmer Language*, volume 20 of *Work Papers*. University of North Dakota, Summer Institute of Linguistics, North Dakota.

Keller, K. C. (1959). The Phonemes of Chontal (Mayan). *International Journal of American Linguistics*, 25(1):44–53.

Kelley, G. (1963). Vowel Phonemes and External Vocalic Sandhi in Telugu. *Journal of the American Oriental Society*, 83:67–73.

Kennedy, N. M. (1960). *Problems of Americans in Mastering the Pronunciation of Egyptian Arabic*. Center for Applied Linguistics, Washington, D.C.

Kennedy, R. J. (1981). Phonology of Kala Lagaw Ya in Saibai Dialect. In *Work Papers of the Summer Institute of Linguistics, Australian Aborigines Branch A 5*, volume 5, pages 103–137. Summer Institute of Linguistics, Darwin.

Kert, G. M. (1971). *Saamskij Jazyk (Kil'dinskij Dialekt): Fonetika, Morfologija, Sintaksis*. Nauka, Leningrad.

Key, H. (1961). The Phonotactics of Cayuvava. *International Journal of American Linguistics*, 27(2):143–150.

Key, M. R. (1968). *Comparative Tacanan Phonology*. Mouton, The Hague.

Key, M. R. (1978). Lingüística comparativa araucana. *VICUS*, 2:45–55.

Khajdakov, S. M. (1966). *Ocherki po Lakskoj Dialektologii*. Nauka, Moscow.

Kifer, M., Bruijn, J. D., Boley, H., and Fensel, D. (2005). A Realistic Architecture for the Semantic Web. In *Rules and Rule Markup Languages for the Semantic Web*, pages 17–29.

Kim, C.-W. (1968). The Vowel System of Korean. *Language*, 44(3):516–527.

Kim, C.-W. (1972). Two Phonological Notes: A-Sharp and B-Flat. In Braume, M. K., editor, *Contributions to Generative Phonology*, pages 155–170. University of Texas Press, Austin.

Kim, J.-m. (1986). *Phonology and Syntax of Korean Morphology*. PhD thesis, University of Southern California, Los Angeles.

Kimball, G. D. (1985). *A Descriptive Grammar of Koasati*. PhD thesis, Tulane University.

Kimball, R. (1996). *The Data Warehouse Toolkit*. John Wiley & Sons, Inc.

Kimball, R. and Ross, M. (2002). *The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling (Second Edition)*. John Wiley & Sons, Inc, Indianapolis, IN.

Kimura, M. (1968). Evolutionary Rate at the Molecular Level. *Nature*, 217:624–626.

Kimura, M. (1983). *The Neutral Theory of Molecular Evolution*. Cambridge University Press, Cambridge, UK.

Kindell, G. (1972). *Kaingang phonemics. Appendix to U. Wiesemann, Die phonologische und grammatische Struktur der Kaingang-Sprache*. Mouton, The Hague / Paris.

Kinkade, M. D. (1963). The Phonology and Morphology of Upper Chehalis 1. *International Journal of American Linguistics*, 29(3):181–195.

Kinser, J. (2009). *Python for Bioinformatics*. Jones and Bartlett Publishers.

Kiparsky, P. (1968). Linguistic Universals and Linguistic Change. In Bach, E. and Harms, R. T., editors, *Universals in Linguistic Theory*, pages 171–202. Holt, Rinehart and Winston, New York.

Kiparsky, P. (1979). *Panini as a Variationist*. MIT Press, Cambridge, MA.

Kirton, J. F. (1967). *Anyula Phonology*, volume 10 of *Pacific Linguistics, Series A*. Australian National University, Canberra.

Kirton, J. F. and Charlie, B. (1978). *Seven Articulatory Positions in Yanyuwa Consonants*, volume 51 of *Pacific Linguistics, Series A*. Australian National University, Canberra.

Klagstad, H. (1958). The Phonemic System of Colloquial Standard Bulgarian. *Slavic and East European Journal*, 16:42–54.

Klein, H. E. M. (1973). *A Grammar of Argentine Toba: Verbal and Nominal Morphology*. PhD thesis, Columbia University.

Kleine, A. (2003). Illustrations of the IPA: Standard Yiddish. *Journal of the International Phonetic Association*, 33(2):261–265.

Kleinschmidt, S. (1851). *Grammatik der grönländischen Sprache.* Reimer, Berlin.

Klingenheben, A. (1966). *Deutsch-Amharischer Sprachführer.* Wiesbaden, Harrassowitz.

Kluckhohn, C. and MacLeish, K. (1955). Moencopi Variations From Whorf's Second Mesa Hopi. *International Journal of American Linguistics*, 21:150–156.

Kochetov, A., Khatib, S. A., and Kosa, L. A. (2008). Areal-typological Constraints on Consonant Place Harmony Systems. In *Paper Presented at 2008 Annual Meeting of the Linguistic Society of America.*

Kodzasov, S. V. (1977). Fonetika archinskogo jazyka. In Kibrik, A., Kodzasov, S., Olovjannikova, I., and Samedov, D., editors, *Opyt Strukturnogo Opisanija Archinskogo Jazyka 1*, pages 185–352. Izdatel'stvo Moskovskogo Universiteta, Moscow.

Kohrt, M. (1986). The Term 'Grapheme' in the History and Theory of Linguistics. In Augst, G., editor, *New Trends in Graphemics and Orthography*, pages 80–96. Berlin: de Gruyter.

Kondo, V. F. and Kondo, R. W. (1967). Guahibo Phonemes. In Waterhouse, V., editor, *Phonemic Systems of Colombian Languages*, volume 14 of *SIL Publications in Linguistics and Related Fields*, pages 89–98. Summer Institute of Linguistics, University of Oklahoma, Norman.

Kondrak, G. (2003). Phonetic Alignment and Similarity. *Computers and the Humanities*, 37(3):273–291.

Kong, F. L. (2004). A Generative Approach to the Verb Morphology of Samba Leekɔ. Master's thesis, University of Yaounde I.

Koops, R. G. (1990). *Aspects of the Grammar of Kuteb.* PhD thesis, University of Colorado at Boulder.

Kooyers, O., Kooyers, M., and Bee, D. (1971). The Phonemes of Washkuk (Kwoma). *Te Reo*, 14:36–41.

Koshal, S. (1979). *Ladakhi Grammar.* Orient Book Distributors.

Kostic, D., Mitter, A., and Krishnamurti, B. (1977). *A Short Outline of Telugu Phonetics.* Indian Statistical Institute, Calcutta.

Kouankem, C. (2003). Complex Constructions in Bàànòò. Master's thesis, University of Yaounde I.

Kouonang, A. (1983). Esquisse phonologique du parler bali-kumbat. Master's thesis, Universite de Yaounde.

Kouwenberg, S. (1994). *A Grammar of Berbice Dutch Creole.* Number 12 in Mouton Grammar Library. Mouton de Gruyter.

Kraft, C. H. (1963). *A Study of Hausa Syntax. Volume 1 and 2.* Department of Linguistics, Hartford Seminary Foundation, Hartford.

Kraft, C. H. and Kirk-Greene, A. H. M. (1973). *Hausa.* Hodder and Stoughton, London.

Kraft, C. H. and Kraft, M. G. (1973). *Introductory Hausa.* University of California Press, Berkeley.

Krauss, M. E. (1965). Eyak: A Preliminary Report. *Canadian Journal of Linguistics/Revue Canadienne de Linguistique*, 10:167–187.

Krauss, M. E. (1975). St. Lawrence Island Eskimo Phonology and Orthography. *Linguistics*, 152:39–72.

Krejnovich, E. A. (1937). *Fonetika Nivkhskogo Jazyka.* Gosudarstvennoe Uchebno-pedagogicheskoe Izdatel'stvo, Moscow and Leningrad.

Krejnovich, E. A. (1958). *Jukagirskij jazyk.* Academy of Sciences of the USSR, Moscow and Leningrad.

Krejnovich, E. A. (1968a). Jukagirskij jazyk. In Skorik, P. J., editor, *Jazyki narodov SSSR. Volume 5: Mongol'skie, tunguso-man'chzhurskie i paleoaziaskie jazyki*, pages 435–452. Nauka, Leningrad / Moscow.

Krejnovich, E. A. (1968b). Ketskij jazyk. In Skorik, P. J., editor, *Jazyki narodov SSSR. Volume 5: Mongol'skie, tunguso-man'chzhurskie i paleoaziaskie jazyki*, pages 453–473. Nauka, Leningrad.

Krishnamurti, B. (1961). *Telugu Verbal Bases.* University of California Press, Berkeley / Los Angeles.

Kroeber, A. L. and Grace, G. W. (1960). *The Sparkman Grammar of Luiseño*, volume 16 of *University of California Publications in Linguistics.* University of California Press, Berkeley.

Kruatrachue, F. (1960). *Thai and English: A Comparative Study of Phonology for Pedagogical Applications.* Indiana University Press, Bloomington.

Krueger, J. R. (1962). *Yukat Manual*, volume 21 of *Indiana University Publications, Uralic and Altaic Series.* Indiana University Press, Bloomington.

Kruger, J. R. (1961). *Chuvash Manual*, volume 7 of *Uralic and Altaic Series.* Indiana University Press, Bloomington.

Kruspe, N. D. (1999). *Semelai.* PhD thesis, University of Melbourne.

Kuipers, A. H. (1960). *Phoneme and Morpheme in Kabardian.* The Hague: Mouton.

Kuipers, A. H. (1967). *The Squamish Language*, volume 73 of *Janua Linguarum: Series Practica.* Mouton, The Hague.

Kuipers, A. H. (1974). *The Shuswap Language: Grammar, Texts, Dictionary.* Mouton, The Hague.

Kula, N. C. (2002). *The Phonology of Verbal Derivation in Bemba.* LOT.

Kum Nang, J. (2002). A Sketch Phonology and a Step Towards the Standardization of Naki. Master's thesis, University of Yaounde I.

Kung, S. S. (2007). *A Descriptive Grammar of Huehuetla Tepehua.* PhD thesis, The University of Texas at Austin.

Kuperus, J. (1985). *The Londo Word: Its Phonological and Morphological Structure*, volume 119. Musée Royal de l'Afrique Centrale (MRAC), Tervuren.

Kutsch Lojenga, C. (1994). Kibudu: A Bantu Language With Nine Vowels. *Africana Linguistica*, 11:127–134.

Kutsch Lojenga, C. (2008). Nine Vowels and ATR Vowel Harmony in Lika, a Bantu Language in DR Congo. *Africana Linguistica*, 14:63–84.

Ladefoged, P. (1964). *A Phonetic Study of West African Languages.* Cambridge University Press, Cambridge.

Ladefoged, P. (1968). *A Phonetic Study of West African Languages – an Auditory-Instrumental Survey.* Cambridge at the University Press.

Ladefoged, P. (1990a). Some Reflections on the IPA. *Journal of Phonetics*, 18:335–346.

Ladefoged, P. (1990b). The Revised International Phonetic Alphabet. *Language*, 63(3):550–552.

Ladefoged, P. (1996). The IPA and a Theory of Phonetic Description. In *UCLA Working Papers in Phonetics*, volume 94, pages 12–19. Department of Linguistics, UCLA.

Ladefoged, P. (1997). Linguistic Phonetic Descriptions. In Hardcastle, W. J. and Laver, J., editors, *The Handbook of Phonetic Sciences*, Blackwell handbooks in linguistic, pages 589–618. Blackwell Publishers, Oxford, UK.

Ladefoged, P. (1999). English, American. *Handbook of the International Phonetic Association*, pages 41–44.

Ladefoged, P. and Johnson, K. (2010). *A Course in Phonetics.* Wadsworth, Cengage Learning.

Ladefoged, P. and Maddieson, I. (1996). *The Sounds of the World's Languages.* Blackwell, Cambridge, UK.

Ladefoged, P. and Roach, P. (1986). Revising the International Phonetic Alphabet: A Plan. *Journal of the International Phonetic Association*, 16:22–29.

Ladefoged, P. and Traill, T. (1980). The Phonetic Inadequacy of Phonological Specifications of Clicks. *UCLA Working Papers in Phonetics*, 49:1–27.

Ladefoged, P., Williamson, K., Elugbe, B., and Uwalaka, A. (1976). The Stops of Owerri Igbo. *Studies in African Linguistics, Supp.*, 6:147–163.

Lagarde, P. L. (1980). *Le Verbe Huron.* Etude Morphologique d'Apres une Description Grammaticale de la Seconde Moitie du XVIIe Siecle. Editions l'Harmattan.

Lahaussois, A. (2002). *Aspects of the Grammar of Thulung Rai: An Endangered Himalayan Language*. PhD thesis, University of California, Berkeley.

Landaburu, J. (1979). *La langue des andoke (Amazonie colombienne) Grammaire*. Société d'Études Linguistiques et Anthropologiques de France, Paris.

Landau, E., Lončarić, M., Horga, D., and Škarić, I. (1995). Illustrations of the IPA: Croatian. *Journal of the International Phonetic Association*, 25(2):83–86.

Langdon, M. H. (1970). *A Grammar of Diegueño: The Mesa Grande Dialect*, volume 66 of *University of California Publications in Linguistics*. University of California Press, Berkeley.

Lapenda, G. (1968). *Estrutura da Lingua Iate: Falada pelos indios Fulnios em Pernambuco.* Universidade Federal de Pernambuco, Recife.

Larsen, R. and Larsen, M. (1977). Orokaiva Phonology and Orthography. In Loving, R., editor, *Workpapers in Papua New Guinea Languages*, volume 19, pages 5–28. Summer Institute of Linguistics.

Larsen, R. S. and Pike, E. V. (1949). Huasteco Intonations and Phonemes. *Language*, 25:268–277.

Larsen, T. W. (1988). *Manifestations of Ergativity in Quiché Grammar.* PhD thesis, University of California, Berkeley.

Lassila, O. and Swick, R. R. (1999). Resource Description Framework (RDF): Model and Syntax Specification (Recommendation). World Wide Web Consortium. Online: `http://www.w3.org/TR/REC-rdf-syntax`.

Lasswell, S. T. (1998). *An Ecological Reference Grammar of Sölring North Frisian.* PhD thesis, University of California at Santa Barbara.

Lastra, Y. (1968). *Cochabamba Quechua Syntax.* Mouton, The Hague.

Latané, B. (1981). The Psychology of Social Impact. *American Psychologist*, 36:343–365.

Law, H. W. (1955). The Phonemes of Isthmus Nahuat. *El México Antiguo*, 8:267–278.

Laycock, D. (1978). A Little Mor. In Wurm, S. A. and Carrington, L., editors, *Second International Conference on Austronesian Linguistics: Proceedings. Fascicle 1: Western Austronesian*, volume 61 of *Pacific Linguistics, Series C*, pages 285–291. Australian National University, Canberra.

Laycock, D. C. (1965). Three Upper Sepik Phonologies. *Oceanic Linguistics*, 4:113–117.

Lazard, G. (2006). *La quête des invariants interlangues: la Linguistique est-elle une science?* Paris: Champion.

le Bris, P. and Prost, A. (1981). *Dictionnaire Bobo-francais.* Société d'Études Linguistiques et Anthropologiques de France, Paris.

Lee, J. R. (1983). *Tiwi Today: A Study of Language Change in a Contact Situation.* PhD thesis, Australian National University, Canberra.

Lee, J. R. (1984). Changes in the Roundedness Feature in Tiwi.

Lee, W.-S. and Zee, E. (2003). Illustrations of the IPA: Standard Chinese (Beijing). *Journal of the International Phonetic Association*, 33(1):109–112.

Leer, J. (1982). Issues in Tanacross Orthography. Manuscript. Alaska Native Language Center Archives.

Lees, R. B. (1961). *The Phonology of Modern Standard Turkish*, volume 6 of *Indiana University Publications, Uralic and Altaic Series*. Indiana University Press, Bloomington.

Lehfeldt, W. (1975). Die Verteilung der Phonemanzahl in den natürlichen Sprachen. *Phonetica*, 31:247–287.

Lehiste, I. and Peterson, G. E. (1961). Transitions, Glides and Diphthongs. *Journal of the Acoustical Society of America*, 33:268–277.

Lehtinen, M. (1964). *Basic Course in Finnish*. Indiana University, Bloomington.

Leitch, M. (2003). Babole (C101). In Nurse, D. and Philippson, G., editors, *The Bantu Languages*, pages 392–421. Routledge.

Leslau, W. (1938). *Lexique soqotri (sudarabique moderne) avec comparaisons et explications étymologiques*. C. Klincksieck, Paris.

Leslau, W. (1968). *Amharic Textbook*. University of California Press, Berkeley.

Levengood de Estrello, M. and Larsen, H. (1982). *Bosquejo descriptivo del quechua de huaylas*. Datos Etnolinguisticos. Instituto Linguístico de Verano, 57 edition.

Lewis, M. P., editor (2009). *Ethnologue: Languages of the World, Sixteenth Edition*. Summer Institute of Linguistics, 16 edition.

Lewis, W. D., Farrar, S., and Langendoen, D. T. (2006). Linguistics in the Internet Age: Tools and Fair Use. In *Proceedings of E-MELD 2006: Tools and Standards: the state of the art*.

Li, F.-k. (1932). A List of Chipewyan Stems. *International Journal of American Linguistics*, 7:122–151.

Li, F.-k. (1933). Chipewyan Consonants. *Bulletin of the Institute of History and Philology, Academia Sinica Ts'ai Yuan P'ei Anniversary Volume*, 1:429–467.

Li, F.-k. (1946). Chipewyan. In Hoijer, H., editor, *Linguistic Structures of native America*, pages 394–423. Wenner-Gren Foundation, New York.

Li, F.-k. (1948). The Distribution of Initials and Tones in the Sui Language. *Language*, 244:160–167.

Li, F.-k. (1964). *The Phonemic System of the Tai Lu Language*, volume 35 of *Bulletin of the Institute of History and Philology.* Academia Sinica, Taipei.

Li, F.-k. (1977a). *A Handbook of Comparative Tai.* University of Hawaii Press, Honolulu.

Li, P. J.-k. (1973). *Rukai Structure.* PhD thesis, [University of Hawaii at Manoa, Taipei.

Li, P. J.-k. (1977b). The Internal Relationships of Rukai. *Bulletin of the Institute of History and Philology, Academia Sinica*, 48(1):1–92.

Liang, M. (1984a). Laihua Yuanyinde Duanchang (Long and Short Vowels in the Lai Dialect). *Yuyan Yanjiu*, 2:57–62.

Liang, M. (1984b). Laiyu Gaikuang (A Brief Description of the Lai Language). *Minzu Yuwen*, 4:64–79.

Liccardi, M. and Grimes, J. (1968). Itonama Intonation and Phonemes. *Linguistics*, 38:36–41.

Liljegren, H. (2008). *Towards a Grammatical Description of Palula: An Indo-Aryan Language of the Hindu Kush.* PhD thesis, Stockholm University.

Liljencrants, J. and Lindblom, B. (1972). Numerical Simulation of Vowel Quality Systems: The Role of Perceptual Contrast. *Language*, 48:839–862.

Lillehaugen, B. D. (2006). *Expressing Location in Tlacolula Valley Zapotec.* PhD thesis, University of California, Los Angeles.

Lindblom, B. and Maddieson, I. (1988). Phonetic Universals in Consonant Systems. In Hyman, L. M. and Li, C. N., editors, *Language, Speech and Mind: Studies in Honour of Victoria A. Fromkin.* London: Routledge.

Lindskoog, J. N. and Brend, R. M. (1962). Cayapa Phonemics. In Elson, B. F., editor, *Studies in Ecuadorian Indian Languages 1*, pages 31–44. Summer Institute of Linguistics, University of Oklahoma, Norman.

Lindstrom, E. (2002). *Topics in the Grammar of Kuot: A Non-Austronesian Language of New Ireland, Papua New Guinea*. PhD thesis, Stockholm University.

LINGUIST List (2009). Multitree: A Digital Library of Language Relationships. Institute for Language Information and Technology (LINGUIST List), Eastern Michigan University. Ypsilanti, MI. Online: `http://multitree.org/`.

Linn, M. S. (2001). *A Grammar of Euchee (Yuchi)*. PhD thesis, University of Kansas.

Liphola, M. M. (2001). *Aspects of Phonology and Morphology of Shimakonde*. PhD thesis, The Ohio State University.

Lisker, L. (1963). *Introduction to Spoken Telugu*. American Council of Learned Societies, New York.

Lisker, L. and Abramson, A. S. (1964). A Cross-language Study of Voicing in Initial Stops: Scoustic Measurements. *Word*, 20:384–422.

Lithgow, D. (1977). Dobu Phonemics. In Loving, R., editor, *Workpapers in Papua New Guinea Languages*, volume 19, pages 73–96. Summer Institute of Linguistics.

Liu, L. (1964). Ching-p'o-yu Kai-k'uang. *Chung Kuo Yu Wen*, 5:407–417.

Lock, A. and Lock, M. (1990). Description of the Phonology of the Abau Language.

Loeweke, E. and May, J. (1964). The Phonological Hierarchy in Fasu. *Anthropological Linguistics*, 7(5):89–97.

Lojenga, C. K. (2006). Bila (D32). In Nurse, D. and Philippson, G., editors, *The Bantu languages*, pages 450–474. Routledge.

Longacre, R. E. (1966). On the Linguistic Affinities of Amuzgo. *International Journal of American Linguistics*, 32(1):46–49.

Losey, W. E. (2002). *Writing Gojri: Linguistic and Sociolinguistic Constraints on a Standardized Orthography for the Gujars of South Asia.* PhD thesis, University of North Dakota.

Lounsbury, F. G. (1953). *Oneida Verb Morphology.* Yale University Publications in Anthropology. Yale University Press, New Haven.

Lovelace, C. A. (1992). Discourse Grammar of Tsuvadi Folktales. Master's thesis, The University of Texas at Arlington.

Lukas, J. (1937). *A Study of the Kanuri Language: Grammar and Vocabulary.* Oxford University Press, London.

Lukas, J. and Willms, A. (1961). Outline of the Language of the Jarawa in Northern Nigeria (Plateau Province). *Afrika und Übersee*, 45(1/2):1–66.

Lunsford, W. A. (2001). An Overview of Linguistic Structures in Torwali, a Language of Northern Pakistan. Master's thesis, University of Texas at Arlington.

Lunt, H. G. (1973). Remarks on Nasality: The Case of Guarani. In Anderson, S. R., editor, *A Festschrift for Morris Halle*, pages 131–139. Holt, Rinehart and Winston, New York.

Lüpke, F. (2005). *A Grammar of Jalonke Argument Structure.* PhD thesis, Radboud Universiteit Nijmegen.

Lupyan, G. and Dale, R. (2010). Language Structure Is Partly Determined by Social Structure. *PLoS ONE*, 5(1):1–10.

Luvsanvandan, S. (1964). The Khalkha-Mongolian Phonemic System. *Acta Orientalia (Academiae Scientiarum Hungaricae)*, 17:175–185.

Lydall, J. (1976). Hamer. In Bender, M. L., editor, *The Non-Semitic Languages of Ethiopia*, pages 393–438. African Studies Center, Michigan State University, East Lansing.

Lynch, J. (1978). *A Grammar of Lanakel*, volume 55 of *Pacific Linguistics, Series B.* Australian National University, Canberra.

Lynch, J. (2000). *A Grammar of Anejom.* Pacific Linguistics.

Lytkin, V. I. (1966). Komi-zyrjanskij jazyk. In Lytkin, V. I. and Majtinskaja, K. E., editors, *Jazyki narodov SSSR. Volume 3: Finno-ugorskie jazyki i samodijskie jazyki*, pages 281–299. Nauka, Moscow / Leningrad.

Mac an Fhailigh, E. (1968). *The Irish of Erris, Co. Mayo.* The Dublin Institute for Advanced Studies, Dublin.

MacDonald, G. E. (1973). The Teberan Language Family. In Franklin, K., editor, *The Linguistic Situation in the Gulf District and Adjacent Areas, Papua New Guinea*, volume 26 of *Pacific Linguistics, Series C*, pages 113–148. Australian National University, Canberra.

Macdonald, R. R. and Soenyono, D. (1967). *Indonesian Reference Grammar.* Georgetown University Press, Washington, D.C.

Mackay, V. (1968). Eloyi. In Kropp-Dakubu, M. E., editor, *West African Langauge Data Sheets*, volume 1, pages 194–207. West African Linguistics Society.

Maddieson, I. (1984). *Pattern of Sounds.* Cambridge University Press, Cambridge, UK.

Maddieson, I. (1986). The Size and Structure of Phonological Inventories: Analysis of UPSID. In Ohala, J. J. and Jaeger, J. J., editors, *Experimental Phonology.* Orlando: Academic Press.

Maddieson, I. (1987). The Margi Vowel System and Labiocoronals. *Studies in African Linguistics*, 18(3):327–355.

Maddieson, I. (1991). Testing the Universality of Phonological Generalizations With a Phonetically Specified Segment Database: Results and Limitations. *Phonetica*, 48:193–206.

Maddieson, I. (2005). Vowel Inventories. In Haspelmath, M., Dryer, M. S., Gil, D., and Comrie, B., editors, *The World Atlas of Language Structures.* Oxford University Press.

Maddieson, I. (2006). Correlating Phonological Complexity: Data and Validation. *Linguistic Typology*, 10(1):106–123.

Maddieson, I. (2007). Issues of Phonological Complexity: Statistical Analysis of the Relationship Between Syllable Structures, Segment Inventories, and Tone Contrasts. In Solé, M.-J., Beddor, P. S., and Ohala, M., editors, *Experimental Approaches to Phonology*. Oxford University Press.

Maddieson, I. (2008a). Consonant Inventories. In Haspelmath, M., Dryer, M. S., Gil, D., and Comrie, B., editors, *The World Atlas of Language Structures Online*. Munich: Max Planck Digital Library.

Maddieson, I. (2008b). Tone. In Haspelmath, M., Dryer, M. S., Gil, D., and Comrie, B., editors, *The World Atlas of Language Structures Online*. Munich: Max Planck Digital Library.

Maddieson, I. (2008c). Vowel Quality Inventories. In Haspelmath, M., Dryer, M. S., Gil, D., and Comrie, B., editors, *The World Atlas of Language Structures Online*. Munich: Max Planck Digital Library.

Maddieson, I. (2011a). Consonant Inventories. In Dryer, M. S. and Haspelmath, M., editors, *The World Atlas of Language Structures Online*. Max Planck Digital Library, Munich.

Maddieson, I. (2011b). Tone. In Dryer, M. S. and Haspelmath, M., editors, *The World Atlas of Language Structures Online*. Max Planck Digital Library, Munich.

Maddieson, I. (2011c). Voicing in Plosives and Fricatives. In Dryer, M. S. and Haspelmath, M., editors, *The World Atlas of Language Structures Online*. Max Planck Digital Library, Munich.

Maddieson, I. (2011d). Vowel Quality Inventories. In Dryer, M. S. and Haspelmath, M., editors, *The World Atlas of Language Structures Online*. Max Planck Digital Library, Munich.

Maddieson, I. and Ladefoged, P. (1985). “Tense” and “Lax” in Four Minority Languages of China. *Journal of Phonetics*, 13:433–454.

Maddieson, I. and Precoda, K. (1990). Updating UPSID. In *UCLA Working Papers in Phonetics*, volume 74, pages 104–111. Department of Linguistics, UCLA.

Madtha, W. (1984). *The Christian Konkani of South Kanara (A Linguistic Analysis)*. Karnatak University, Dharwad.

Maganga, C. and Schadeberg, T. C. (1992). *Kinyamwezi Grammar, Texts, Vocabulary*, volume 1 of *East African Languages and Dialects*. Rüdiger Köppe Verlag.

Magier, D. S. (1983). *Topics in the Grammar of Marwari*. PhD thesis, University of California, Berkeley.

Major, M. M. (1979). *Phonology of the Christian Goan Konkani Dialect (As Spoken in Nagpur)*. Nagpur Vidyapeeth Mudranalaya, Nagpur.

Malecot, A. (1963). Luiseno, a Structural Analysis 1: Phonology. *International Journal of American Linguistics*, 29:89–95.

Malou, J. (1988). *Dinka Vowel System*, volume 82 of *Summer Institute of Linguistics Publications in Linguistics*. Summer Institute of Linguistics and University of Texas at Arlington, Dallas.

Manandhar, T. L. (1986). *Newari-English Dictionary: Modern Language of Kathmandu Valley. Edided by Anne Vergati*. Agam Kala Prakashan, Delhi.

Manessy, G. (1981). Les langues voltaïques. In *Les langues de l'Afrique subsahariene*. Peeters Publishers, Leuven, Belgium.

Manessy, G. and Sauvageot, S. (1963). *Notes Preliminaires to Wolof Et Serer: Etudes De Phonetique Et De Grammaire Descriptive*. Universite de Dakar, Dakar.

Manley, T. M. (1972). *Outline of the Sre Structure*, volume 12 of *Oceanic Linguistics Special Publications*. University of Hawaii Press, Honolulu.

Manning, M. and Saggers, N. (1977). A Tentative Phonemic Analysis of Ningil. In Loving, R., editor, *Phonologies of five P.N.G. languages*, volume 19, pages 49–72. Summer Institute of Linguistics.

Mansen, R. A. (1967). Guajiro Phonemes. In Waterhouse, V., editor, *Phonemic Systems of Colombian Languages*, volume 14 of *Summer Institute of Linguistics Publications in Linguistics and Related Fields*, pages 49–59. Summer Institute of Linguistics of the University of Oklahoma, Norman.

Mao, Z., Meng, C., and Zheng, Z. (1982). *A Study of the Yao Minority Language*. Institute of Cultural Minorities Language Series. Minzu Chubanshe, Beijing.

Marlett, S. A. (2005). Illustrations of the IPA: Seri. *Journal of the International Phonetic Association*, 35(1):117–121.

Marshall, C. C. and Shipman, F. M. (2003). Which Semantic Web? In *HYPERTEXT '03: Proceedings of the Fourteenth ACM Conference on Hypertext and Hypermedia*, pages 57–66, New York, NY. ACM.

Martens, M. and Tuominen, S. (1977). A Tentative Phonemic Statement in Yil in West Sepik Province. In Loving, R., editor, *Workpapers in Papua New Guinea Languages*, volume 19. Summer Institute of Linguistics.

Martin, S. E. (1951). Korean Phonemics. *Language*, 27(4):519–533.

Martin, S. E. (1952). *Morphophonemics of Standard Colloquial Japanese*, volume 47 of *Language Dissertation*. Linguistic Society of America, Baltimore.

Martin, S. E. (1954). *Korean Morphophonemics*. Waverly Press, Baltimore.

Martin, S. E. and Lee, Y.-S. C. (1969). *Beginning Korean*. Yale University Press, New Haven.

Martinet, A. (1955). *Économie des changements phonétiques*. A. Francke.

Martinet, A. (1968). Phonetics and Linguistic Evolution. In Malmberg, B., editor, *Manual of Phonetics*. North Holland, Amsterdam.

Masaquiza, F. C. and Marlett, S. A. (2008). Illustrations of the IPA: Salasaca Quichua. *Journal of the International Phonetic Association*, 38(2):223–227.

Maslova, E. (2000). A Dynamic Approach to the Verification of Distributional Universals. *Linguistic Typology*, 4:307–333.

Maslova, E. (2002). Distributional Universals and the Rate of Type Shifts: Towards a Dynamic Approach to Probability Sampling. Lecture given at the 3rd Winter Typological School, Moscow. Online: http://anothersumma.net/Publications/Sampling.pdf.

Maslova, E. (2003a). A Case of Implicational Universals. *Linguistic Typology*, 7(1):101–107.

Maslova, E. (2003b). *A Grammar of Kolyma Yukaghir*. Number 27 in Mouton Grammar Library. Mouton de Gruyter.

Maslova, E. and Nikitina, T. (2008). Stochastic Universals and Dynamics of Cross-linguistic Distributions: The Case of Alignment Types. Online: http://www.anothersumma.net/Publications/Ergativity.pdf.

Mathangwane, J. K. (1999). *Ikalanga Phonetics and Phonology: A Synchronic and Diachronic Study*. CSLI.

Matisoff, J. A. (1973). *A Grammar of Lahu*, volume 75 of *University of California Publications in Linguistics*. University of California Press, Berkeley.

Matson, D. M. (1964). *A Grammatical Sketch of Juang: A Munda Language*. PhD thesis, University of Wisconsin.

Mazaudon, M. (1973). *Phonologie Tamang (Nepal)*, volume 4 of *Collection Tradition Orale*. Société d'Études Linguistiques et Anthropologiques de France, Paris.

Mbah, N. M. (2003). *Phonological Sketch of Bangolan*. NACALCO, Yaounde.

Mbuagbaw, T. E. (1996). *Denya Phonology*. Cameroon Bible Translation Association.

Mbuagbaw, T. E. (2000). *Kenyang Segmental Phonology*. Cameroon Association for Bible Translation and Literacy.

Mc Laughlin, F. (2005). Voiceless Implosives in Seereer-Siin. *Journal of the International Phonetic Association*, 35(2):201–214.

McCarthy, J. (1988). Feature Geometry and Dependency: A Review. *Phonetica*, 45:84–108.

McConnel, U. H. (1945). Wikmunkan Phonetics. *Oceania*, 15(4):353–375.

McElhanon, K. A. (1970a). *Selepet Phonology*, volume 14 of *Pacific Linguistics, Series B.* Australian National University, Canberra.

McElhanon, K. A. (1970b). Stops and Fricatives: Non-unique Solutions in Selepet. *Linguistics*, 60:49–62.

McGuinness, D. L. and van Harmelen, F. (2004). *OWL Web Ontology Language Overview.*

McIntosh, J. B. (1945). Huichol Phonemes. *International Journal of American Linguistics*, 11(1):31–35.

McIntosh, M. (1984). *Fulfulde Syntax and Verbal Morphology.* University of Port Harcourt Press, Boston.

McKaughan, H. P. (1958). *The Inflection and Syntax of Maranao Verbs.* Publication of the Institute of National Language. Bureau of Printing, Manila.

McKaughan, H. P. and Macaraya, B. A. (1967). *A Maranao Dictionary.* University of Hawaii Press, Honolulu.

McLaughlin, J. E. (1987). *A Phonology and Morphology of Panamint.* PhD thesis, University of Kansas.

McWhorter, J. H. (2001). The World's Simplest Grammars Are Creole Grammars. *Linguistic Typology*, 5:125–166.

Meader, R. E. (1967). *Iranxe: notas gramaticais e lista vocabular*, volume 2 of *Publicacoes serie diversos linguistica.* Museu Nacional, Rio de Janeiro.

Mecklenburg, C. (1974). Phonology of Faiwol. In Loving, R., editor, *Workpapers in Papua New Guinea Languages: Studies in Languages of the OK Family*, volume 7. Summer Institute of Linguistics.

Meinhof, C. and Jones, D. (1928). Principles of Practical Orthography for African Languages. *Africa: Journal of the International African Institute*, 1(2):228–239.

Menovshchikov, G. A. (1968). Aleutskij jazyk. In Skorik, P. J., editor, *Jazyki narodov SSSR. Volume 5: Mongol'skie, tunguso-man'chzhurskie i paleoaziatskie jazyki*, pages 386–406. Nauka, Leningrad.

Merrill, E. D. (2008). Illustrations of the IPA: Tilquiapan Zapotec. *Journal of the International Phonetic Association*, 38(1):107–114.

Metcalfe, C. D. (1971). *A Tentative Phonetic Statement of the Bardi Language*, volume 38 of *Papers on the languages of the Australian Aborigines, AAS.* Australian Institute of Aboriginal and Torres Strait Islander Studies (AIATSIS), Canberra.

Mielke, J. (2004). *The Emergence of Distinctive Features.* PhD thesis, The Ohio State University.

Mielke, J. (2008). *The Emergence of Distinctive Features.* Oxford University Press.

Mielke, J. (2009). Segment Inventories. *Language and Linguistics Compass*, 3/2:700–718.

Mielke, J. and Hume, E. (2006). Distinctive Features. Encylopedia of Langauge. Elsevier.

Mierau, E. (1965). *A Descriptive Grammar of Ukrainian Low German.* PhD thesis, Indiana University.

Miestamo, M., Bakker, D., and Arppe, A. (2011). Sampling for Variety. In *Paper Presented at the Association for Linguistic Typology 9, Hong Kong, Jul 21–24.*

Miestamo, M., Sinnemäki, K., and Karlsson, F. (2008). *Language Complexity: Typology, Contact, Change.* John Benjamins, Amsterdam.

Migliazza, B. (1998a). *A Grammar of So – A Mon-Kher Language of Northeast Thailand.* PhD thesis, Mahidol University.

Migliazza, B. (1998b). *A Grammar of So – a Mon Khmer Language of Northeast Thailand.* PhD thesis, Mahidol University.

Migliazza, E. and Grimes, J. E. (1961). Shiriana Phonology. *Anthropological Linguistics*, 3:31–41.

Miller, M. T. (2007). *A Grammar of West Coast Bajau*. PhD thesis, University of Texas at Arlington.

Miller, W. R. (1966). *Acoma Grammar and Texts*, volume 40 of *University of California Publications in Linguistics*. University of California Press, Berkeley / Los Angeles.

Minch, A. (1992). *Amanab Grammar Essentials*. Summer Institute of Linguistics.

Miotti, R. (2002). Illustrations of the IPA: Friulian. *Journal of the International Phonetic Association*, 32(2):237–247.

Miret, F. (1998). Some Reflections on the Notion of Diphthong. *Contrastive Linguistics*, 34:27–51.

Mitchell, T. F. (1962). *Colloquial Arabic: The Living Language of Egypt*. The English Universities Press, London.

Mithen, S. (2003). *After the Ice*. Orion Books, London, UK.

Mithun, M. (1999). *The Languages of North America*. Cambridge University Press.

Mohanan, K. P., Archangeli, D., and Pulleyblank, D. (2009). The Emergence of Optimality Theory. In Uyechi, L. and Wee, L. H., editors, *Reality Exploration and Discovery: Pattern Interaction in Language and Life*. CSLI Publications, Stanford, CA.

Mokhtarian, B. (2004). *Die Maqāmen des Ḥarīrī in tabarisher Übersetzung*. PhD thesis, Universität Tübingen.

Mongui Sánchez, J. R. (1981). *La lengua kamentzá: fonética-fonología-fextos*. Publicaciones del Instituto Caro y Cuervo LIX, Bogotá.

Monino, Y. and Roulon, P. (1972). *Phonologie du Gbaya Kara 'Bodoe de Ndongue Bongowen (Région de Bouar, République Centrafricaine)*, volume 31. Société d'Études linguistiques et anthropologiques de France, Paris.

Monod-Becquelin, A. (1975). *La pratique linguistique des indiens trumai (Haut-Xingu, Mato Grosso, Brésil)*. Centre National de la Recherche Scientifique, Paris.

Montgomery, C. (1970). Problems in the Development of an Orthography for the Sebei Language of Uganda. *Journal of the Language Association of East Africa*, 1:48–55.

Montler, T. (2005). *An Outline of the Morphology and Phonology of Saanich, North Straits Salish*. University of Montana.

Moran, S. (2008). *A Grammatical Sketch of Western Sisaala*. Verlag Dr. Müller.

Moran, S. (2009). An Ontology for Accessing Transcription Systems (OATS). In *Proceedings of the First Workshop on Language Technologies for African Languages (AfLaT 2009)*, Athens, Greece. Association for Computational Linguistics.

Moran, S. (2012). *Phonetics Information Base and Lexicon*. PhD thesis, University of Washington.

Moran, S., McCloy, D., and Wright, R. (2012). Revisiting Population Size vs. Phoneme Inventory Size. *Language*.

Morev, L. N., Moskalev, A. A., and Plam, Y. Y. (1979). *The Lao Language*. Nauka, Moscow.

Morgenstierne, G. (1945). Notes on Burushaski Phonology. *Norsk Tidsskrift for Sprogvidenskap*, 13:61–95.

Morphy, F. (1983). Djapu, a Yolngu Dialect. In Dixon, R. M. W. and Blake, B. J., editors, *Handbook of Australian Languages 3*, pages 1–188. John Benjamins, Amsterdam.

Morris, C. (1984). *Tetun - English Dictionary*, volume 83 of *Pacific Linguistics, Series C*. Australian National University, Canberra.

Morse, M. L. A. (1976). *A Sketch of the Phonology and Morphology of Bobo (Upper Volta)*. PhD thesis, Columbia University, New York.

Moseley, C., Asher, R. E., and Tait, M., editors (1994). *Atlas of the World's Languages*. Routledge.

Moshinsky, J. (1974). *A Grammar of Southeastern Pomo*, volume 72 of *University of California Publications in Linguistics*. University of California Press, Berkeley and Los Angeles.

Mott, B. (2007). Illustrations of the IPA: Chistabino (Pyrenean Argonese). *Journal of the International Phonetic Association*, 37(1):103–114.

Moulton, W. G. (1962). *The Sounds of English and German*. University of Chicago Press, Chicago.

Mous, M. (2006). Nen (A44). In Nurse, D. and Philippson, G., editors, *The Bantu Languages*, pages 283–306. Routledge.

Mpayimana, P. (2003). Phonologie et morphologie du kinyarwanda. Master's thesis, Universite de Yaounde I.

Mukherjee, A., Choudhury, M., Basu, A., and Ganguly, N. (2008). Modeling the Cooccurrence Principles of the Consonant Inventories: A Complex Network Approach. *International Journal of Modern Physics C (IJMPC)*, 18:281–295.

Mukherjee, A., Choudhury, M., Basu, A., and Ganguly, N. (2010). Modelling the Redundancy of Human Speech Sound Inventories: An Information Theoretic Approach. *Journal of Quantitative Linguistics*, 17(4):317–343.

Mulder, J. G. (1988). *Ergativity in Coast Tsimshian (Sm'algyax)*. PhD thesis, University of California at Los Angeles.

Munroe, R. L., Fought, J. G., and Macaulay, R. K. S. (2009). Warm Climates and Sonority Classes: Not Simply More Vowels and Fewer Consonants. *Cross-Cultural Research*, 43(2):123–133.

Munroe, R. L., Munroe, R. H., and Winters, S. (1996). Cross-cultural Correlates of the Consonant-vowel (CV) Syllable. *Cross-Cultural Research*, 30:60–83.

Munroe, R. L. and Silander, M. (1999). Climate and the Consonant-Vowel (CV) Syllable: A Replication Within Language Families. *Cross-Cultural Research*, 33:43–62.

Murkelinskij, G. B. (1967). Lakskij jazyk. In Bokarev, E. A. and Lomtatidze, K. V., editors, *Jazyki narodov SSSR. Volume 4: Iberijskokavkazskie jazyki*, pages 489–507. Nauka, Leningard / Moskva.

Murock, G. P. (1967). *Ethnographic Atlas*. University of Pittsburgh Press, Pittsburgh, PA.

Mutonyi, N. (2000). *Aspects of Bukusu Morphology and Phonology*. PhD thesis, The Ohio State University.

Muzenga, J. G. K. (1980). *Esquisse de grammaire kete*, volume 104. Musée Royal de l'Afrique Centrale (MRAC), Tervuren.

Naden, A. J. (1973a). *The Grammar and Semantics of Bisa*. Summer Institute of Linguistics and School of Oriental and African Studies, London.

Naden, A. J. (1973b). *The Grammar of Bisa - a Synchronic Description of the Lebir Dialect*. PhD thesis, University of London, School of Oriental and African Studies.

Nagai, T. (2006). *Agentive and Patientive Verb Bases in North Alaskan Iñupiaq*. PhD thesis, University of Alaska Fairbanks.

Najlis, E. (1966). *Lengua Abipona. Tomo 1-2*. Centro de Estudios, Universidad de Buenos Aires, Buenos Aires.

Nakagawa, H. (2006). *Aspects of the Phonetic and Phonological Structure of the G|ui Language*. PhD thesis, University of Witwatersrand, Johannesburg.

Nardi, D. and Brachman, R. J. (2003). An Introduction to Description Logics. In Baader, F., Calvanese, D., McGuinness, D., Nardi, D., and Patel-Schneider, P., editors, *The Description Logic Handbook: Theory, Implementation, and Applications*, pages 1–39. Cambridge University Press.

Naruemon, C. (1995). A Phonological Study of Pwo Karen at Huay-Hom-Nok Village, Tambon Tha-Mae-Lob, Mae-Tha District, Lamphun Province. Master's thesis, Mahidol University.

Nater, H. F. (1984). *The Bella Coola Language*, volume 92 of *National Museum of Man Mercury Series, Canadian Ethnology Service Paper.* National Museums of Canada, Ottawa.

Naumann, C. (Forthcoming). The Phoneme Inventory of Taa (West !Xoon Dialect). In Vossen, R. and Haacke, W. H., editors, *Essays in Memory of Anthony Traill.* Rüdiger Köppe, Cologne.

Navarro, T. T. (1961). *Manual de Pronunciacion Espanola, Consejo Superior de Investigaciones Cientificas.* Publicaciones de la Revista de Filologia Espanola III, Madrid.

Nchare, A. L. (2005). Une analyse minimaliste et derivationnelle de la morphosyntaxe du Shupamem. Master's thesis, Universite de Yaounde I.

N'diaye, G. (1970). *Structure du dialécte basque de maya.* Mouton, The Hague.

Ndokobai (2003). Etude phonologique du cuvok et principes orthographiques. Master's thesis, Universite de Yaounde I.

Neisser, F. (1953). *Studien zur Georgischen Wortbildung.* Deutsche Morgenlandische Gesellschaft, Wiesbaden.

Netting, R. (1973). Kofyar. In Kropp-Dakubu, M. E., editor, *West African Langauge Data Sheets*, volume 1, pages 344–349. West African Linguistics Society.

Nettle, D. (1995). Segmental Inventory Size, Word Length, and Communicative Efficiency. *Linguistics*, 33:359–367.

Nettle, D. (1996). Language Diversity in West Africa: An Ecological Approach. *Journal of Anthrpological Archaeology*, 15:403–438.

Nettle, D. (1999a). Is the Rate of Linguistic Change Constant? *Lingua*, 108:119–136.

Nettle, D. (1999b). *Linguistic Diversity.* Oxford University Press, Oxford, UK.

Nettle, D. (1999c). Using Social Impact Theory to Simulate Language Change. *Lingua*, 108:95–117.

Nettle, D. (2007). Language and Genes: A New Perspective on the Origins of Human Cultural Diversity. *PNAS*, 104(26):10755–10756.

Newman, P. (1970). *A Grammar of Tera*, volume 57 of *University of California Publications in Linguistics*. University of California Press, Berkeley.

Newman, P. (1974). *The Kanakuru Language*. The Institute of Modern English Language Studies, University of Leeds, in association with the West African Linguistic Society, Leeds.

Newman, S. (1947). Bella Coola I: Phonology. *International Journal of American Linguistics*, 13(3):129–134.

Newman, S. (1950). Review of Boas' Kwakiutl Grammar. *International Journal of American Linguistics*, 16(2):99–101.

Newman, S. (1965). *Zuni Grammar*. University of New Mexico Press, Albuquerque.

Newmark, L. (1957). *Structural Grammar of Albanian*, volume 23 of *International Journal of American Linguistics*. Indiana University, Bloomington.

Newmeyer, F. J. (2005). *Possible and Probably Languages: A Generative Perspective on Linguistic Typology*. Oxford University Press, New York, NY.

Newmeyer, F. J. (2007). Linguistic Typology Requires Crosslinguistic Formal Categories. *Linguistic Typology*, 11:133–157.

Newmeyer, F. J. (2010). On Comparative Concepts and Descriptive Categories: A Reply to Haspelmath. *Language*, 86(3):688–695.

Newton, B. (1972). *The Generative Interpretation of Dialect: A Study of Modern Greek Phonology*, volume 8 of *Cambridge Studies in Linguistics*. Cambridge University Press, Cambridge.

Nforgwei, S. T. (2004). *A Study of the Phonological, Morphological and Syntactic Processes in the Standardisation of Limbum*. PhD thesis, Univeresity of Yaounde I.

Nganganu, K. L. (2001). *The Phonology of Ambele*. PhD thesis, University of Yaounde I.

Nganmou, A. (1991). *Modalites Verbales: Temps, Aspect Et Mode en Medumboc*. PhD thesis, Universite de Yaounde.

Ngeloh Takwe, J. (2002). Structural Phonology of Bamunka. Master's thesis, University of Yaounde I.

Ngoran, L. L. (1999). A Sketch Outline of the Phonology of Ndemli. Master's thesis, University of Yaounde I.

Ngouagna, J. P. (1988). Esquisse Phonologique Du Ngomba. Master's thesis, Universite de Yaounde.

Ngue um, E. (2002). Morphologie verbale du mvùmbɔ̀. Master's thesis, Universite de Yaounde I.

Nguyen, P. C. (1974). A Contribution to the Phonological Interpretation of the Diphthongs in Modern Vietnamese. *Phonetica Pregensia*, 4:133–142.

Nichols, J. (1992). *Linguistic Diversity in Space and Time*. Chicago: University of Chicago Press.

Nichols, J. (1996a). Chechen. In Smeets, R., editor, *North East Caucasian Languages: Part 2*, number 4 in The Indigenous Languages of the Caucasus. Caravan Books, Delmar, NY.

Nichols, J. (1996b). Ingush. In Smeets, R., editor, *The Indigenous Languages of the Caucasus*, volume 4 of *The Indigenous Languages of the Caucasus*. Caravan Press, Delmar, NY.

Nichols, J. (2003). Diversity and Stability in Language. In Janda, R. D. and Joseph, B. D., editors, *Handbook of Historical Linguistics*, pages 283–310. Blackwell, London, UK.

Nichols, J. (2007). What, if Anything, is Typology? *Linguistic Typology*, 11:231–238.

Nigam, R. and Neethivanan, J. (1971). *Survey of Kanauri in Himachal Pradesh*, volume 3. Lang. Monograph, Calcutta.

Nihalani, P. (1995). Illustrations of the IPA: Sindhi. *Journal of the International Phonetic Association*, 25(2):95–98.

Nnomo, T. A. and Mbezele, L. E. (1982). *Elements de Grammaire Ewondo*. Collège Liberman, Donala.

Nordlinger, R. (1990). *A Sketch Grammar of Bilinarra*. PhD thesis, University of Melbourne.

Noss, R. B. (1954). *An Outline of Siamese Grammar*. PhD thesis, Yale University, New Haven.

Noss, R. B. (1964). *Thai Reference Grammar*. Foreign Service Institute, Department of State, United States Government, Washington, D.C.

Noukeu, S. (2002). Esquisse phonologique du kada (Guidar). Master's thesis, University of Yaounde I.

Novikova, K. A. (1960). *Ocherki Dialektov Evenskogo Jazyka: Ol'skij Govor 1*. Academy of Sciences of the USSR, Moscow / Leningrad.

Nowak, A., Szamrej, J., and Latané, B. (1990). From Private Attitude to Public Opinion: A Dynamical Theory of Social Impact. *Psychological Review*, 97:362–376.

Nuchanart, W. (1998a). Thavung Phonology at Muang Khamkert, Bolikhamxai Province, Lao P.D.R. Master's thesis, Mahidol University.

Nuchanart, W. (1998b). Thavung Phonology at Muang Khamkert, Bolikhamxai Province, Lao P.D.R. Master's thesis, Mahidol University.

Nurse, D. (1986). Reconstruction of Dahalo History Through Evidence From Loanwords. *Sprache und Geschichte in Afrika*, 7(2):267–305.

Oates, L. F. (1964a). *A Tentative Description of the Gunwinggu Language (Of Western Arnhem Island)*, volume 10 of *Oceania Linguistic Monographs*. University of Sydney, Sydney.

Oates, L. F. (1964b). Distribution of Phonemes and Syllables in Gugu-Yalanji. *Anthropological Linguistics*, 6(1):23–6.

Oates, W. J. and Oates, L. F. (1964). Gugu-Yalanji Linguistic and Anthropological Data. In et al., W. O., editor, *Gugu-Yalanji and Wik Munkan Language Studies*, volume 2 of *Occasional Papers in Aboriginal Studies*, pages 1–17. AIATSIS, Canberra.

Oatridge, D. and Oatridge, J. (1973). Phonemes of Binumarien. In Watson, J. B., editor, *Anthropological Studies in the Eastern Highlands of New Guinea: The Languages of the Eastern Family of the East New Guinea Highland Stock*, volume 1. University of Washington Press.

Obata, K. (2003). *A Grammar of Bilua: A Papuan Language of the Solomon Islands.* Pacific Linguistics.

Oberly, S. I. (2008). *A Phonetic Analysis of Southern Ute With a Discussion of Southern Ute Language Policies and Revitalization.* PhD thesis, University of Arizona.

Obolensky, S., Panah, K. Y., and Nouri, F. K. (1963). *Persian Basic Course.* Center for Applied Linguistics, Washington, D.C.

Ochoa Peralta, M. A. (1984). *El idioma huasteco de Xiloxuchil, Veracruz.* Instituto Nacional de Antropologiá e Historia, México.

O'Conner, J. D. (1973). *Phonetics.* Penguin, Middlesex.

O'Connor, M. C. (1987). *Topics in Northern Pomo Grammar.* PhD thesis, University of California, Berkeley.

Odden, D. (2003). Rufiji-Ruvume (N10, P10-20). In Nurse, D. and Philippson, G., editors, *The Bantu Languages*, pages 529–580. Routledge.

Odden, D. (2005). *Introducing Phonology.* Cambridge University Press.

O'Grady, G. (1964). *Nyangumata*, volume 9 of *Oceania Linguistic Monographs.* University of Sidney, Sidney.

Ohala, M. (1983). *Aspects of Hindi Phonology.* Motilal Banarsidass, Delhi.

Okell, J. (1969). *A Reference Grammar of Colloquial Burmese (Two Volumes).* Oxford University Press, London.

Olmsted, D. L. (1964). *A History of Palaihnihan Phonology.* University of California Press, Berkeley / Los Angeles.

Olmsted, D. L. (1966). *Achumawi Dictionary.* University of California Press, Berkeley.

Olson, K. S. (2004). Illustrations of the IPA: Mono. *Journal of the International Phonetic Association*, 34(2):233–238.

Omar, A. H. (1981). *The Iban Language of Sarawak: A Grammatical Description.* Kementarian Pelajaran Malaysia, Kuala Lampur.

Oranuch, S.-a. (1984). Phrases to Sentences in Kuay (Surin). Master's thesis, Mahidol University.

Osbiston, R. (1975). Mampruli. In Kropp-Dakubu, M. E., editor, *West African Langauge Data Sheets*, volume 2, pages 115–123. West African Linguistics Society.

Osborn, H. A. (1948). Amahuaca Phonemes. *International Journal of American Linguistics*, 14:188–190.

Osborn, H. A. J. (1966). Warao I: Phonology and Morphophonemics. *International Journal of American Linguistics*, 32:108–123.

Osborne, C. R. (1974a). *The Tiwi Language*, volume 55 of *Australian Aboriginal Studies.* Australian Institute of Aboriginal Studies, Canberra.

Osborne, C. R. (1974b). *The Tiwi Language.* Australian Aboriginal Studies No. 55. Linguistic Series No. 21. Australian Institute of Aboriginal Studies.

Osterhout, L., Wright, R. A., and Allen, M. (2007). The Psychology of Linguistic Form. In Hogan, P. C., editor, *The Cambridge Encyclopedia of the Language Sciences.* Cambridge University Press.

Ott, W. and Ott, R. (1967). Phonemes of the Ignaciano Language. *Linguistics*, 35:56–60.

Ouyang, J. and Zheng, Y. (1963). Laiyu Gaikuang (A Brief Description of Lai). *Zhongkuo Yuwen*, 5:432–433.

Ouyang, J. and Zheng, Y., editors (1980). *Liyu Jianzhi (A Brief Guide to the Li Language).* National Institute of Minorities, Beijing.

Ozanne-Rivierre, F. (1976). *Le iaai. Langue mélanésienne d'Ouvéa (Nouvelle-Calédonie), Phonologie, Morphologie, Esqujisse Syntaxique.* Société d'Études linguistiques et anthropologiques de France, Paris.

Padayodi, C. M. (2008). Illustrations of the IPA: Kabiye. *Journal of the International Phonetic Association*, 38(2):215–221.

Painter, C. (1974). Hill Guang. In Kropp-Dakubu, M. E., editor, *West African Langauge Data Sheets*, volume 2, pages 62–69. West African Linguistics Society.

Palmer, B. (1999). *A Grammar of the Kokota Language, Santa Isabel, Solomon Islands.* PhD thesis, The University of Sydney.

Palmer, F. R. (1962). *The Morphology of the Tigre Noun.* Oxford University Press, London.

Palmer, F. R. (1968). *Selected Papers of J.R. Firth 1952–1959.* London: Longmans.

Panadda, B. (1993). A Phonological Study of Akha in Pa-kha-suk-jai Village, Tambol Mae-saː-nok, King Amphur Mae-fa-luang, Chiang Rai Province. Master's thesis, Mahidol University.

Panfilov, V. Z. (1962). *Grammatika nivxskogo jazyka.* Nauka, Moscow.

Panfilov, V. Z. (1968). Nivxskij jazyk. In Skorik, P. J., editor, *Jazyki narodov SSSR. Volume 5: Mongol'skie, tungusko-man'chzhurskie i paleoaziatskie jazyki*, pages 408–434. Nauka, Moskva.

Parker, G. J. (1977). Review of Lastra, Y.: Cochabamba Quechua Syntax. *Language*, 45(3):702–703.

Parker, J. and Parker, D. (1974). A Tentative Phonology of Baining (Kakat Dialect). In *Phonologies of Four Papua New Guinea Languages, Workpapers in Papua New Guinea Languages*, volume 4. Summer Institute of Linguistics, Ukarumpa.

Parker, S. G. (1991). *Estudios sobre la fonologia des chamicuro*, volume 30 of *Serie Lingüística Peruana*. Instituto Linguistico de Verano.

Pasch, H. (1986). *Die Mba-Sprachen: Die Nominalklassensysteme und die genetische Gliederung einer Gruppe von Ubangi-Sprachen*, volume 6 of *Sprache und Geschichte in Afrika, Beiheft*. Helmut Buske, Hamburg.

Passy, P. (1888). Our Revised Alphabet. *The Phonetic Teacher*, pages 57–60.

Paster, M. (2006). Aspects of Maay Phonology and Morphology. *Studies in African Linguistics*, 35(1):73–120.

Patrie, J. (1982). *The Genetic Relationship of the Ainu Language*. University of Hawaii Press, Honolulu, Hawaii.

Paulian, C. (1975). *Le kukya, langue Teke du Congo*, volume 49–50 of *Bibliotheque de la SELAF*. Société d'Études Linguistiques et Anthropologiques de France, Paris.

Payne, D. (1985). *Aspects of the Grammar of Yagua: A Typological Approach*. PhD thesis, University of California at Los Angeles.

Payne, D. L. (1981). *The Phonology and Morphology of Axininca Campa*, volume 66 of *Summer Institute of Linguistics Publications in Linguistics*. Summer Institute of Linguistics, Dallas.

Peasgood, E. T. (1972). Carib Phonology. In Grimes, J. E., editor, *Languages of the Guianas*, pages 35–41. Summer Institute of Linguistics, University of Oklahoma, Norman.

Pellegrino, F., Marsico, E., Chitoran, I., and Coupé, C. (2009). *Approaches to Phonological Complexity*. Mouton de Gruyter, Berlin.

Pence, A. (1966). Kunimaipa Phonology: Hierarchical Levels. In *Papers in New Guinea Linguistics 5*, volume 7 of *Pacific Linguistics, Series A*, pages 49–97. Australian National University, Canberra.

Penzl, H. (1955). *A Grammar of Pashto; A Descriptive Study of the Dialect of Kandahar*. American Council of Learned Societies, Washington, D.C.

Pepandze, N. J. (2005). *The Morpho-Syntax of Baba*. PhD thesis, University of Yaounde I.

Percival, W. K. (1964). *A Grammar of Toba-Batak*. PhD thesis, Yale University.

Pericliev, V. (2004). There Is No Correlation Between the Size of a Community Speaking a Language and the Size of the Phonological Inventory of That Language. *Linguistic Typology*, 8(3):376–383.

Pericliev, V. (2010). *Machine-Aided Linguistic Discovery: An Introduction and Some Examples*. London: Equinox.

Pericliev, V. and Valdés-Pérez, R. E. (2002). Differentiating 451 Languages in Terms of Their Segment Inventories. *Studia Linguistica*, 56(1):1–27.

Perkins, R. D. (1980). *The Evolution of Culture and Grammar*. PhD thesis, SUNY, Buffalo, NY.

Perkins, R. D. (1988). The Covariation of Culture and Grammar. In Hammond, M., Moravcsik, E. A., and Wirth, J., editors, *Studies in Syntactic Typology*, pages 359–378. Benjamins, Amsterdam.

Perkins, R. D. (1992). *Deixis, Grammar, and Culture*. Benjamins, Amsterdam.

Perrin, M. and Hill, M. (1969). *Mambila (Parler D'Atta): Description Phonologique*. Université Fédérale du Cameroun, Section de Linguistique Appliquée, Yaoundé.

Peterson, J. (2006). Kharia. Unpublished Habilitationsschrift.

Pfeiffer, M. (1972). *Elements of Kurux Historical Phonology*. E. J. Brill, Leiden.

Pharris, N. J. (2006). *Winuunsi Tm Talapaas: A Grammar of the Molalla Language.* PhD thesis, University of Michigan.

Philipp, M. (1974). *Phonologie des Deutschen.* W. Kohlhammer, Stuttgart.

Phillips, P. J. (1976). *Wahgi Phonology and Morphology*, volume 36 of *Pacific Linguistics, Series B.* Australian National University, Canberra.

Phunsap, I. N. (1984). A Relational Grammar Analysis of the Buriram Dialect of Northern Khmer. Master's thesis, Mahidol University.

Picanço, G. L. (2005). *Mundurukú: Phonetics, Phonology, Synchrony, Diachrony.* PhD thesis, University of British Columbia.

Pichl, W. J. (1973a). Krim. In Kropp-Dakubu, M. E., editor, *West African Langauge Data Sheets*, volume 1, pages 374–383. West African Linguistics Society.

Pichl, W. J. (1973b). Mmani. In Kropp-Dakubu, M. E., editor, *West African Langauge Data Sheets*, volume 2, pages 136–141. West African Linguistics Society.

Pichl, W. J. (1973c). Ndut-Falor. In Kropp-Dakubu, M. E., editor, *West African Langauge Data Sheets*, volume 2, pages 159–166. West African Linguistics Society.

Pichl, W. J. (1973d). Sherbro. In Kropp-Dakubu, M. E., editor, *West African Langauge Data Sheets*, volume 2, pages 223–229. West African Linguistics Society.

Pierrehumbert, J. B. and Talkin, D. (1992). Lenition of /h/ and Glottal Stop. In Docherty, G. J. and Ladd, D. R., editors, *Gesture, Segment, Prosody: Papers in Laboratory Phonology II.* Cambridge University Press, Cambridge, UK.

Pike, E. and Diatta, B. (1994). The Phonology of Joola Huluf. *Journal of West African Languages*, 24(2):77–84.

Pike, E. V. (1951). Tonemic-intonemic Correlation in Mazahua (Otomi). *International Journal of American Linguistics*, 17:37–41.

Pike, E. V. and Pike, K. L. (1947). Immediate Constituents of Mazateco Syllables. *International Journal of American Linguistics*, 13:78–91.

Pike, K. L. (1947). *Phonemics: A Technique for Reducing Languages to Writing*. University of Michigan Press.

Pinnow, H.-J. (1959). *Versuch einer Historischen Lautlehre der Kharia-Sprache*. Harrassowitz, Wiesbaden.

Pinnow, H.-J. (1964). Bemerkungen zur Phonetik und Phonemik des Kurukh. *Indo-Iranian Journal*, 8:32–59.

Pinnow, H.-J. (1972). Über die Vokale im Hindi. *Zeitschrift fur Phonetik und Allgemeine Sprachwissenschaft*, 7:43–53.

Plank, F. (2003). There's More Than One Way to Make Sense of One-way Implications – and Sense They Need to Be Made Of. *Linguistic Typology*, 7(1):128–140.

Platiel, S. (1979). *Description du parler samo de toma*. PhD thesis, Sorbonne.

Ploykaew, P. (2001). *Samre Grammar*. PhD thesis, Mahidol University.

Plungian, V. A. and Tembiné, I. (1994). Stratégies communicatives au Mali: langues régionales, bambara, français. In Dumestre, G., editor, *Vers une description sociolinguistique du pays Dogon: attitudes linguistiques et problémes de standardisation*, pages 163–195. Didier Erudition, Paris.

Plunkett, G. C. (2009). An Overview of Foodo. *Journal of West African Languages*, 36(1–2):107–138.

Poletto, R. E. (1998). *Topics in Runyankore Phonology*. PhD thesis, Ohio State University.

Polome, E. C. (1967). *Swahili Language Handbook*. Center for Applied Linguistics, Washington, D.C.

Poornima, S. and Good, J. (2010). Modeling and Encoding Traditional Wordlists for Machine Applications. In *Proceedings of the 2010 Workshop on NLP and Linguistics: Finding the Common Ground*, pages 1–9. Association for Computational Linguistics.

Poppe, N. (1964). *Bashkir Manual*, volume 36 of *Indiana University Publications, Uralic and Altaic Series*. Indiana University, Bloomington.

Port, R. F. and Leary, A. (2005). Against Formal Phonology. *Language*, 81:927–64.

Post, U. R. (1966). The Phonology of Tiruray. *Philippine Journal of Science*, 95(3):563–575.

Powell, J. V. (1975). Proto-Chimakuan: Materials for a Reconstruction. In *Working Papers in Linguistics, University of Hawaii*, volume 7. University of Hawaii Press, Honolulu.

Prasse, K.-G. (1972). *Manuel de grammaire touarègue (tahaggart). Part 1-3: Phonétique, écriture, pronon.* Akademisk Forlag, Copenhagen.

Price, N. (1989). *Notes on Mada Phonology.* The Summer Institute of Linguistics, Inc, Dallas.

Price, P. D. (1976). Southern Nambiquara Phonology. *International Journal of American Linguistics*, 42(4):338–348.

Pride, K. (1965). *Chatino Syntax.* Summer Institute of Linguistics, Norman, Oklahoma.

Priest, P. (1968). Phonemes of the Sirionó Language. *Linguistics*, 41(6):102–108.

Prince, A. and Smolensky, P. (1993). *Optimality Theory: Constraint Interaction in Generative Grammar.* Rutgers Optimality Archive; ROA-537–0802.

Pring, J. T. (1967). *A Grammar of Modern Greek on a Phonetic Basis.* Hodder & Stoughton, London.

Prost, G. R. (1967). Chacobo. In Matteson, E., editor, *Bolivian Indian Grammars 1*, pages 285–359. Summer Institute of Linguistics, Norman, Oklahoma.

Prost, R. P. A. (1956). *La langue sonay et ses dialectes*, volume 47 of *Memoires de l'Institut Français d'Afrique Noire.* Institut Français d'Afrique Noire, Dakar.

Prud'Hommeaux, E. and Seaborne, A. (2006). SPARQL Query Language for RDF. Technical report, W3C.

Pugh, S. M. and Press, I. (1999). *Ukrainian: A Comprehensive Grammar.* Routledge.

Pukui, M. K. and Elbert, S. H. (1965). *Hawaiian-English Dictionary.* University of Hawaii Press, Honolulu.

Pullum, G. K. and Ladusaw, W. A. (1996). *Phonetic Symbol Guide.* University of Chicago, Chicago, IL, second edition edition.

Purnell, H. C. (1965). *Phonology of a Yao Dialect.* Hartford Seminary Foundation, Hartford.

Purnell, H. C. (1972). *Miao and Yao Linguistic Studies. Selected Articles in Chinese, Translated by Chang Yu-hung and Chu Kwo-ray*, volume 7 of *Linguistics Series.* Department of Asian Studies, Cornell University, Ithaca.

Pyne, P. C. (1972). Anyin. In Kropp-Dakubu, M. E., editor, *West African Langauge Data Sheets*, volume 1, pages 45–56. West African Linguistics Society.

Qiu Efeng, N. X. (1980). Wayu Gaikuang (A Brief Description of the Va/Wa Language). *Minzu Yuwen*, 1:58–69.

Quaireau, A. (1987). *Description de l'agni des parlers moronou, ndénié et bona.* PhD thesis, Université de Grenoble III.

Quigley, E. C. (2003). Awara Phonology. Master's thesis, The University of North Dakota.

R Development Core Team (2011). R: A Language and Environment for Statistical Computing. Vienna: R Foundation for Statistical Computing. Online: http://www.r-project.org.

Rabel, L. (1961). *Khasi, a Language of Assam*, volume 10 of *Louisiana State University Studies, Humanities Series.* Louisiana State University Press, Baton Rouge.

Ramaswami, N. (1982). Brokskat.

Rand, E. (1968). The Structural Phonology of Alabaman, a Muskogean Language. *International Journal of American Linguistics*, 34:94–103.

Raudenbush, S. W. and Bryk, A. S. (2002). *Hierarchical Linear Models: Applications and Data Analysis Methods*. Sage Publications, Inc.

Ray, P. S. (1967). Dafla Phonology and Morphology. *Anthropological Linguistics*, 9:9–14.

Redden, J. E. (1979). *A Descriptive Grammar of Ewondo*, volume 4 of *Occasional Papers on Linguistics*. Southern Illinois University at Carbondale, Department of Linguistics, Carbondale, Illinois.

Reeder, J. (1998). Pagibete, a Northern Bantu Borderlands Language: A Grammatical Sketch. Master's thesis, The University of Texas at Arlington.

Reesink, G. P. (1987). *Structures and Their Functions in Usan: A Papuan Language of Papua New Guinea*. John Benjamins, Amsterdam.

Reetz, H. (2005). HTML Interface to the UCLA Phonological Segment Inventory Database (UPSID). Online: http://web.phonetik.uni-frankfurt.de/upsid.html.

Regueira, X. L. (1996). Illustrations of the IPA: Galician. *Journal of the International Phonetic Association*, 26(2):119–122.

Reh, M. (1996). *Anywa Language: Description and Internal Reconstructions*. Rüdiger Köppe Verlag.

Rehg, K. (1984a). Nasal Substitution Rules in Ponapean. In Bender, B. W., editor, *Studies in Micronesian Linguistics*, volume 80 of *Pacific Linguistics, Series C*, pages 317–337. Australian National University, Canberra.

Rehg, K. (1984b). On the History of Ponapean Phonology. In Bender, B. W., editor, *Studies in Micronesian Linguistics*, volume 80 of *Pacific Linguistics, Series C*, pages 281–316. Australian National University, Canberra.

Rehg, K. L. (1981). *Ponapean Reference Grammar*. University of Hawaii Press, Honolulu.

Reilly, E. M. (2002). A Survery of Texistepec Popoluca Verbal Morphology. B.A. Thesis.

Relich, R. R. (1973). A General Grammar of Koba Baga With Texts and Translations. Master's thesis, Duquesne University.

Remijsen, A. C. L. (2002). *Word-prosodic Systems of Raja Ampat Languages*. PhD thesis, University of Leiden.

Renck, G. L. (1967). A Tentative Statement of the Phonemes of Yagaria. In *Pacific Linguistics, Series A*, volume 12, pages 19–48. Australian National University, Canberra.

Renck, G. L. (1975). *A Grammar of Yagaria*, volume 40 of *Pacific Linguistics, Series B*. Australian National University, Canberra.

Rialland, A. and Djamouri, R. (1984). Harmonie vocalique, consonantique et structures de dépendance dans le mot en mongol khalkha. *Bulletin de la Société de Linguistique de Paris*, 79(1):333–383.

Rice, K. and Avery, P. (1993). Segmental Complexity and the Structure of Inventories. In *Toronto Working Papers in Linguistics*, volume 12. University of Toronto.

Rich, F. (1963). Arabela Phonemes and High-level Phonology. In Waterhouse, V. G., editor, *Studies in Peruvian Indian Languages*, volume 1, pages 193–206. Summer Institute of Linguistics, University of Oklahoma, Norman.

Riehl, A. K. and Jauncey, D. (2005). Illustrations of the IPA: Tamambo. *Journal of the International Phonetic Association*, 35(2):255–259.

Rijkhoff, J. and Bakker, D. (1998). Language Sampling. *Linguistic Typology*, 2:263–314.

Rijkhoff, J., Bakker, D., Hengeveld, K., and Kahrel, P. (1993). A Method of Language Sampling. *Studies in Language*, 17(1):169–203.

Rischel, J. (1974). *Topics in West Greenlandic Phonology*. Akademisk Forlag, Copenhagen.

Ristinen, E. K. (1960). *An East Cheremis Phonology*, volume 1 of *Uralic and Altaic Series*. Indiana University Press, Bloomington.

Ristinen, E. K. (1965). On the Phonemes of Nenets. *Ural-Altaische Jahrbücher*, 40:22–44.

Ristinen, E. K. (1968). Problems Concerning Vowel Length in Nenets. *Ural-Altaisches Journal*, 40:22–44.

Roach, P. (2004). Illustrations of the IPA: English, British: Received Pronunciation. *Journal of the International Phonetic Association*, 34(2):239–245.

Robbins, F. E. (1961). Quiotepec Chinantec Syllable Patterning. *International Journal of American Linguistics*, 27:237–250.

Robbins, F. E. (1968). *Quiotepec Chinantec Grammar*, volume 4 of *Papeles de la Chinantla*. Museo Nacional de Antropologí?a, México.

Robbins, F. E. (1975). Nasal Words Without Phonetic Vowels in Quiotepec Chinantec. In Brend, R. M., editor, *Studies in Tone and Intonation by Members of the Summer Institute of Linguistics*, Bibliotheca Phonetica, pages 126–130. Karger, Basel.

Roberts, J. R. (1987). *Amele*. Croom Helm Descriptive Grammar Series. Croom Helm, London.

Robins, R. H. (1953). The Phonology of the Nasalized Verbal Forms in Sundanese. *Bulletin of the School of Oriental and African Studies*, 15:138–145.

Robins, R. H. (1957). *Vowel Nasality in Sundanese: A Phonological and Grammatical Study*. Studies in Linguistics. Blackwell, Oxford.

Robins, R. H. (1958). *The Yurok Language: Grammar, Texts, Lexicon*. University of California Press.

Robins, R. H. and Waterson, N. (1952). Notes on the Phonetics of the Georgian Word. *Bulletin of the School of Oriental and African Studies*, 15:55–72.

Robinson, J. O. S. (1976). His and Hers Morphology: The Strange Case of the Tarok Possessives. *Studies in African Linguistics Supplement*, 6:201–209.

Robinson, L. C. (2008). *Dupaningan Agta: Grammar, Vocabulary, and Texts*. PhD thesis, University of Hawai'i.

Robinson, S. (2006). The Phoneme Inventory of the Aita Dialect of Rotokas. *Oceanic Linguistics*, 45(1):206–209.

Roddy, K. M. (2007). A Sketch Grammar of Satawalese, the Language of Satawal Island, Yap State, Micronesia. Master's thesis, University of Hawai'i.

Rodegem, F. (1967). *Précis de grammaire rundi*. E. Story-Scientia.

Rodrigues, A. D. (1980). Contribuicoes das linguas Brasileiras para a fonetica e a fonologia. In *XII Reuniao Brasileira de Antropologia, Rio de Janeiro*.

Rogers, D. and d'Arcangeli, L. (2004). Illustrations of the IPA: Italian. *Journal of the International Phonetic Association*, 34(1):117–121.

Rood, D. S. (1975). The Implications of Wichita Phonology. *Language*, 51:315–337.

Roop, D. H. (1970). *A Grammar of the Lisu Language*. PhD thesis, Yale University.

Rose, P. and Morphy, F. (1982). *Yolngu Sounds (Tape and Listening Exercise)*. Australian National University, Canberra.

Rosen, N. (2007). *Domains in Michif Phonology*. PhD thesis, University of Toronto.

Rosendall, E. P. (1998). Aspects of Gbari Grammar. Master's thesis, University of Texas at Arlington.

Rosenfelder, I., Fruehwald, J., Evanini, K., and Yuan, J. (2011). FAVE (Forced Alignment and Vowel Extraction) Program Suite. Online: http://fave.ling.upenn.edu.

Ross, M. (1980). Some Elements of Vanimo, a New Guinea Tone Language. In *Papers in New Guinea Linguistics 20*, volume 56 of *Pacific Linguistics, Series A*, pages 77–109. Australian National University, Canberra.

Rottland, F. (1977). Reflexes of Proto-Bantu Phonemes in Yanzi (B85). *Africana Linguistica*, 7:375–390.

Rousselot, P. J. (1897). *Principes De Phonétique Experimentale*. Paris: H. Welter.

Rowe, K. (2005). *Siar-Lak Grammar Essentials*, volume 50 of *Data Papers on Papua New Guinea Languages*. SIL Ukarumpa.

Rubino, C. R. G. (1997). *A Reference Grammar of Ilocano*. PhD thesis, University of California, Santa Barbara.

Ruhlen, M. (1973). *Rumanian Phonology*. PhD thesis, Stanford University.

Ruhlen, M. (1975). *A Guide to the Languages of the World*. Language Universals Project, Stanford University.

Ruhlen, M. (1987). *A Guide to the World's Languages. Volume 1: Classification*. Edward Arnold, London, UK.

Rupp, J. (1980). *Chinanteco de San Juan Lealao, Oaxaca*. Archivo de lenguas indígenas de México. Centro de Investigación para la Integración Social, Mexico, México.

Rupp, J. (1983). *Huave de San Mateo del Mar*. Archivo de lenguas indígenas de México. Centro de Investigación para la Integración Social, Mexico, México.

Rycroft, D. K. and Ngcobe, A. B. (1979). *Say It in Zulu*. School of Oriental and African Studies, London.

Sachnine, M. (1982). *Le lamé: Un parler zime du Nord-Cameroun (langue tchadique): phonologie-grammaire*. Societe d'Etudes Linguistiques et Anthropologiques de France and l'Agence de Cooperation Culturelle et Technique, Paris.

Sagey, E. (1986). *The Representation of Features and Relations in Non-Linear Phonology*. PhD thesis, Massachusetts Institute of Technology.

Sagey, E. (1990). *The Representation of Features in Nonlinear Phonology: The Articulator Node Hierarchy*. Garland, New York, NY.

Saint, R. and Pike, K. L. (1962). Auca Phonemics. In Elson, B., editor, *Studies in Ecuadorian Indian Languages*, volume 1, pages 2–30. Summer Institute of Linguistics, University of Oklahoma, Norman.

Sakel, J. (2004). *A Grammar of Mosetén*. Number 33 in Mouton Grammar Library. Mouton de Gruyter.

Salser, J. K. (1971). Cubeo Phonemics. *Linguistics*, 75:74–79.

Samarin, W. J. (1966). *The Gbeya Language: Grammar, Texts and Vocabularies*, volume 44 of *University of California Publications in Linguistics*. University of California Press, Berkeley.

Samarin, W. J. (1967a). *A Grammar of Sango*. Mouton, The Hague.

Samarin, W. J. (1967b). *Basic Course in Sango*. Grace Theological Seminary and College, Winona Lake, IN.

Sampson, D. (1985a). The Phonology of Banda-Tangbago. *Studies in African Linguistics*, 9:269–274.

Sampson, G. (1985b). *Writing Systems*. Stanford University Press, Stanford, CA.

Sampson, G., Gil, D., and Trudgill, P., editors (2009). *Language Complexity as an Evolving Variable*. Oxford University Press.

Sandalo, M. F. (1995). *A Grammar of Kadiweu*. PhD thesis, University of Pittsburgh.

Sanders, A. and Sanders, J. (1994). *Kamasau (Wand Tuan) Grammar*. Summer Institute of Linguistics.

Sands, A. K. (1989). A Grammar of Garadjari, Western Australia. B.A. Thesis.

Sanou, D. J.-F. (1978). *La langue bobo de tondogosso*. PhD thesis, Universite Paris V.

Santandrea, S. (1976). *The Kresh Group: Aja and Baka Languages (Sudan)*. Istituto Universitario Orientale, Napoli.

Santos, R. (1977). *Phonologie et Morphotonologie de la Langue Wey (Konyagi)*, volume 69 of *Les Langues Africaines au Senegal*. Centre de Linguistique Appliquée de Dakar, Dakar.

Sapir, E. (1912). Language and Environment. *American Anthropologist*, 14:226–242.

Sapir, E. (1923). The Phonetics of Haida. *International Journal of American Linguistics*, 3-4:143–58.

Sapir, E. (1925). Sound Patterns in Language. *Language*, 1(2):37–51.

Sapir, E. and Hoijer, H. (1967). *The Phonology and Morphology of the Navaho Language*, volume 50 of *University of California Publications in Linguistics*. University of California Press, Berkeley.

Sapir, E. and Swadesh, M. (1939). *Nootka Texts: Tales and Ethnological Narratives, With Grammatical Notes and Lexical Materials*. William Dwight Whitney Linguistic Series. Linguistic Society of America, Philadelphia.

Sapir, E. and Swadesh, M. (1955). Native Accounts of Nootka Ethnography. *International Journal of American Linguistics*, 21(4):1–8.

Sapir, E. and Swadesh, M. (1960). *Yana Dictionary*, volume 22 of *University of California Publications in Linguistics*. University of California Press, Berkeley.

Sapir, J. D. (1965). *A Grammar of Diola-Fogny*, volume 3 of *West African Language Monographs*. Cambridge University Press, in association with The West African Languages Survey and The Institute of African Studies, Ibadan.

Saporta, S. and Contreras, H. (1962). *A Phonological Grammar of Spanish*. University of Washington Press, Seattle.

Sastry, J. V. (1972). *Telugu Phonetic Reader*. Central Institute of Indian Languages, Mysore.

Sat, S. C. (1966). Tuvinskij jazyk. In Baskakov, N. A., editor, *Jazyki narodov SSSR. Volume 2: Tjurkskie jazyki*, pages 387–402. Nauka, Moscow and Leningrad.

Sauvageot, S. (1965). *Description synchronique d'un dialecte wolof: le parler du dyolof*, volume 73 of *Mémoires de l'Institut Fondamental d'Afrique Noire*. Institut Français de l'Afrique Noire, Dakar.

Sawyer, J. O. (1965). *English–Wappo Vocabulary*. University of California Press, Berkeley.

Saxton, D. (1963). Papago Phonemes. *International Journal of American Linguistics*, 29:29–35.

Sayers, B. J. and Godfrey, M. (1964). Outline Description of the Alphabet and Grammar of a Dialect of Wik-Munkan Spoken at Coen, North Queensland. In Oates, W. J., editor, *Gugu-Yalanji and Wik-Munkan Language Studies*, Occasional Papers in Aboriginal Studies, Canberra, pages 49–78. Australian Institute of Aboriginal Studies, Canberra.

Scatton, E. A. (1984). *A Reference Grammar of Modern Bulgarian.* Slavica Publishers, Columbus, Ohio.

Schachter, P. and Fromkin, V. (1968). A Phonology of Akan: Akuapem, Asante, Fante. In *UCLA Working Papers in Phonetics*, volume 9, Los Angeles. Phonetics Laboratory, University of California.

Schachter, P. and Otanes, F. T. (1972). *Tagalog Reference Grammar.* University of California Press, Berkeley. Reprinted in 1983.

Schadeberg, T. C. (1981a). *A Survey of Kordofanian. Volume 1: The Heiban Group.* Helmut Buske, Hamburg.

Schadeberg, T. C. (1981b). *A Survey of Kordofanian. Volume 2: The Talodi Group.* Helmut Buske, Hamburg.

Schaefer, R. (1975). Frafra. In Kropp-Dakubu, M. E., editor, *West African Langauge Data Sheets*, volume 2, pages 42–47. West African Linguistics Society.

Schauer, S. and Schauer, J. G. (1967). Yucuna Phonemics. In Waterhouse, V., editor, *Phonemic Systems of Colombian languages*, volume 14 of *Summer Institute of Linguistics Publications in Linguistics and Related Fields*, pages 61–71. Summer Institute of Linguistics of the University of Oklahoma, Norman.

Schlegel, S. A. (1971). *Tiruray-English Lexicon.* University of California Press, Berkeley.

Schlicter, M.-A. (1985). *The Yukian Language Family.* PhD thesis, University of California, Berkeley.

Schoenhals, A. and Schoenhals, L. C. (1965). *Vocabulario Mixe De Totontepec.* Instituto Lingüístico de Verano, México.

Schuh, R. G. (1972). *Aspects of Ngizim Syntax.* PhD thesis, University of California at Los Angeles.

Schuh, R. G. (1995). Aspects of Avatime Phonology. *Studies in African Linguistics*, 24(1):31–67.

Schulze, C. and Stauffer, D. (2005). Monte Carlo Simulation of the Rise and the Fall of Languages. *International Journal of Modern Physics C*, 16(5):781–787.

Schulze, C., Stauffer, D., and Wichmann, S. (2008). Birth, Survival and Death of Languages by Monte Carlo Simulation. *Communications in Computational Physics*, 3(2):271–294.

Schumann, O. and Garcia de Leon, A. (1966). El dialecto nahual de Almomoloa. *Tlalocan*, 5:178–192.

Schutz, A. J. (1981). A Reanalysis of the Hawaiian Vowel System. *Oceanic Linguistics*, 20(1):1–43.

Scobbie, J. (2002). Flexibility in the Face of Incompatible English VOT Systems. In *Paper Presented at Eighth Conference on Laboratory Phonology (LABPHON 8).*

Scorza, D. (1985). A Sketch of Au Morphology and Syntax. In *Papers in New Guinea Linguistics*, volume 22, pages 215–273.

Scott, N. C. (1957). Notes on the Pronunciation of Sea Dayak. *Bulletin of the School of Oriental and African Studies*, 20(1):509–512.

Scripture, E. W. (1902). *The Elements of Experimental Phonetics.* New York, NY: Charles Scribner's Sons.

Sebeok, T. A. and Ingemann, F. J. (1961). *An Eastern Cheremis Manual*, volume 9 of *Research and Studies in Uralic and Altaic Languages.* Indiana University.

Seglenmej, S. F. (1979). Tuvinskie perednejazychnye v tverdorjadnykh slovoformakh. In Nadeljaev, V. M., editor, *Fonetika Sibirskikh Jazykov*, pages 35–44. Akademija Nauk SSSR, Sibirskoe Otdelenie, Institut Istorii, Filologii i Filosofii, Novosibirsk.

Seiden, W. (1960). Chamorro Phonemes. *Anthropological Linguistics*, 2(4):6–35.

Seiler, W. (1985). *Imonda, a Papuan Language*, volume 93 of *Pacific Linguistics*. The Australian National University.

Selmer, E. W. (1935). *Georgische Experimentalstudien*. Jacob Dybwad, Oslo.

Senn, A. (1966). *Handbuch der litauischen Sprache: Band I: Grammatik*. Carl Winter, Heidelberg.

Shackle, C. (1976). *The Siraiki Language of Central Pakistan: A Reference Grammar*. Unwin Brothers Limited.

Shafeev, D. A. (1964). *A Short Grammatical Outline of Pashto (Translated and Edited by Herbert H. Paper*, volume 33 of *Indiana Research Center in Anthropology, Folklore and Linguistics Publications*. Indiana University, Bloomington.

Shannon, C. E. (1948). A Mathematical Theory of Communication. *Bell System Technical Journal*, 27:379–423, 623–656.

Sharpe, M. C. (1972). *Alawa Phonology and Grammar*, volume 37 of *Australian Aboriginal Studies*. Australian Institute of Aboriginal Studies, Canberra.

Sheldon, S. N. (1974). Some Morphophonemic and Tone Perturbation Rules in Múra-Pirahã. *International Journal of American Linguistics*, 40:279–282.

Sherman, D. (1975). Stop and Fricative Systems: A Discussion of Paradigmatic Gaps and the Question of Language Sampling. In *Working Papers on Language Universals*, volume 17, pages 1–31. Stanford University.

Sherman, D. and Vihman, M. (1972). The Language Universals Phonological Archiving Project: 1971–1972. In *Working Papers on Language Universals*, volume 9, pages 163–179. Stanford University.

Sherzer, J. (1983). *Kuna Ways of Speaking: An Ethnographic Perspective.* University of Texas Press, Austin.

Shibatani, M. (1990). *The Languages of Japan.* Cambridge Language Surveys. Cambridge University Press, Cambridge.

Shimizu, K. (1971). The Kente Dialect of Kpan. *Research Notes Department of Linguistics and Nigerian Languages, University of Ibadan*, 4(2–3):1–36.

Shimizu, K. (1983). *The Zing Dialect of Mumuye: A Descriptive Grammar.* Helmut Buske Verlag, Hamburg.

Shipley, W. F. (1956). The Phonemes of Northeastern Maidu. *International Journal of American Linguistics*, 22:233–237.

Shipley, W. F. (1964). *Maidu Grammar.* University of California Press, Berkeley.

Shklovsky, K. (2005). *Person Marking in Petalcingo Tzeltal.* PhD thesis, Reed College.

Shosted, R. K. (2006). Correlating Complexity: A Typological Approach. *Linguistic Typology*, 10:1–40.

Sievers, E. (1876). *Grundzüge der Lautphysiologie zur Einfuhrung in das Studium der Lautlehere der Indogermanischen Sprachen.* Leipzig: Breitkopf and Hartel.

Silver, S. (1964). Shasta and Karok: A Binary Comparison. In Bright, W., editor, *Studies in California Linguistics*, pages 170–181. University of California Press, Berkeley / Los Angeles.

Silverman, K., Beckman, M., Pitrelli, J., Ostendorf, M., Wightman, C., Price, P., Pierrehumbert, J., and Hirschberg, J. (1992). ToBI: A Standard for Labeling English Prosody. In *Second International Conference on Spoken Language Processing. ISCA.*

Simeon, G. (1969). Hokkaido Ainu Phonemics. *Journal of the American Oriental Society*, 89:751–757.

Simons, G. F. (1989). Working With Special Characters. In *Occasional Publications in Academic Computing*, Occasional Publications in Academic Computing. Summer Institute of Linguistics, Dallas, TX.

Simpson, A. P. (1999). Fundamental Problems in Comparative Phonetics and Phonology. Does UPSID Help to Solve Them? In *Proceedings of the XIVth ICPhS*, San Francisco, CA.

Singler, J. V. (1979). The Segmental Phonology of Verb Suffixes in Talo Klao (Kru). Master's thesis, University of California at Los Angeles.

Singnoi, U. (1988a). A Comparative Study of Pray and Mal Phonology. Master's thesis, Mahidol University.

Singnoi, U. (1988b). A Comparative Study of Pray and Mal Phonology. Master's thesis, Mahidol University.

Sinnemaeki, K. (2011). *Language Universals and Linguistic Complexity: Three Case Studies in Core Argument Marking*. PhD thesis, University of Helsinki.

Sirin, E., Parsia, B., Grau, B. C., Kalyanpur, A., and Katz, Y. (2007). Pellet: A Practical OWL-DL Reasoner. *Web Semantics: Science, Services and Agents on the World Wide Web*, 5:51–53.

Sischo, W. R. (1979). Michoacán Nahual. In *Modern Aztec Grammatical Sketches*, volume 2 of *Studies in Uto-Aztecan Grammar*. Summer Institute of Linguistics, 56 edition.

Sivertsen, E. (1956). Pitch Problems in Kiowa. *International Journal of American Linguistics*, 22:117–130.

Sjoberg, A. F. (1962). The Phonology of Standard Uzbek. In Poppe, N., editor, *American Studies in Altaic Linguistics*, volume 13 of *Uralic and Altaic Series*, pages 237–262. Indiana University Press, Bloomington.

Sjoberg, A. F. (1963). *Uzbek Structural Grammar*, volume 18 of *Uralic and Altaic Series*. Indiana University Press, Bloomington.

Skorik, P. I. (1968). Chukotskij jazyk. In Vinogradov, V. V., editor, *Jazyki Narodov SSSR. Volume 5: Mongol'skie, tunguso-man'chzhurskie i paleoaziaskie jazyki*, pages 248–270. Nauka, Moscow / Leningrad.

Skorik, P. J. (1961). *Grammatika chukotskogo jazyka. Chast' pervaja.* Akademija Nauk SSSR, Moscow / Leningrad.

Smith, J. and Weston, P. (1974). Mianmin Phonemes and Tonemes. In Loving, R., editor, *Workpapers in Papua New Guinea Languages: Studies in Languages of the OK Family*, volume 7, pages 35–142. Summer Institute of Linguistics.

Smith, K. D. (1968). Laryngealization and Delaryngealization in Sedang Phonemics. *Linguistics*, 38:52–69.

Smith, M. K., Welty, C., and McGuinness, D. L. (2004). OWL Web Ontology Language Guide. Technical report, W3C.

Snijders, T. A. and Bosker, R. J. (1999). *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling.* SAGE Publications.

Snyder, W. A. (1968). *Southern Puget Sound Salish: Phonology and Morphology*, volume 8 of *Sacramento Anthropological Society.* Sacramento Anthropology Society, Sacramento State College, Sacramento.

Snyman, J. W. (1970). *An Introduction to the !Xu (!Kung) Language.* A. A. Balkema, Capetown.

Snyman, J. W. (1975). *Zu|'hoasi Fonologie en Woordeboek.* Balkema, Cape Town.

Soderberg, C. D. and Olson, K. S. (2008). Illustrations of the IPA: Indonesian. *Journal of the International Phonetic Association*, 38(2):209–213.

Sommer, B. A. (1969). *Kunjen Phonology: Synchronic and Diachronic*, volume 11 of *Pacific Linguistics, Series B.* Australian National University, Canberra.

Sommer, G. (2003). Western Savanna (K,R). In Nurse, D. and Philippson, G., editors, *The Bantu Languages*, pages 529–580. Routledge.

Sommerfelt, A. (1964). Consonant Clusters or Single Phonemes in Northern Irish. In Abercrombie, D., editor, *In Memory of Daniel Jones*, pages 368–373. Longmans, Green and Co Ltd, London.

Song, Z. (1982). Woguo Tuwayu Yinxi Chutan. [A Preliminary Study of the Phonological System of the Tuva Language in China]. *Minzu Yuwen*, 6:58–65.

Sonnenschein, A. H. (2004). *A Descriptive Grammar of San Bartolomé Zoogocho Zapotec.* PhD thesis, University of Southern California.

Sowa, J. F. (2000). *Knowledge Representation.* Brookes/Cole.

Spaulding, C. and Spaulding, P. (1994). *Phonology and Grammar of Nankina*, volume 41 of *Data Papers on Papua New Guinea Languages.* Summer Institute of Linguistics.

Spotts, H. (1953). Vowel Harmony and Consonant Sequences in Mazahua (Otomi). *International Journal of American Linguistics*, 19:253–258.

Sproat, R. (2000). *A Computational Theory of Writing Systems.* Cambridge University Press, Cambridge, UK.

Stahlke, H. (1970). *Topics in Ewe Phonology.* PhD thesis, University of California at Los Angeles.

Stairs Kreger, G. A. and de Stairs, E. F. S. (1981). *Diccionario Huave de San Mateo del Mar*, volume 24 of *Vocabularios y Diccionarios Indígenas "Mariano Silva y Aceves" volume.* Instituto Lingüístico de Verano, México.

Stanford, R. and Lyn (1970). *Collected Field Reports on the Phonology and Grammar of Chakosi.* Institute of African Studies, University of Ghana.

Steinitz, W. (1950). *Geschichte des Ostjakischen Vokalismus.* Akademic-Verlag, Berlin.

Stell, N. N. (1972). *Fonologia de la lengua Axluxlaj*, volume 8 of *Cuadernos de Linguistica Indigena.* Centro de Estudios Linguisticos, University of Buenos Aires, Buenos Aires.

Steltenpool, J. (1969). *Ekagi-Dutch-English-Indonesian Dictionary.* Koninklijk Instituut voor Taal-, Land- en Volkenkunde. Martinus Nijhoff, The Hague.

Sten, H. (1963). *Manuel de Phonetique Francaise.* Munksgaard, Copenhagen.

Stennes, L. H. (1967). *A Reference Grammar of Adamawa Fulani.* African Studies Center, Michigan State University.

Stenzel, K. S. (2004). *A Reference Grammar of Wanano.* PhD thesis, University of Colorado.

Stevens, J. P. (2009). *Applied Multivariate Statistics for the Social Sciences.* Routledge, New York, NY.

Stevenson, R. C. (1957). A Survey of the Phonetics and Grammatical Structure of the Nuba Mountain Languages, With Particular Reference to Otoro, Katcha and Nyimang. *Afrika und Übersee*, 40, 41:73–84, 93–115, 117–152, 171–196.

Stewart, J. M. (1967). Tongue Root Position in Akan Vowel Harmony. *Phonetica*, 16:185–204.

Stokhof, W. A. L. (1979). *Woisika 2: Phonemics*, volume 59 of *Pacific Linguistics, Series B.* Australian National University, Canberra.

Stolte, J. and Stolte, N. (1971). A Description of Northern Barasano Phonology. *Linguistics*, 75:86–92.

Story, G. L. and Naish, C. M. (1973). *Tlingit Verb Dictionary.* Alaska Native Language Center, University of Alaska, Fairbanks.

Straight, H. S. (1976). *The Acquisition of Maya Phonology, Variation in Yucatec Child Language.* Garland Publishing, Inc., New York / London.

Street, C. S. and Mollinjin, G. P. (1981). The Phonology of Murinbata. In Waters, B., editor, *Australian Phonologies: Collected Papers*, volume 5 of *Work Papers of SIL-AAB, Series A*, pages 183–244. Summer Institute of Linguistics, Australian Aborigines Branch, Darwin.

Street, J. C. (1963). *Khalkha Structure*, volume 24 of *Uralic and Altaic Series*. Indiana University Press, Bloomington.

Suárez, J. A. (1975). *Estudios huaves*, volume 22 of *Colección científica Lingüística*. Instituto Nacional de Antropología e Historia, México.

Suárez, J. A. (1983). *La lengua tlapaneca de Malinaltepec*. Universidad Nacional Autónoma de México, México.

Sumner, C. (1957). *Étude expérimentale de l'amharique moderne: d'après la prononciation d'Abraha François*. University College Press, Addis Ababa.

Susnik, B. J. (1974). *Estudios Guayaki: Sistema fonetico y tematico*. Museo Etnografico "Andres Barbero", Asunción, Paraguay.

Ŝuŝtarŝiĉ, R., Komar, S., and Petek, B. (1995). Illustrations of the IPA: Slovene. *Journal of the International Phonetic Association*, 25(2):86–90.

Svantesson, J.-O. (1983). *Kammu Phonology and Morphology*, volume 18 of *Travaux de L'Institut de Linguistique de Lund*. CWK Gleerup, Malmö.

Svantesson, J.-O. (1985). Vowel Harmony Shift in Mongolian. *Lingua*, 67:283–327.

Swadesh, M. (1971). *The Origin and Diversification of Language. Edited Post Mortem by Joel Sherzer*. Aldine, Chicago, IL.

Swanton, J. R. (1909). *Tlingit Myths and Texts*, volume 39 of *Bureau of American Ethnology Bulletin*. U.S. Government Printing Office, Washington, D.C.

Swanton, J. R. (1911). Tlingit. In Boas, F., editor, *Handbook of American Indian Languages 1*, volume 40 of *Bureau of American Ethnology Bulletin*, pages 159–204. Smithsonian Institution, Washington, D.C.

Sweet, H. (1881). Sound Notation. *Transactions of the Philological Society*, pages 177–235.

Swift, L. B. (1963). *A Reference Grammar of Modern Turkish*. Indiana University Press, Bloomington.

Swift, L. B., Ahaghotu, A., and Ugorji, E. (1962). *Igbo Basic Course.* Foreign Service Institute, US Department of State, Washington, D.C.

Swiri, R. A. (1998). Lexical Expansion in the Mankon Language. Master's thesis, University of Yaounde I.

Takizala, A. (1974). *Studies in the Grammar of Kihungan.* PhD thesis, University of California, San Diego.

Tanenbaum, A. S. (2003). *Computer Networks.* Prentice Hall.

Tantiwithipakorn, W. (1998). A Phonological Study Wa at Ban Santisuk Moo 19, Tambol Patung, Mae-Chan District Chiengrai Province. Master's thesis, Mahidol University.

Tarpent, M.-L. (1987). *A Grammar of the Nisgha Language.* PhD thesis, University of Victoria.

Tataru, A. (1978). *The Pronunciation of Rumanian and English: Two Basic Contrastive Analyses.* Haag and Herchen Verlag, Frankfurt am Main.

Taweeporn, S. (1998). The Phonology of the Nyeu Language. Master's thesis, Mahidol University.

Taylor, A. R. (1969). *A Grammar of Blackfoot.* PhD thesis, The University of California at Berkeley.

Taylor, D. (1955). Phonemes of the Hopkins (British Honduras) Dialect of Island Carib. *International Journal of American Linguistics*, 21(3):233–241.

Taylor, F. W. (1953). *A Grammar of the Adamawa Dialect of the Fulani Language (Fulfulde).* Oxford University Press.

Taylor, F. W. (1959). *A Practical Hausa Grammar*, volume 2. Clarendon Press, Oxford.

Teeter, K. V. (1964). *The Wiyot Language.* University of California Press, Berkeley.

Teferra, A. (1991). A Sketch of Shabo Grammar. In Bender, M. L., editor, *Proceedings of the Fourth Nilo-Saharan Linguistics Colloquium, Nilo-Saharan Linguistic Analyses and Documentation*, volume 7, pages 371–388. Helmut Buske Verlag Hamburg.

Tereshchenko, N. M. (1966a). Neneckij jazyk. In Lytkin, V. I. and Majtinskaja, K. E., editors, *Jazyki narodov SSSR. Volume 3: Finno-ugorskie i samodijskie jazyki*, pages 376–395. Nauka, Moscow / Leningrad.

Tereshchenko, N. M. (1966b). Nganasanskij jazyk. In Lytkin, V. I. and Majtinskaja, K. E., editors, *Jazyki Narodov SSSR. Volume 3: Finno-Ugorskie jazyki i samodijskie jazyki*, pages 416–437. Nauka, Moscow / Leningrad.

Tereshchenko, N. M. (1979). *Nganasanskij Jazyk.* Nauka, Leningrad.

Ternes, E. (1970). *Grammaire Structurale du Breton de l'Ile de Groix.* Carl Winter, Heidelberg.

Terrill, A. (1999). *A Grammar of Lavukaleve: A Papuan Language of the Solomon Islands.* PhD thesis, Australian National University.

Thalbitzer, W. (1904). *A Phonetical Study of the Eskimo Language*, volume 31 of *Meddelelser om Gronland.* Reimer, Copenhagen.

Thanamteun, O. (2000). A Phonological Study of Pa-O (Taungthu) at Ban Huay Salop, Tambon Huay Pha, Muang District, Mae Hon Song Province. Master's thesis, Mahidol University.

The International Phonetic Association (1999). *Handbook of the International Phonetic Association: A Guide to the Use of the International Phonetic Alphabet.* Cam.

The Unicode Consortium (2007). The Unicode Standard, Version 5.0.0, Defined By: The Unicode Standard, Version 5.0.

Thelwall, R. (1981). Lexicostatistical Subgrouping and Lexical Reconstruction of the Daju Group. In Schadeberg, T. and Bender, L., editors, *Nilo-Saharan: Proceedings of the First*

Nilo-Saharan Linguistics Colloquium, Leiden, September 8-10, 1980, pages 167–185. Foris, Dordrecht.

Thieberger, N. A. (2004). *Topics in the Grammar and Documentation of South Efate, an Oceanic Language of Central Vanuatu*. PhD thesis, University of Melbourne.

Thomas, D. D. (1971). Chrau Grammar. In Grace, G. W., McKaughan, H. P., Bender, B. W., and Topping, D. M., editors, *Oceanic Linguistics Special Publications*. University of Hawaii Press.

Thomason, S. (1994). The Editor's Department. *Language*, 70(2):409–413.

Thompson, E. D. (1976). Nera. In Bender, M. L., editor, *The Non-Semitic Languages of Ethiopia*, pages 484–494. African Studies Center, Michigan State University, East Lansing.

Thompson, L. C. (1965). *A Vietnamese Grammar*. University of Washington Press, Seattle.

Thongkum, T. L. (1979). The Distribution of the Sounds of Bruu. *Mon-Khmer Studies*, 8:221–293.

Thongkum, T. L. (1984). *Nyah Kur (Chao Bon)-Thai-English Dictionary*, volume 2 of *Monic Language Studies*. Chulalongkorn University, Linguistic Research Unit, Faculty of Arts, Bangkok.

Thornes, T. J. (2003). *A Northern Paiute Grammar With Texts*. PhD thesis, University of Oregon.

Thurman, R. C. (1970). Chuave Phonemic Statement. Ms.

Timyan, J. (1976). *A Discource-Based Grammar of Baule: The Kode Dialect*. PhD thesis, The City University of New York.

Timyan, J. (1977). *A Discourse-based Grammar of Baule: The Kode Dialect*. PhD thesis, City University of New York.

Tobbler, S. J. (1983). *The Grammar of Karipuna Creole*, volume 10. Summer Institute of Linguistics.

Todaeva, B. X. (1973). *Mongorskij jazyk: issledovanie, teksty, slovar'.* Nauka, Moscow.

Todd, E. M. (1975). The Solomon Language Family. In Wurm, S., editor, *Papuan Languages and the New Guinea Linguistic Scene, New Guinea Area Languages and Language Study 1*, volume 38 of *Pacific Linguistics, Series C*, pages 805–846. Australian National University.

Todd, T. L. (1985). *A Grammar of Dimili (Also Known as Zaza).* PhD thesis, University of Michigan.

Tolsma, G. J. (1999). *A Grammar of Kulung.* PhD thesis, Universiteit te Leiden.

Tomiche, N. (1964). *Le Parler Arabe Du Caire.* Mouton, Paris.

Tomlin, R. S. (1986). *Basic Word Order: Functional Principles.* Croom Helm, London.

Topping, D. M. (1973). *Chamorro Reference Grammar.* University of Hawaii Press, Honolulu.

Topping, D. M. (1980). *Spoken Chamorro.* University Press of Hawaii, Honolulu.

Tracy, F. V. (1972). Wapishana Phonology. In Grimes, J. E., editor, *Languages of the Guianas*, pages 78–84. Summer Institute of Linguistics, University of Oklahoma, Norman.

Trager, F. H. (1971). The Phonology of Picuris. *International Journal of American Linguistics*, 37(1):29–33.

Traill, A. (1985). *Phonetic and Phonological Studies of !Xõo Bushman.* Amsterdam: John Benjamins Publishing Company.

Triulzi, A., Dafallah, A. A., and Bender, M. L. (1976). Berta. In Bender, M. L., editor, *The Non-Semitic Languages of Ethiopia*, pages 513–532. African Studies Center, Michigan State University, East Lansing.

Trnka, B. (1968). *A Phonological Analysis of Present-day Standard English*, volume 17 of *Alabama Linguistic and Philological Series.* University of Alabama Press, Alabama.

Trubetzkoy, N. (1939). *Grundzüge der Phonologie.* Travaux du cercle linguistique de Prague 7.

Trudgill, P. (1974). Linguistic Change and Diffusion: Description and Explanation in Sociolinguistic Dialect Geography. *Language in Society*, 3:215–246.

Trudgill, P. (1996). Dialect Typology: Isolation, Social Network and Phonological Structure. In Guy, G. R., Feagin, C., Schiffrin, D., and Baugh, J., editors, *Towards a Social Science of Language: Papers in Honour of William Labov, Volume 1: Variation and Change in Language and Society*, pages 3–21. Amsterdam: Benjamins.

Trudgill, P. (1997). Typology and Sociolinguistics: Linguistic Structure, Social Structure and Explanatory Comparative Dialectology. *Folia Linguistica*, 31(3–4):349–360.

Trudgill, P. (2002). Linguistic and Social Typology. In Chambers, J. K., Trudgill, P., and Schilling-Estes, N., editors, *The Handbook of Language Variation and Change*. Blackwell Publishers, Oxford, UK.

Trudgill, P. (2004a). Linguistic and Social Typology: The Austronesian Migrations and Phoneme Inventories. *Linguistic Typology*, 8(3):305–320.

Trudgill, P. (2004b). On the Complexity of Simplifcation. *Linguistic Typology*, 8(3):384–288.

Tryon, D. T. (1968). *Iai Grammar*, volume 8 of *Pacific Linguistics, Series B*. Australian National University, Canberra.

Tryon, D. T. (1970). *An Introduction to Maranungku (Northern Australia)*, volume 15 of *Pacific Linguistics, Series B*. The Australian National University, Canberra.

Tryon, D. T. (1974). *Daly Family Languages*, volume 32 of *Pacific Linguistics, Series C*. Australian National University, Canberra.

Tschenkéli, K. (1958). *Einführung in die georgische Sprache*. Amirami Verlag, Zürich.

Tucker, A. N. (1967). *The Eastern Sudanic Languages*. International African Instutite, London.

Tucker, A. N., Bryan, M., and Woodburn, J. (1977). The East African Click Languages: A Phonetic Comparison. In Möhlig, W. J. G., Rottland, F., and Heine, B., editors, *Zur*

Sprachgeschichte und Ethnohistorie in Afrika, Neue Beiträge afrikanistischer Forschungen, pages 301–323. Dietrich Reimer Verlag, Berlin.

Tucker, A. N. and Bryan, M. A. (1966). *Linguistic Analyses: The Non-Bantu Languages of North-Eastern Africa*. Oxford University Press, London.

Tucker, A. N. and Hackett, P. E. (1959). *Le groupe linguistique Zande*, volume 8 of *Annales du Musee Royal de l'Afrique Centrale*. Musee Royale de l'Afrique Centrale, Tervuren.

Tucker, A. N. and Mpaayei, J. T. O. (1955). *A Maasai Grammar With Vocabulary*, volume 2 of *Publications of the African Institute, Leyden*. Longmans, Green & Co, London.

Tucker, A. N. and Tucker, M. A. (1966). *Linguistic Analyses: The Non-Bantu Languages of North-Eastern Africa*. Oxford University Press, London.

Tung, T.-H. (1964). *A Descriptive Study of the Tsou Language, Formosa*, volume 48 of *Special Publications*. Institute of History and Philology, Academia Sinica, Taipei.

Turner, B. (1986). *A Teaching Grammar of the Manam Language*. Number 34 in On Papua New Guinea Languages. Summer Institute of Linguistics: Ukarumpa, Papua New Guinea.

Turton, D. and Bender, M. L. (1976). Mursi. In Bender, M. L., editor, *The Non-Semitic Languages of Ethiopia*, volume 5 of *Committee on Ethiopian Studies: Occasional Papers Series*, pages 533–561. African Studies Center, Michigan State University, East Lansing, Michigan.

Tuttle, S. G. and Sandoval, M. (2002). Jicarilla Apache. *Journal of the International Phonetic Association*, 32(1):105–112.

Twaddell, W. F. (1935). On Defining the Phoneme. *Language*, 11(1):5–62.

Tyler, S. A. (1969). *Koya: An Outline Grammar*, volume 54 of *University of California Publications in Linguistics*. University of California Press, Berkeley.

Ubrjatova, E. I. (1966). Jakutskij jazyk. In Vinogradov, V. V., editor, *Jazyki narodov SSSR. Volume 2: tjurkskie jazyki*, pages 403–427. Nauka, Moscow / Leningrad.

Uhlenbeck, E. M. (1949). *De structuur van het Javaanse morpheem.* PhD thesis, University of Leiden, Bandoeng.

Uhlenbeck, E. M. (1963). Review of E.C. Horne "Beginning Javanese". *Lingua*, 12:69–76.

Uldall, E. (1956). Guarani Sound System. *International Journal of American Linguistics*, 20:341–342.

Underhill, R. (1976). *Turkish Grammar.* MIT Press, Cambridge, Mass.

Ungsitipoonporn, S. (2001). A Phonological Comparison Between Khlongphlu Chong and Wangkraphrae Chong. Master's thesis, Mahidol University.

Urquía Sebastián, R. and Marlett, S. A. (2008). Illustrations of the IPA: Yine. *Journal of the International Phonetic Association*, 38(3):365–369.

Urua, E.-A. E. (2004). Illustrations of the IPA: Ibibio. *Journal of the International Phonetic Association*, 34(1):105–109.

van den Berg, R. (1989). *A Grammar of the Muna Language.* Foris Publications.

van der Laan, M. and Rose, S. (2010). Statistics Ready for a Revolution: Next Generation of Statisticians Must Build Tools for Massive Data Sets. *Amstat News.*

van der Stap, P. A. M. (1966). *Outline of Dani Morphology*, volume 48 of *Verhandelingen van het Koninklijk Instituut voor Taal, Land en Volkenkunde.* Nijhoff, The Hague.

van der Tuuk, H. N. (1971). *A Grammar of Toba Batak.* Martinus Nijhoff, The Hague. Reprint of 1864.

van der Voort, H. (2004). *A Grammar of Kwaza.* Number 29 in Mouton Grammar Library. Mouton de Gruyter.

van Gijn, E. (2006). *A Grammar of Yurakarè.* PhD thesis, Radboud Universiteit Nijmegen.

Van Syoc, W. B. (1959). *The Phonology and Morphology of the Sundanese Language.* PhD thesis, University of Michigan.

Van Wynen, D. and de Van Wynen, M. G. (1962). *Tacana y Castellano*, volume 2 of *Vocabularios Bolivianos*. Instituto Lingüístico de Verano, Cochabamba.

Vanvik, A. (1972a). A Phonetic-phonemic Analysis of Standard Eastern Norwegian. *Norwegian Journal of Linguistics*, 26(2):119–164.

Vanvik, A. (1972b). A Phonetic-phonemic Analysis of Standard Eastern Norwegian. *Norwegian Journal of Linguistics*, 27(2):11–139.

Vaux, B. (2009). The Role of Features in a Symbolic Theory of Phonology. In Raimy, E. and Cairns, C. E., editors, *Contemporary Views on Architecture and Representations in Phonology*. MIT Press.

Vaux, B. and Samuels, B. (2005). Aspiration and Laryngeal Markedness. *Phonology*, 23:395–436.

Veena, C. (1980). A Description of Moken: A Malay-Polynesian Language. Master's thesis, Mahidol University.

Verguin, J. (1967). *Le Malais*. Mouton, The Hague.

Verheijen, J. A. J. (1986). *The Sama/Bajau Language in the Lesser Sunda Islands*, volume 70 of *Pacific Linguistics, Series D*. Australian National University, Canberra.

Verhoeven, J. (2005). Illustrations of the IPA: Belgian Standard Dutch. *Journal of the International Phonetic Association*, 35(2):243–247.

Vermeer, H. J. and Sharma, A. (1966). *Hindi-Lautlehre mit Einführung in die Devnagari-Schrift*. Julius Groos Verlag, Heidelberg.

Vidal, A. (2001a). *Pilagá Grammar (Guaykuruan Family, Argentina)*. PhD thesis, University of Oregon.

Vidal, A. (2001b). *Pilagá Grammar (Guaykuruan Family, Argentina)*. PhD thesis, University of Oregon.

Vihman, M. (1974). Excerpts From the Phonology Archive Coding Manual (Preliminary Edition). In *Working Papers on Language Universals*, volume 15, pages 141–153. Stanford University.

Voegelin, C. and Voegelin, F. (1966). Languages of the World: Indo-Pacific Fascicle 8. *Anthropological Linguistics*, 8(4):10–14.

Voegelin, C. F. (1946). Delaware, an Eastern Algonquian Language. In Hoijer, H., editor, *Linguistic Structures of Native America*, volume 6 of *Viking Fund Publication in Anthropology*, pages 130–157. Wenner-Gren Foundation.

Voegelin, C. F. (1956). Phonemicizing for Dialect Study With Reference to Hopi. *Language*, 32:116–135.

Voegelin, C. F. and Voegelin, F. M. (1977). *Classification and Index of the World's Languages*. Elsevier, New York, NY.

Vogel, A. R. (2003). *Jarawara Verb Classes*. PhD thesis, University of Pittsburgh.

Vogt, H. (1938). Esquisse d'une grammaire du georgien moderne. *Norsk Tidsskrift for Sprogvidenskap*, 9:5–114.

Vogt, H. (1939). Alternances vocaliques en géorgien. *Norsk Tidsskrift for Sprogvidenskap*, 11:118–135.

Vogt, H. (1958). Structure phonemique du georgien. *Norsk Tidsskrift for Sprogvidenskap*, 18:5–90.

Vogt, H. (1971). *Grammaire de la langue géorgienne*. The Institute for Comparative Research in Human Culture, Oslo.

Volk, E. (2011). *Mijikenda Tonology*. PhD thesis, Tel Aviv University.

Volodin, A. P. (1976). *Itel'menskij jazyk*. Nauka, Leningrad.

Voorhoeve, C. (1985). Some Notes on the Arandai Language, South Bird's Head, Irian Jaya. *Irian*, 13:3–40.

Voorhoeve, C. L. (1965). *The Flamingo Bay Dialect of the Asmat Language*, volume 46 of *Verhandelingen van het Koninklijk Instituut voor Taal, Land en Volkenkunde.* Martinus Nijhoff, The Hague.

Voorhoeve, C. L. (1971). Miscellaneous Notes on Languages in West Irian, New Guinea. In et al., T. D., editor, *Papers in New Guinea Linguistics 14*, volume 28 of *Pacific Linguistics, Series A*, pages 47–114. Australian National University, Canberra.

Voorhoeve, C. L. (1982). The West Makian Language, North Moluccas, Indonesia: A Field Report. In Voorhoeve, C. L., editor, *The Makian Languages and Their Neighbours*, volume 46 of *Pacific Linguistics, Series D*, pages 1–74. Australian National University, Canberra.

Voorhoeve, C. L. (1989). Notes on Irarutu (An Austronesian Language Spoken in the Centre of the Bomberai Peninsula, Southwest Irian Jaya). *IRIAN*, 17:107–119.

Voutsa, L. (2003). Morphologie verbale du ngombale. Master's thesis, University of Yaounde I.

Vries, L. d. (1992). *The Morphology of Wambon of the Irian Jaya Upper-Digul Area.* KITLV Press.

Walker, W. (1972). Toward the Sound Pattern of Zuni. *International Journal of American Linguistics*, 38(4):240–259.

Walker, W. (1975). Cherokee. In Crawford, J. M., editor, *Studies in Southestern Indian Languages*, pages 189–236. University of Georgia Press, Athens.

Walton, J. and Walton, J. (1967). Phonemes of Muinane. In Waterhouse, V. G., editor, *Phonemic Systems of Colombian Languages*, pages 37–47. Summer Institute of Linguistics, University of Oklahoma, Norman.

Wang, F. (1983). Miaoyu Fangyan Huafen Wenti (On the Division of Miao Dialects). *Minzu Yuwen*, 5:1–22.

Wang, F. (1985). *Miaoyu Jianzhi. Brief Guide to Miao Language.* Minzu Chubanshe, Beijing.

Wangler, H. H. (1972). *Physiologische Phonetik. Eine Einführung.* N. G. Elwert Verlag, Marburg.

Wannemacher, M. W. (1998). *Aspects of Zaiwa Phonology: An Autosegmental Account.* Number 129 in Publications in Linguistics. Summer Institute of Linguistics and The University of Texas at Arlington.

Ward, I. C. (1936). *An Introduction to the Ibo Language.* W. Heffer and Sons, Cambridge, U. K.

Ward, I. C. (1963). A Short Phonetic Study of Wolof (Jolof) as Spoken in the Gambia and in Senegal. In Manessy, G. and Sauvageot, S., editors, *Wolof et sérèr; études de phonétique et de grammaire descriptive*, volume 12 of *Publications de la Section de langues et littératures*, pages 57–63. Université de Dakar, Faculté des lettres et sciences humaines, Dakar.

Warren, D. M. (N.D.). *Vocabulary of the Akan (Twi-Fante) Language of Ghana.* Number 6 in Indiana University Publications, African Series. Indiana University Press, Bloomington.

Wash, S. (2001). *Adverbial Clauses in Barbareño Chumash Narrative Discourse.* PhD thesis, The University of California at Santa Barbara.

Watkins, L. J. and McKenzie, P. (1984). *A Grammar of Kiowa.* Studies in the Anthropology of North American Indians. University of Nebraska Press, Lincoln / London.

Watson, K. (2007). Illustrations of the IPA: English, Liverpool. *Journal of the International Phonetic Association*, 37(3):351–360.

Watson, R. (1964). Pacoh Phonemes. *Mon-Khmer Studies*, 1:135–148.

Watt, D. and Allen, W. (2003). Illustrations of the IPA: English, Tyneside. *Journal of the International Phonetic Association*, 33(2):267–271.

Watters, D. E. (2006). Notes on Kusunda Grammar: A Language Isolate of Nepal. *Himalayan Linguistics Archive*, 3:1–182.

Watters, J. K. (1988). *Topics in Tepehua Grammar*. PhD thesis, University of California, Berkeley.

Watters, J. R. (1981). *A Phonology and Morphology of Ejagham - With Notes on Dialect Variation*. PhD thesis, University of California, Los Angeles.

Watuseke, F. S. (1976). West Makian, a Language of the North-Halmahera Group of the West-Irian Phylum. *Anthropological Linguistics*, 18:274–285.

Webb, T. (1974). Urii Phonemes. In Loving, R., editor, *Phonologies of four Papua New Guinea Languages*, volume 4 of *Workpapers in Papua New Guinea Languages*. Summer Institute of Linguistics.

Weber, D. J. (1983). *A Grammar of Huallaga (Huanuco) Quechua*. PhD thesis, University of California, Los Angeles.

Wedekind, K. (1972). *An Outline of the Grammar of Busa (Nigeria)*. PhD thesis, Christian-Albrechts-Universität zu Kiel.

Weiers, M. (1971). *Die Sprache der Moghol der Provinz Herat in Afghanistan (Sprachmaterial, Grammatik, Wortliste)*. Westdeutscher Verlag, Opladen.

Weimer, H. and Weimer, N. (1972). Yareba Phonemes. *Te Reo*, 15:52–57.

Wellens, I. H. W. (2003). *An Arabic Creole in Africa: The Nubi Language of Uganda*. PhD thesis, Katholieke Universiteit Nijmegen.

Wells, J. C. (n.d.). Computer-coding the IPA: A Proposed Extension of SAMPA. Unpublished Manuscript.

Welmers, W. E. (1946). A Descriptive Grammar of Fanti. *Language*, 22(3):3–78.

Welmers, W. E. (1950). Notes on Two Languages of the Senufo Group. Part 1: Senadi. *Language*, 26:126–146.

Welmers, W. E. (1952). Notes on the Structure of Bariba. *Language*, 28:82–103.

Welmers, W. E. (1962). The Phonology of Kpelle. *Journal of African Languages*, 1:69–93.

Welmers, W. E. (1973). *African Language Structures*. University of California Press, Berkeley / Los Angeles.

Welmers, W. E. (1976). *A Grammar of Vai*, volume 84 of *University of California Publications in Linguistics*. University of California Press.

Wendel, T. D. (1993). A Preliminary Grammar of Hanga Hundi. Master's thesis, University of Texas at Arlington.

Werle, J.-M. and Gbalehi, D. J. (1976). *Phonologie et morphonologie du Bete de la Region de Guiberoua*. Institut de Linguistique Appliquée and Société Internationale de Linguistique, Abidjan.

Werner, O. (1972). *Phonemik des Deutschen*, volume 108 of *Sammlung Metzler*. J.B. Metzler Verlagsbuchhandlung, Stuttgart.

Wheatley, J. (1969). Bakairi Verb Structure. *Linguistics*, 47:80–100.

Wheatley, J. (1973). Pronouns and Nominal Elements in Bacairi Discourse. *Linguistics*, 104:105–115.

Wheeler, A. and Wheeler, M. (1962). Siona Phonemics. In Elson, B., editor, *Studies in Ecuadorian Indian Languages 1*, pages 96–111. Summer Institute of Linguistics and University of Oklahoma, Norman.

Whiteley, W. H. (1958). *A Short Description of Item Categories in Iraqw*. East African Institute of Social Research, Kampala.

Whorf, B. L. (1946). The Hopi Language, Toreva Dialect. In Osgood, C., editor, *Linguistic Structures of Native America*, volume 6 of *Viking Fund Publications in Anthropology*, pages 158–183. Viking Fund Publications, New York.

Wichmann, S. and Holman, E. W. (2009). Population Size and Rates of Language Change. *Human Biology*, 81:259–274.

Wichmann, S., Stauffer, D., Schulze, C., and Holman, E. W. (2008). Do Language Change Rates Depend on Population Size? *Advances in Complex Systems*, 11(3):357–369.

Wichser, M. (1994). *Description Grammaticale Du Kar*. PhD thesis, Ecole Pratique des Hautes Etudes: Sciences Historiques et Philologiques Linguistique Africaine.

Widmann, T. and Bakker, P. (2006). Does Sampling Matter? A Test in Replicability, Concerning Numerals. *Linguistic Typology*, 10:83–95.

Wiering, E. (1974). The Indicative Verb in Doowaayaayo. *Linguistics*, 124:33–56.

Wiesemann, U. (1972). *Die phonologische und grammatische Struktur der Kaingáng-Sprache*. Mouton, The Hague.

Williams, W. L. and Williams, H. W. (1965). *First Lessons in Maori*. Whitcombe and Tombs Limited.

Williamson, K. (1965). *A Grammar of the Kolokuma Dialect of Ijo*, volume 2 of *West African Language Monographs*. Cambridge University Press, Cambridge.

Williamson, K. (1967). Songhai Word List (Gao Dialect). *Department of Linguistics and Nigerian Languages, University of Ibadan, Research Notes*, 3:1–34.

Williamson, K. (1969). Igbo. In Dunstan, E., editor, *Twelve Nigerian Languages: A Handbook on Their Sound Systems for Teachers of English*, pages 85–96. Longmans; Africana Publishing, London.

Williamson, K. (1970). *Reading and Writing Ogbia*. Rivers Readers Project, Institute of African Studies, University of Ibadan.

Williamson, K. (1972). Assimilation in Ogbia. *Department of Linguistics and Nigerian Languages, University of Ibadan, Research Notes*, 5(2–3):1–5.

Williamson, K. and Blench, R. (2000). Niger-Congo. In Heine, B. and Nurse, D., editors, *African Languages: An Introduction*, pages 10–42. Cambridge University Press.

Wilson, D. (1969). Suena Phonology. In *Papers in New Guinea Linguistics 9*, volume 18 of *Pacific Linguistics, Series A*, pages 87–93. Australian National University, Canberra.

Wilson, J. E. (1996). A Phonological Grammar of Kuche. Master's thesis, The University of Texas at Arlington.

Wilson, J. P. (2002). Binandere Verbal Structures.

Wilson, P. R. (1973). Abulas Sentences. In Healy, A., editor, *Workpapers in Papua New Guinea Linguistics: Three studies in sentence structure*, volume 1, pages 21–164. SIL, Ukarumpa.

Wilson, W. A. A. (1961). *An Outline of the Temne Language*. School of Oriental and African Studies, London.

Wilson, W. A. A. and Bendor-Samuel, J. T. (1969). The Phonology of the Nominal in Dagbani. *Linguistics*, 52:56–82.

Wise, M. R. (1958). Diverse Points of Articulation of Allophones in Amuesha (Arawak). *Miscellanea Phonetica*, 3:15–21.

Wolff, H. (1959). Subsystem Typologies and Area Linguistics. *Anthropological Linguistics*, 1(7):1–88.

Wonderly, W. L. (1951). Zoque II: Phonemes and Morphophonemics. *International Journal of American Linguistics*, 17:105–123.

Wongnoppharalert, S. (1993). Nam Sod Khmu Syntactic Structure: A Study in Tagmemics, Transformational and Case Grammar. Master's thesis, Mahidol University.

Woodward, M. F. (1964). Hupa Phonemics. In Bright, W., editor, *Studies in California Linguistics*, pages 199–216. University of California Press, Berkeley / Los Angeles.

Wozna, B. and Wilson, T. (2005). Seimat Grammar Essentials. In *Data Papers on Papua New Guinea Languages*, volume 48. SIL Ukarumpa.

Wright, R. A. (1996). *Consonant Clusters and Cue Preservation in Tsou.* PhD thesis, UCLA.

Wroughton, J. R. (1996). *Gramatica y textos del quechua shausha huanca.* Documento de Trabajo. Instituto Linguistico de Verano, 30 edition.

Wu, C. F. J. (1986). Jackknife, Bootstrap and Other Resampling Methods in Regression Analysis. *The Annals of Statistics*, 14(4):1261–1295.

Wurm, S. A. (1972a). *Languages of Australia and Tasmania.* Mouton, The Hague.

Wurm, S. A. (1972b). Notes on the Indication of Possession With Nouns in Reef and Santa Cruz Islands Languages. In *Papers in Linguistics of Melanesia 3*, volume 35 of *Pacific Linguistics, Series A*, pages 85–113. Australian National University, Canberra.

Wurm, S. A. (1977). Missionary lingue franche: Kiwai. In Wurm, S. A., editor, *Language, Culture, Society, and the Modern World 2*, volume 3 of *New Guinea Area Languages and Language Study.* Australian National University, Canberra.

Yadava, Y. P. (2000). A Study of the Dhangar Language. Preliminary report for the Endangered Language Fund.

Yallop, C. (1977). *Alyawarra: An Aboriginal Language of Central Australia.* Number 10 in Australian Aboriginal Studies: Research and Regional Studies. Advocate Press Pty Ltd.

Yensi, A. M. (1996). The Noun Class System of Oku. Master's thesis, University of Yaounde I.

Yumitani, Y. (1998). *A Phonology and Morphology of Jemez Towa.* PhD thesis, University of Kansas.

Zakhar'in, B. A. (1974). *Problemy Fonologii Jazyka Kashmiri.* Nauka, Moscow.

Zakhar'in, B. A. and Edelman, D. I. (1971). *Jazyk kashmiri.* Jazyki narodov Azii i Afriki. Nauka, Moscow.

Zavala, R. (2000). *Inversion and Other Topics in the Grammar of Olutec (Mixean).* PhD thesis, University of Oregon.

Zewen, P. F. (1987). *Introduction à la langue des îles marquises.* Haere Po no Tahiti.

Zhirkov, L. I. (1936). *Avarsko-russkij slovar' (s kratkim grammaticheskim ocherkom Avarskogo jazyka).* Nauka, Moscow.

Zhirkov, L. I. (1955). *Lakskij jazyk: fonetika i morfologija.* Izdatel'stvo Akademii Nauk SSSR, Moscow.

Zhukova, A. N. (1980). *Jazyk Palanskikh Korjakov.* Nauka, Leningrad.

Zigmond, M. L., Booth, C. G., and Munro, P. (1988). *Kawaiisu: A Grammar and Dictonary With Texts*, volume 119 of *University of California Publications in Linguistics.* University of California Press, Berkeley.

Zinder, L. and Matusevich, M. (1937). Eksperimental'noe issledovanie fonem nivkhskogo iazyka. Appendix to Kreinovich, E.A., 1937, Fonetika nivkhskogo (gilyatskogo) iazyka.

Zuraw, K. (2006). Using the Web as a Phonological Corpus: A Case Study From Tagalog. In *Proceedings of the 2nd International Workshop on Web as Corpus.*

Zwarts, J. (2003). *The Phonology of Endo: A Southern Nilotic Language of Kenya.* Lincom Europa.

Appendix A

GENEALOGICAL COVERAGE OF SEGMENT INVENTORIES IN PHOIBLE

To assess the genealogical representation of inventories in PHOIBLE, I compared PHOIBLE's contents with language families in the Ethnologue 15th edition (Gordon, 2005), as encoded and disseminated through Multitree (LINGUIST List, 2009).[1] There are 114 named groups of languages, either genealogically related (100; e.g. Indo-European, Niger-Congo) or geographically categorized (14; e.g. African Deaf Sign Languages, Central American Language Isolates, European Unclassified Languages). Below I list all groups, their Multitree four-digit family codes, the number of representatives in PHOIBLE, the total number of languages in the language family and PHOIBLE's coverage of each language family. Note that in some cases a language is classified in Multitree under two different root nodes in two different language families. For example, Jamaican Creole [jam] is listed under Central American Pidgins and Creoles [capc], North American Pidgins and Creoles [napc] and European Pidgins and Creoles [eupc]. In this case, if there is an inventory in PHOIBLE, it is counted for each category. A few languages in PHOIBLE are now extinct. They are not counted here if the language is not listed in the Ethnologue.

Language family	Fam code	PHOIBLE	Total	Coverage %
African Deaf Sign Language	adsl	0	23	0.00
African Language Isolates	afis	0	1	0.00
African Unclassified Languages	afun	0	11	0.00
Afro-Asiatic	afas	57	375	15.20
Alacalufan	alac	1	2	50.00
Algic	algi	6	44	13.64
Altaic	altc	15	66	22.73

[1] See Section 4.4.

Language family	Fam code	PHOIBLE	Total	Coverage %
Amto-Musan	amto	0	2	0.00
Andamanese	anda	2	13	15.38
Arauan	arau	1	8	12.50
Araucanian	arcn	1	2	50.00
Arawakan	arwk	14	64	21.88
Arutani-Sape	arus	0	2	0.00
Asian Language Isolates	asis	4	5	80.00
Asian Unclassified Languages	asun	0	10	0.00
Australian	aust	36	263	13.69
Austro-Asiatic	ausa	27	169	15.98
Austronesian	anes	78	1271	6.14
Aymaran	ayma	2	3	66.67
Barbacoan	barb	4	7	57.14
Basque	basq	1	4	25.00
Bayono-Awbono	baya	0	2	0.00
Caddoan	cadd	2	5	40.00
Cahuapanan	cahu	1	2	50.00
Cant	cant	0	1	0.00
Carib	cari	7	32	21.88
Central American Language Isolates	cais	1	1	100.00
Central American Unclassified	caun	0	4	0.00
Chapacura-Wanham	chaw	1	5	20.00
Chibchan	chib	6	22	27.27
Chimakuan	chmn	1	2	50.00
Choco	choc	1	12	8.33
Chon	chon	1	2	50.00
Chukchi-Kamchatkan	chka	3	5	60.00
Chumash	chum	1	7	14.29

Language family	Fam code	PHOIBLE	Total	Coverage %
Coahuiltecan	coah	1	1	100.00
Dravidian	drav	11	73	15.07
East Bird's Head	ebir	0	3	0.00
East Papuan	epap	8	36	22.22
Eskimo-Aleut	eska	4	11	36.36
European Unclassified Languages	euun	0	3	0.00
Geelvink Bay	geba	1	33	3.03
Guahiban	guah	1	5	20.00
Gulf	gulf	1	4	25.00
Harakmbet	hara	0	2	0.00
Hibito-Cholon	hibi	0	2	0.00
Hmong-Mien	hmom	1	35	2.86
Hokan	hoka	11	28	39.29
Huavean	huav	1	4	25.00
Indo-European	ieur	54	450	12.00
Iroquoian	iroq	4	11	36.36
Japanese	japo	1	12	8.33
Jivaroan	jiva	1	4	25.00
Kartvelian	kart	3	5	60.00
Katukinan	katk	0	3	0.00
Keres	kere	1	2	50.00
Khoisan	khoi	4	27	14.81
Kiowa Tanoan	kita	5	6	83.33
Kwomtari-Baibai	kwba	0	6	0.00
Left May	lema	0	6	0.00
Lower Mamberamo	lmam	0	2	0.00
Lule-Vilela	luvi	0	1	0.00
Macro Ge	macg	5	30	16.67

Language family	Fam code	PHOIBLE	Total	Coverage %
Maku	maku	2	6	33.33
Mascoian	masc	0	5	0.00
Mataco-Guaicuru	mgua	6	12	50.00
Mayan	maya	12	70	17.14
Misumalpan	misu	0	4	0.00
Mixe-Zoque	mizo	6	17	35.29
Mura	mura	1	1	100.00
Muskogean	musk	4	6	66.67
Na-Dene	nadn	11	47	23.40
Nambiquaran	namb	2	5	40.00
Niger-Congo	ncon	332	1516	21.90
Nilo-Saharan	nsah	55	204	26.96
North American Language Isolates	nais	2	3	66.67
North American Unclassified Languages	naun	0	2	0.00
North Caucasian	ncau	8	34	23.53
Oto-Manguean	otma	14	174	8.05
Pacific Language Isolates	ocis	0	6	0.00
Pacific Unclassified Languages	paun	0	7	0.00
Panoan	pano	4	28	14.29
Peba-Yaguan	pbya	1	2	50.00
Penutian	penu	9	33	27.27
Quechuan	quch	6	46	13.04
Salishan	sali	10	27	37.04
Salivan	slvn	1	3	33.33
Sepik-Ramu	sepr	10	101	9.90
Sino-Tibetan	sitb	36	411	8.76
Siouan	siou	3	17	17.65
Sko	skoo	3	7	42.86

Language family	Fam code	PHOIBLE	Total	Coverage %
South American Language Isolates	saso	12	24	50.00
South American Unclassified Languages	saun	0	39	0.00
Subtiaba-Tlapanec	subt	1	5	20.00
Tacanan	taca	2	6	33.33
Tai-Kadai	taik	10	87	11.49
Tarascan	tara	1	2	50.00
Torricelli	torr	6	53	11.32
Totonacan	toto	3	11	27.27
Trans-New Guinea	trng	48	565	8.50
Tucanoan	tucn	6	25	24.00
Tupi	tupi	10	76	13.16
Uralic	urlc	9	39	23.08
Uru-Chipaya	urch	0	2	0.00
Uto-Aztecan	utaz	15	61	24.59
Wakashan	waka	2	5	40.00
West Papuan	wpap	3	26	11.54
Witotoan	wita	2	6	33.33
Yanomam	yano	2	4	50.00
Yeniseian	yeos	1	2	50.00
Yukaghir	yuka	2	2	100.00
Yuki	yuki	2	2	100.00
Zamucoan	zamu	0	2	0.00
Zaparoan	zapa	1	7	14.29

Appendix B

LIST OF LANGUAGES, ISO 639-3 CODES AND THEIR SOURCES IN PHOIBLE

This appendix lists each inventory in PHOIBLE by its ISO 639-3 language name identifier, language name (as it is given in the language description or the source database), source (indicating from which database it was taken), and bibliographic citation(s). Bibliographic citations can be looked up in the bibliography. For readers interested in finding out if a particular language is represented in PHOIBLE, I suggest that they first look up the ISO 639-3 language name identifier for that language by searching the online version of the Ethnologue (Lewis, 2009).[1] Then check if the ISO 639-3 code is in the table below. They are listed alphabetically.

ISO 639-3	Language Name	Source	Reference
aal	KOTOKO	UPSID	Bouny 1977
aau	Abau	PHOIBLE	Lock and Lock 1990
abi	Abidji	AA	Hartell 1993; Chanard 2006
abn	Abua	PHOIBLE	Gardner 1966
abt	Abulas	PHOIBLE	Wilson 1973
acd	Gechode	PHOIBLE	Cleal 1973a
ace	Acehnese	PHOIBLE	Asyik 1987
acv	ACHUMAWI	UPSID	Olmsted 1964, 1966
ada	Dangme	AA	Hartell 1993; Chanard 2006
adj	Adioukrou	AA	Hartell 1993; Chanard 2006
adn	Adang	PHOIBLE	Haan 2001
adz	Adzera	SPA	Holzknecht 1973
adz	ADZERA	UPSID	Holzknecht 1973

[1] http://www.ethnologue.com/

ISO 639-3	Language Name	Source	Reference
ael	Ambele	PHOIBLE	Nganganu 2001
aey	AMELE	UPSID	Roberts 1987
afo	Eloyi	PHOIBLE	Mackay 1968
aft	Afitti	PHOIBLE	de Voogt 2009
agm	ANGAATIHA	UPSID	Huisman 1973; Huisman et al. 1981; Huisman and Lloyd 1981
agq	AGHEM	UPSID	Hyman 1979
ags	Esimbi	PHOIBLE	Fointein 1986
ahg	Agaw	PHOIBLE	Hetzron 1969a
ahk	Akha	PHOIBLE	Panadda 1993
ahl	Ahlõ	AA	Hartell 1993; Chanard 2006
ahp	AIZI	UPSID	Herault 1971
aht	AHTNA	UPSID	Kari and Buck 1975; Kari 1979
aig	Antiguan Creole	PHOIBLE	Farquhar 1974
ain	Ainu	SPA	Simeon 1969
ain	AINU	UPSID	Simeon 1969; Patrie 1982
aja	Aja	PHOIBLE	Santandrea 1976
ajg	Adja (Bénin)	AA	Hartell 1993
ajg	Adja (BéNin)	AA	Chanard 2006
ajg	Adja (Togo)	AA	Hartell 1993; Chanard 2006
aka	Akan	SPA	Welmers 1946; Ladefoged 1964; Stewart 1967; Schachter and Fromkin 1968
aka	AKAN	UPSID	Welmers 1946; Ladefoged 1964; Stewart 1967; Schachter and Fromkin 1968; Dolphyne 1988a
aka	Akan	AA	Hartell 1993; Chanard 2006
aka	Akan	PHOIBLE	Dolphyne 1988b
ake	AKAWAIO	UPSID	Edwards 1978

ISO 639-3	Language Name	Source	Reference
akl	Aklan	PHOIBLE	Chai 1971
akp	Siwu	PHOIBLE	Iddah 1975
akz	Alabama	SPA	Rand 1968
akz	ALABAMA	UPSID	Rand 1968
alc	QAWASQAR	UPSID	Clairis 1977
ald	ALLADIAN	UPSID	Duponchel 1971; Dumestre and Duponchel 1971
ale	Aleut	SPA	Jakobson 1944; Bergsland 1956, 1959; Menovshchikov 1968
ale	ALEUT	UPSID	Jakobson 1944; Bergsland 1956, 1959; Menovshchikov 1968
alh	Alawa	SPA	Sharpe 1972
alh	ALAWA	UPSID	Sharpe 1972
als	Albanian	SPA	Newmark 1957
als	ALBANIAN	UPSID	Newmark 1957
aly	Alyawarra	PHOIBLE	Yallop 1977
amc	Amahuaca	SPA	Osborn 1948
amc	AMAHUACA	UPSID	Osborn 1948
ame	Amuesha	SPA	Fast 1953
ame	AMUESHA	UPSID	Fast 1953; Wise 1958
amf	HAMER	UPSID	Lydall 1976
amh	Amharic	SPA	Sumner 1957; Klingenheben 1966; Leslau 1968
amh	AMHARIC	UPSID	Sumner 1957; Klingenheben 1966; Leslau 1968
amn	Amanab	PHOIBLE	Minch 1992
amo	AMO	UPSID	Di Luzio 1972; Anderson 1980
amp	ALAMBLAK	UPSID	Bruce 1984

ISO 639-3	Language Name	Source	Reference
amu	AMUZGO	UPSID	Bauernschmidt 1965; Longacre 1966
anc	Angas	SPA	Burquest 1971
anc	ANGAS	UPSID	Burquest 1971
anc	Angas	AA	Hartell 1993; Chanard 2006
ann	Obolo	PHOIBLE	Faraclas 1984
ano	ANDOKE	UPSID	Landaburu 1979
ant	Western Desert	SPA	Douglas 1955, 1964
ant	WESTERN DESERT	UPSID	Douglas 1955, 1964
anu	Anywa	PHOIBLE	Reh 1996
anv	Denya	PHOIBLE	Mbuagbaw 1996
any	Anyi	PHOIBLE	Pyne 1972
any	Anyi Sanvi	PHOIBLE	Ahua 2004
any	Agni Djuablin	PHOIBLE	Ahua 2004
aoj	Muhiang	PHOIBLE	Conrad 1978
aon	Arapesh	PHOIBLE	Fortune 1942
apd	Arabe	AA	Hartell 1993; Chanard 2006
apj	Jicarilla Apache	PHOIBLE	Tuttle and Sandoval 2002
apm	Chiricahua Apache	PHOIBLE	Hoijer 1944
apn	Apinaye	SPA	Burgess and Ham 1968
apn	APINAYE	UPSID	Burgess and Ham 1968
apq	ANDAMANESE	UPSID	Brown 1914; Voegelin and Voegelin 1966
apu	Apurinã	PHOIBLE	Facundes 2000
aqc	ARCHI	UPSID	Kodzasov 1977
are	Arrarnte	PHOIBLE	Anderson 2000
arg	Aragonese	PHOIBLE	Mott 2007
arh	Ika	PHOIBLE	Franks 1985

ISO 639-3	Language Name	Source	Reference
arl	ARABELA	UPSID	Rich 1963
arn	Araucanian	SPA	Echeverría and Contreras 1965
arn	ARAUCANIAN	UPSID	Echeverría and Contreras 1965; Key 1978
arr	Karo	PHOIBLE	Gabas 1999
ary	Moroccan Arabic	SPA	Harrell 1962, 1965; Abdel-Massih 1973
arz	Egyptian Arabic	SPA	Kennedy 1960; Mitchell 1962; Tomiche 1964
arz	ARABIC	UPSID	Kennedy 1960; Mitchell 1962; Tomiche 1964
atb	Zaiwa	PHOIBLE	Wannemacher 1998
aty	Aneityum	PHOIBLE	Lynch 2000
auc	AUCA	UPSID	Saint and Pike 1962
auy	Auyana	SPA	Bee 1965b
ava	AVAR	UPSID	Zhirkov 1936; Charachidzé 1981
avn	Avatime	PHOIBLE	Schuh 1995
avt	Au	PHOIBLE	Scorza 1985
avu	Avokaya	AA	Hartell 1993; Chanard 2006
awn	Awiya	SPA	Hetzron 1969b
awn	AWIYA	UPSID	Hetzron 1969b
awx	Awara	PHOIBLE	Quigley 2003
axb	ABIPON	UPSID	Najlis 1966
ayg	Genyanga	PHOIBLE	Cleal 1973b
ayl	Lebanese Arabic	PHOIBLE	Elfitoury 1976
ayz	Maybrat	PHOIBLE	Dol 1999
azj	Azerbaijani	SPA	Householder Jr and Lofti 1965
azj	AZERBAIJANI	UPSID	Householder Jr and Lofti 1965
azo	Awing	PHOIBLE	Gisele 1994

ISO 639-3	Language Name	Source	Reference
bak	BASHKIR	UPSID	Poppe 1964
bam	BAMBARA	UPSID	Bird et al. 1977
bam	Bambara	AA	Hartell 1993; Chanard 2006
bao	Barasano	SPA	Stolte and Stolte 1971
bao	BARASANO	UPSID	Stolte and Stolte 1971
bas	Basaa	AA	Hartell 1993; Chanard 2006
bav	Babungo	AA	Hartell 1993; Chanard 2006
bax	Shupamem	PHOIBLE	Nchare 2005
baz	Nen	PHOIBLE	Mous 2006
bba	BARIBA	UPSID	Welmers 1952
bba	Bariba	AA	Hartell 1993; Chanard 2006
bbc	Batak	SPA	van der Tuuk 1971
bbc	BATAK	UPSID	van der Tuuk 1971
bbc	Toba-Batak	PHOIBLE	Percival 1964
bbk	Babanki	PHOIBLE	Akumbu 1999
bbl	BATS	UPSID	Desheriev 1953
bbo	BOBO-FING	UPSID	Morse 1976; le Bris and Prost 1981
bbo	Bobo	AA	Hartell 1993; Chanard 2006
bbp	Banda, West Central	PHOIBLE	Sampson 1985a
bbw	Baba	PHOIBLE	Pepandze 2005
bbx	Bubia	PHOIBLE	Fiensong S. Chia 1993
bby	Befang	PHOIBLE	Gueche 2004
bca	BAI	UPSID	Dell 1981
bce	Mamenyan	PHOIBLE	Forku 2000
bch	Bariai	PHOIBLE	Gallagher and Baehr 2005
bci	Baoulé	AA	Hartell 1993; Chanard 2006
bci	Baule	PHOIBLE	Timyan 1977
bcj	BARDI	UPSID	Metcalfe 1971

ISO 639-3	Language Name	Source	Reference
bcs	KOHUMONO	UPSID	Cook 1969a
bcw	Bana	PHOIBLE	Hofmann 1990
bdh	Baka	PHOIBLE	Santandrea 1976
bdi	Burun	PHOIBLE	Andersen 2006
bdl	SAMA	UPSID	Verheijen 1986
bdr	Bajau, West Coast	PHOIBLE	Miller 2007
bdu	Lukundu	PHOIBLE	Atta 1993
bdy	BANDJALANG	UPSID	Cunningham 1969
beh	Biali	AA	Hartell 1993; Chanard 2006
bej	BEJA	UPSID	Hudson 1974, 1976
bem	Bemba	PHOIBLE	Kula 2002
ben	Bengali	SPA	Ferguson and Chowdhury 1960
ben	BENGALI	UPSID	Ferguson and Chowdhury 1960
beq	Beembe	SPA	Jacquot 1962
beq	BEEMBE	UPSID	Jacquot 1962, 1981
bev	BETE	UPSID	Werle and Gbalehi 1976
bex	Jur Mödö	AA	Hartell 1993; Chanard 2006
bfd	Bafut	AA	Hartell 1993; Chanard 2006
bfl	Banda (CAF)	AA	Hartell 1993; Chanard 2006
bfl	Banda (Sudan)	AA	Hartell 1993; Chanard 2006
bfm	Mmen	PHOIBLE	Bangha 2003
bfw	Remo	PHOIBLE	Fernandez 1968
bgj	Bangolan	PHOIBLE	Mbah 2003
bgo	Baga Koga	PHOIBLE	Relich 1973
bhg	Binandere	PHOIBLE	Wilson 2002
bhw	Biak	PHOIBLE	Heuvel 2006
bhy	Bhele	AA	Hartell 1993; Chanard 2006
bib	BISA	UPSID	Naden 1973a,b

ISO 639-3	Language Name	Source	Reference
bib	Bisa	AA	Hartell 1993; Chanard 2006
bim	Bimoba	AA	Hartell 1993; Chanard 2006
bin	ẸDo	AA	Hartell 1993; Chanard 2006
bip	Bila	PHOIBLE	Lojenga 2006
biw	Bikele	PHOIBLE	Begné II 1979
bjr	Binumarien	PHOIBLE	Oatridge and Oatridge 1973
bjz	Baruga	PHOIBLE	Farr et al. 1996
bkc	Baka	AA	Hartell 1993; Chanard 2006
bkh	Bakoko	PHOIBLE	Edika 1990
bkk	Brokskat	PHOIBLE	Ramaswami 1982
bkm	Kom	AA	Hartell 1993; Chanard 2006
bkq	BAKAIRI	UPSID	Wheatley 1969, 1973
bkv	Bekwarra	AA	Hartell 1993; Chanard 2006
bla	Blackfoot	PHOIBLE	Taylor 1969
blb	Bilua	PHOIBLE	Obata 2003
blc	BELLA COOLA	UPSID	Newman 1947; Nater 1984
blk	Pa-O, Taungthu	PHOIBLE	Thanamteun 2000
bll	Biloxi	PHOIBLE	Einaudi 1974
blr	Blang	PHOIBLE	Block 1994
bmo	Bambalang	PHOIBLE	Fozoh 2002
bmr	MUINANE	UPSID	Walton and Walton 1967
bnm	Banoo	PHOIBLE	Kouankem 2003
bod	Tibetan	PHOIBLE	Cha 1995
boi	Barbareño	PHOIBLE	Beeler 1970; Wash 2001
bol	Bole	PHOIBLE	Gimba 2000
bom	BIROM	UPSID	Wolff 1959; Bouquiaux 1970
bom	Berom	AA	Hartell 1993; Chanard 2006
bor	BORORO	UPSID	Crowell 1979

ISO 639-3	Language Name	Source	Reference
bot	Bongo	PHOIBLE	Abessolo Eto 1990
bqc	Boussa (Boko)	AA	Hartell 1993; Chanard 2006
bqp	Busa	PHOIBLE	Wedekind 1972
bqx	Kambari	AA	Hartell 1993; Chanard 2006
brb	BRAO	UPSID	Keller 1976
brc	Berbice Dutch	PHOIBLE	Kouwenberg 1994
bre	Breton	SPA	Ternes 1970
bre	BRETON	UPSID	Ternes 1970; Bothorel 1982
brh	BRAHUI	UPSID	Emeneau 1937, 1962; De Armond 1975
brv	BRUU	UPSID	Thongkum 1979
brx	BODO	UPSID	Bhat 1968
bsk	Burushaski	SPA	Morgenstierne 1945
bsk	BURUSHASKI	UPSID	Morgenstierne 1945
bsp	Baga Sitem	PHOIBLE	Ganong 1998
bsq	Bassa	AA	Hartell 1993; Chanard 2006
bss	Akɔɔse (Bakossi)	AA	Hartell 1993; Chanard 2006
btg	BéTé	AA	Hartell 1993; Chanard 2006
bud	Bassar	AA	Hartell 1993; Chanard 2006
bud	N'Cam	PHOIBLE	Badie 1995
bul	Bulgarian	SPA	Klagstad 1958; Aronson 1968
bul	BULGARIAN	UPSID	Klagstad 1958; Aronson 1968; Bidwell 1968; Scatton 1984
bum	Bulu	AA	Hartell 1993; Chanard 2006
bun	Sherbro	PHOIBLE	Pichl 1973d
buu	Kibudu	PHOIBLE	Kutsch Lojenga 1994
buy	Mmani	PHOIBLE	Pichl 1973b
bvm	Bamunka	PHOIBLE	Ngeloh Takwe 2002
bvr	BURARRA	UPSID	Glasgow and Glasgow 1967

ISO 639-3	Language Name	Source	Reference
bvx	Babole	PHOIBLE	Leitch 2003
bwi	Kurripako	PHOIBLE	Granadillo 2006
bwo	Boro	PHOIBLE	Bhattacharya 1977
bwq	Bobo	PHOIBLE	Sanou 1978
bwt	Bafo	PHOIBLE	Ebah Ebude 1990
bxk	Bukusu	PHOIBLE	Mutonyi 2000
byn	Blin	PHOIBLE	Fallon 2006
byv	Medumba	PHOIBLE	Nganmou 1991
byx	BAINING	UPSID	Parker and Parker 1974
bza	Bandi (Gbande)	AA	Hartell 1993; Chanard 2006
bzd	BRIBRI	UPSID	Arroyo 1972
bzf	Boiken	PHOIBLE	Freudenburg and Freudenburg 1974
bzj	Belizean Creole	PHOIBLE	Greene 1994
bzx	Bozo	AA	Hartell 1993; Chanard 2006
cad	CADDO	UPSID	Chafe 1976
cag	ASHUSLAY	UPSID	Stell 1972
cao	Chacobo	SPA	Prost 1967
caq	NICOBARESE	UPSID	Das 1977
car	Carib	SPA	Hoff 1968; Peasgood 1972
car	CARIB	UPSID	Hoff 1968; Peasgood 1972
cas	Moseten	PHOIBLE	Sakel 2004
cat	Catalan	PHOIBLE	Carbonell and Llisterri 1992
cav	Cavinena	PHOIBLE	Guillaume 2004
cbi	Cayapa	SPA	Lindskoog and Brend 1962
cbi	CAYAPA	UPSID	Lindskoog and Brend 1962
cbn	NYAH KUR	UPSID	Diffloth 1984; Thongkum 1984
cbv	CACUA	UPSID	Cathcart 1979; Anderton 1989
ccc	Chamicuro	PHOIBLE	Parker 1991

ISO 639-3	Language Name	Source	Reference
cce	Copi	PHOIBLE	Gowlett 2003
cdm	Chepang	PHOIBLE	Caughley 1982
cdo	FUZHOU	UPSID	Jiahua 1960
ces	Czech	PHOIBLE	Dankovičová 1997
cha	Chamorro	SPA	Topping 1973
cha	CHAMORRO	UPSID	Costenoble 1935; Seiden 1960; Topping 1980, 1973
che	Chechen	PHOIBLE	Nichols 1996a
chf	Chontal	SPA	Keller 1959
cho	Choctaw	PHOIBLE	Broadwell 2006
chp	Chipewyan	SPA	Li 1932, 1933, 1946
chp	CHIPEWYAN	UPSID	Li 1932, 1933, 1946
chq	HIGHLAND CHINANTEC	UPSID	Robbins 1961, 1968, 1975
chr	CHEROKEE	UPSID	Bender and Bender 1946; Walker 1975; Cook 1979
chv	Chuvash	SPA	Kruger 1961; Andreev 1966
chv	CHUVASH	UPSID	Kruger 1961
cic	Chickasaw	PHOIBLE	Gordon et al. 2000
cid	Chimariko	PHOIBLE	Jany 2007
cja	Cham	SPA	Blood 1967
cja	CHAM	UPSID	Blood 1967
cjh	UPPER CHEHALIS	UPSID	Kinkade 1963
cjv	CHUAVE	UPSID	Thurman 1970
cko	Anufɔ	AA	Hartell 1993; Chanard 2006
cko	Anufɔ	PHOIBLE	Stanford and Lyn 1970
ckt	Chukchi	SPA	Skorik 1961, 1968
ckt	CHUKCHI	UPSID	Skorik 1961, 1968

ISO 639-3	Language Name	Source	Reference
cku	Koasati	PHOIBLE	Kimball 1985
cle	Chinanteco	PHOIBLE	Rupp 1980
cme	Cerma	AA	Hartell 1993; Chanard 2006
cmn	Mandarin Chinese	SPA	Karlgren 1926; Chao 1968; Dow 1972; Cheng 1973a
cmn	MANDARIN	UPSID	Karlgren 1926; Jiahua 1960; Chao 1968; Dow 1972; Cheng 1973a
cmn	Standard Chinese; Mandarin	PHOIBLE	Lee and Zee 2003
cnh	LAI	UPSID	Ouyang and Zheng 1963, 1980; Liang 1984b,a
cni	Campa	SPA	Dirks 1953
cni	CAMPA	UPSID	Dirks 1953; Payne 1981
cof	Colorado; Tsafiki	PHOIBLE	Dickinson 2002
cog	Chong	PHOIBLE	Ungsitipoonporn 2001
com	Comanche	PHOIBLE	Charney 1993
con	COFAN	UPSID	Borman 1962
coo	Comox	PHOIBLE	Harris 1981
cou	KONYAGI	UPSID	Santos 1977
cpn	Hill Guang	PHOIBLE	Painter 1974
cqd	HMONG	UPSID	Purnell 1972; Wang 1983, 1985
crb	Island Carib	SPA	Taylor 1955
crb	ISLAND CARIB	UPSID	Taylor 1955
crd	Coeur d'Alene	PHOIBLE	Doak 1997
crg	Michif	PHOIBLE	Rosen 2007
cro	Crow	PHOIBLE	Kaschube 1967
crw	Chrau	PHOIBLE	Thomas 1971
csk	Joola Huluf	PHOIBLE	Pike and Diatta 1994

ISO 639-3	Language Name	Source	Reference
cso	Sochiapan Chinantec	PHOIBLE	Foris 1993
ctd	TIDDIM CHIN	UPSID	Henderson 1965
ctp	CHATINO	UPSID	Pride 1965
ctu	Tila, Chiapas	PHOIBLE	Alvarez 2002
cub	CUBEO	UPSID	Salser 1971
cup	Cupeno	PHOIBLE	Hill 2005
cuv	Cuvok	PHOIBLE	Ndokobai 2003
cyb	CAYUVAVA	UPSID	Key 1961
daa	DANGALEAT	UPSID	Fedry 1977
daf	DAN	UPSID	Bearth and Zemp 1967
daf	Dan	AA	Hartell 1993; Chanard 2006
dag	Dagbani	SPA	Wilson and Bendor-Samuel 1969
dag	DAGBANI	UPSID	Wilson and Bendor-Samuel 1969
dag	Dagbani	AA	Hartell 1993; Chanard 2006
daj	DAJU	UPSID	Tucker and Bryan 1966; Thelwall 1981
dal	DAHALO	UPSID	Tucker et al. 1977; Nurse 1986
dap	Dafla	SPA	Ray 1967
dap	DAFLA	UPSID	Ray 1967
dbj	Ida'an	PHOIBLE	Goudswaard 2005
dbl	DYIRBAL	UPSID	Dixon 1972
dbq	Daba	AA	Hartell 1993; Chanard 2006
deu	German	SPA	Moulton 1962; Werner 1972; Philipp 1974
deu	GERMAN	UPSID	Moulton 1962; Wangler 1972
dga	Dagaare	AA	Hartell 1993; Chanard 2006
dgh	Dghwede	PHOIBLE	Frick 1973
dgi	Dagara	AA	Hartell 1993; Chanard 2006
dib	Dinka, South Central	PHOIBLE	Andersen 1987b

ISO 639-3	Language Name	Source	Reference
dic	Dida	AA	Hartell 1993; Chanard 2006
dif	DIYARI	UPSID	Austin 1981
dih	Digueno	SPA	Langdon 1970
dih	DIEGUENO	UPSID	Langdon 1970
dip	DINKA	UPSID	Andersen 1987c; Malou 1988
diq	Dimili	PHOIBLE	Todd 1985
diu	Diriku	PHOIBLE	Sommer 2003
dje	Zarma	AA	Hartell 1993; Chanard 2006
dni	DANI	UPSID	Bromley 1961; van der Stap 1966
dob	Dobu	PHOIBLE	Lithgow 1977
doc	KAM	UPSID	Guoqiao and Yang 1988
dop	Lokpa	AA	Hartell 1993; Chanard 2006
dow	DOAYO	UPSID	Wiering 1974
dow	Doayo	AA	Hartell 1993; Chanard 2006
dru	RUKAI	UPSID	Li 1973, 1977b
dta	DAGUR	UPSID	Anonymous 1982
dts	DOGON	UPSID	Bendor-Samuel et al. 1989
dts	Dogon	AA	Hartell 1993; Chanard 2006
dua	Duala	AA	Hartell 1993; Chanard 2006
dug	Duruma	AA	Hartell 1993; Chanard 2006
duj	YOLNGU	UPSID	Rose and Morphy 1982; Morphy 1983
duo	Dupaningan Agta	PHOIBLE	Robinson 2008
dyo	DIOLA	UPSID	Sapir 1965
dyo	Joola	AA	Hartell 1993; Chanard 2006
dyu	Dioula	AA	Hartell 1993; Chanard 2006
dzg	Daza	AA	Hartell 1993; Chanard 2006
efi	EFIK	UPSID	Cook 1969b
ega	Ega	PHOIBLE	Connell et al. 2002

ISO 639-3	Language Name	Source	Reference
ego	Eggon	PHOIBLE	Blench and Hepburn 2006
ekg	EKARI	UPSID	Doble 1962; Steltenpool 1969; Doble 1987
ekm	Nulibie	PHOIBLE	Ekambi 1990
ekp	Ekpeye	PHOIBLE	Blench 2006c
ell	Modern Greek	SPA	Householder et al. 1964; Newton 1972; Kaisse 1975, 1976
ell	GREEK	UPSID	Householder et al. 1964; Pring 1967; Newton 1972
ema	Emai	AA	Hartell 1993; Chanard 2006
emk	Manding	AA	Hartell 1993; Chanard 2006
enb	Endo	PHOIBLE	Zwarts 2003
eng	English	SPA	Gimson 1962; Trnka 1968; O'Conner 1973; Halle 1973; Fudge 1975
enn	Engenni	AA	Hartell 1993; Chanard 2006
erk	South Efate	PHOIBLE	Thieberger 2004
esi	Iñupiaq	PHOIBLE	Nagai 2006
ess	YUPIK	UPSID	Krauss 1975
ets	Etsako	PHOIBLE	Elimelech 1976
etu	EJAGHAM	UPSID	Watters 1981
etu	Ejagham De L'Ouest	AA	Hartell 1993; Chanard 2006
eus	Basque	SPA	Gavel 1929; N'diaye 1970
eus	BASQUE	UPSID	Gavel 1929; N'diaye 1970
eve	Even	SPA	Novikova 1960
eve	EVEN	UPSID	Novikova 1960
ewe	Ewe	SPA	Berry 1951a; Ansre 1961; Ladefoged 1964; Stahlke 1970

ISO 639-3	Language Name	Source	Reference
ewe	EWE	UPSID	Berry 1951a; Ladefoged 1964; Stahlke 1970
ewe	Eue	AA	Hartell 1993; Chanard 2006
ewo	EWONDO	UPSID	Abega 1970; Redden 1979; Nnomo and Mbezele 1982
ewo	Ewondo	AA	Hartell 1993; Chanard 2006
eya	EYAK	UPSID	Krauss 1965
faa	FASU	UPSID	Loeweke and May 1964
fai	Faiwol	PHOIBLE	Mecklenburg 1974
fan	Fan	PHOIBLE	Eko 1974
fap	Ndut-Falor	PHOIBLE	Pichl 1973c
ffm	Fulfulde (Mali)	AA	Hartell 1993; Chanard 2006
fia	Mahas-Fiyadikka	SPA	Bell 1971
fia	NUBIAN	UPSID	Stevenson 1957; Bell 1968, 1971
fij	FIJIAN	UPSID	Dixon 1988
fil	Filipino	PHOIBLE	Cubar and Cubar 1994
fin	Finnish	SPA	Lehtinen 1964; Harms 1964, 1966; Austerlitz 1967; Kiparsky 1968; Hammarberg 1974
fin	FINNISH	UPSID	Harms 1964; Lehtinen 1964; Harms 1966; Austerlitz 1967; Hammarberg 1974
flr	Fuliru	AA	Hartell 1993; Chanard 2006
fmp	FE?FE?	UPSID	Hyman 1972
fmp	Fe’efe’fe	AA	Hartell 1993; Chanard 2006
fod	Foodo	PHOIBLE	Plunkett 2009
fon	Fɔn	AA	Hartell 1993; Chanard 2006
fra	French	SPA	Sten 1963

ISO 639-3	Language Name	Source	Reference
fra	FRENCH	UPSID	Sten 1963
frr	Frisian	PHOIBLE	Lasswell 1998
fub	Fulfulde (Cameroon)	AA	Hartell 1993; Chanard 2006
fub	Fulfulde (Cameroon)	PHOIBLE	Taylor 1953
fub	Fulfulde (Fuunaangere)	PHOIBLE	Bickoe 2000
fub	Adamawa Fulfulde	PHOIBLE	Stennes 1967
fuc	Pulaar	AA	Hartell 1993; Chanard 2006
fuf	Pular	AA	Hartell 1993; Chanard 2006
fuh	Fulfulde (Burkina Faso)	AA	Hartell 1993; Chanard 2006
fun	IATE	UPSID	Lapenda 1968
fuq	Fulfulde (Niger)	AA	Hartell 1993; Chanard 2006
fur	Friulian	PHOIBLE	Miotti 2002
fuu	Furu	PHOIBLE	Boyeldieu 2000
fuv	Fula (Nigeria)	PHOIBLE	Arnott 1968a
fuv	Fulfulde (NGA)	PHOIBLE	McIntosh 1984
fvr	FUR	UPSID	Tucker and Tucker 1966; Beaton 1968
fwa	PO-AI	UPSID	Li 1977a
fwe	Fwe	PHOIBLE	Baumbach 1997a
gaa	Ga	SPA	Berry 1951b
gaa	GA	UPSID	Berry 1951b
gaa	Ga	AA	Hartell 1993; Chanard 2006
gaj	Gadsup	SPA	Frantz and Frantz 1966
gaj	GADSUP	UPSID	Frantz and Frantz 1966
gay	Gayo	PHOIBLE	Eades and Hajek 2006
gbc	GARAWA	UPSID	Furby 1974
gbd	Garadjari	PHOIBLE	Sands 1989

ISO 639-3	Language Name	Source	Reference
gbo	Grebo	AA	Hartell 1993; Chanard 2006
gbp	Gbeya	SPA	Samarin 1966
gbp	GBEYA	UPSID	Samarin 1966; Monino and Roulon 1972
gbp	Gbaya (Bossangoa, CAR)	AA	Hartell 1993; Chanard 2006
gbr	GWARI	UPSID	Hyman and Magaji 1970
gbs	Gbesi	PHOIBLE	Capo 1991
gby	Gbari	PHOIBLE	Rosendall 1998
gde	Guɗe	AA	Hartell 1993; Chanard 2006
gej	Gɛn-Mina (Benin)	AA	Hartell 1993; Chanard 2006
gej	Gɛn-Mina (Togo)	AA	Hartell 1993; Chanard 2006
gid	Kada	PHOIBLE	Noukeu 2002
gio	GELAO	UPSID	He 1981, 1983
gjn	Gonja	AA	Hartell 1993; Chanard 2006
gju	Gojri	PHOIBLE	Losey 2002
gkp	Kpɛlɛwoo	AA	Hartell 1993; Chanard 2006
gld	NANAI	UPSID	Avrorin 1968
gle	Irish Gaelic	SPA	Mac an Fhailigh 1968; Burke 1970
gle	IRISH	UPSID	Brothers 1905; Sommerfelt 1964; Mac an Fhailigh 1968
glg	Galician	PHOIBLE	Regueira 1996
gmm	Mbodomo	PHOIBLE	Boyd 1997
gmo	KULLO	UPSID	Allan 1976b
gnd	Zulgo	AA	Hartell 1993; Chanard 2006
gng	Gangam	AA	Hartell 1993; Chanard 2006
god	Godié	AA	Hartell 1993; Chanard 2006
grg	Ma'di	PHOIBLE	Blackings and Fabb 2003a
grt	Garo	SPA	Burling 1961

ISO 639-3	Language Name	Source	Reference
gub	Guajajara	PHOIBLE	Bendor-Samuel 1966
guc	Goajiro	SPA	Holmer 1949
guc	GUAJIRO	UPSID	Holmer 1949; Mansen 1967
gug	Guarani	SPA	Uldall 1956; Gregores and Suárez 1967; Lunt 1973
gug	GUARANI	UPSID	Uldall 1956; Gregores and Suárez 1967; Lunt 1973
guh	GUAHIBO	UPSID	Kondo and Kondo 1967
gum	GUAMBIANO	UPSID	Caudmont 1954; Branks and Branks 1973
gup	Gunwinggu	PHOIBLE	Oates 1964a
guq	ACHE	UPSID	Susnik 1974
gur	Frafra	PHOIBLE	Schaefer 1975
gux	Gulmancema	AA	Hartell 1993; Chanard 2006
gvc	Wanano	PHOIBLE	Stenzel 2004
gvf	Golin	PHOIBLE	Evans et al. 2005
gvn	GUGU-YALANDYI	UPSID	Oates 1964b; Oates and Oates 1964; Wurm 1972a
gya	Gbaya (Northwest, Car)	AA	Hartell 1993; Chanard 2006
hag	Hanga	AA	Hartell 1993; Chanard 2006
hak	Hakka	SPA	Hashimoto 1973
hak	HAKKA	UPSID	Hashimoto 1973
hau	Hausa	SPA	Greenberg 1941; Hodge 1947; Abraham 1959a,b; Kraft 1963; Hodge and Umaru 1963; Brauner and Ashiwaju 1965; Kraft and Kraft 1973; Kraft and Kirk-Greene 1973

ISO 639-3	Language Name	Source	Reference
hau	HAUSA	UPSID	Abraham 1934; Greenberg 1941; Hodge 1947; Abraham 1959b,a; Taylor 1959; Hodge and Umaru 1963; Kraft 1963; Brauner and Ashiwaju 1965; Kraft and Kirk-Greene 1973; Kraft and Kraft 1973
hau	Hausa (Niger)	AA	Hartell 1993; Chanard 2006
hau	Hausa (Nigeria)	AA	Hartell 1993; Chanard 2006
haw	Hawaiian	SPA	Pukui and Elbert 1965
haw	HAWAIIAN	UPSID	Pukui and Elbert 1965; Elbert and Pukui 1979; Schutz 1981
hay	Haya	PHOIBLE	Byarushengo 1977
hbb	Kilba	PHOIBLE	Greive 1973
hch	Huichol	PHOIBLE	McIntosh 1945
hdn	Haida	SPA	Sapir 1923
hdn	HAIDA	UPSID	Sapir 1923
heb	Modern Hebrew	SPA	Cohen and Zafrani 1968; Chayen 1973
her	Herero	PHOIBLE	Elderkin 2003
hin	Hindi-Urdu	SPA	Pinnow 1972; Vermeer and Sharma 1966; Kelkar 1968; Harms 1969
hin	HINDI-URDU	UPSID	Pinnow 1972; Vermeer and Sharma 1966; Kelkar 1968; Harms 1969; Ohala 1983
hix	HIXKARYANA	UPSID	Derbyshire 1985
hoa	Hoava	PHOIBLE	Davis 2003
hop	Hopi	SPA	Whorf 1946; Voegelin 1956
hop	HOPI	UPSID	Whorf 1946; Kluckhohn and MacLeish 1955; Voegelin 1956
hrv	Croatian	PHOIBLE	Landau et al. 1995

ISO 639-3	Language Name	Source	Reference
hts	HADZA	UPSID	Tucker et al. 1977
hum	Kihungan	PHOIBLE	Takizala 1974
hun	Hungarian	SPA	Hall 1938, 1944; Banhidi et al. 1965; Kalman 1972
hun	HUNGARIAN	UPSID	Hall 1938, 1944; Banhidi et al. 1965; Kalman 1972
hup	Hupa	SPA	Woodward 1964; Golla 1970
hup	HUPA	UPSID	Woodward 1964; Golla 1970
hur	Chilliwak Halkomelem	PHOIBLE	Galloway 1977
hus	HUASTECO	UPSID	Larsen and Pike 1949; Ochoa Peralta 1984
hus	Huastec	PHOIBLE	Edmonson 1988
huv	HUAVE	UPSID	Suárez 1975; Stairs Kreger and de Stairs 1981
huv	Huave, San Mateo del Mar	PHOIBLE	Rupp 1983
hye	Armenian	SPA	Allen 1950
hye	ARMENIAN	UPSID	Allen 1950
iai	Iai	SPA	Tryon 1968; Haudricourt 1971
iai	IAI	UPSID	Tryon 1968; Haudricourt 1971; Ozanne-Rivierre 1976
iba	IBAN	UPSID	Scott 1957; Omar 1981
ibb	Ibibio	PHOIBLE	Urua 2004
ibo	Igbo	SPA	Ward 1936; Carnochan 1948; Swift et al. 1962; Ladefoged 1968; Williamson 1969

ISO 639-3	Language Name	Source	Reference
ibo	IGBO	UPSID	Ward 1936; Carnochan 1948; Swift et al. 1962; Ladefoged 1964; Williamson 1969; Ladefoged et al. 1976
ibo	Igbo	AA	Hartell 1993; Chanard 2006
iby	Ibani	PHOIBLE	Blench 2005a
ife	Ifɛ̀	AA	Hartell 1993; Chanard 2006
igb	Ebira	AA	Hartell 1993; Chanard 2006
ige	Igede	AA	Hartell 1993; Chanard 2006
ign	Moxo	SPA	Ott and Ott 1967
ign	MOXO	UPSID	Ott and Ott 1967
igo	Ngomba	PHOIBLE	Ngouagna 1988
ijc	KOLOKUMA IJO	UPSID	Williamson 1965
ijn	Kalabari	PHOIBLE	Harry 2003
ijs	Eastern Ijo (Okrika)	AA	Hartell 1993; Chanard 2006
ikx	IK	UPSID	Heine 1975a
ilo	Ilocano	PHOIBLE	Rubino 1997
imn	Imonda	PHOIBLE	Seiler 1985
ind	Indonesian	PHOIBLE	Soderberg and Olson 2008
inh	Ingush	PHOIBLE	Nichols 1996b
irh	IRARUTU	UPSID	Voorhoeve 1989
irk	Iraqw	SPA	Whiteley 1958
irk	IRAQW	UPSID	Whiteley 1958
irn	IRANXE	UPSID	Meader 1967
isl	Icelandic	SPA	Einarsson 1949; Haugen 1958
iso	ISOKO	UPSID	Donwa 1982
ita	Italian	PHOIBLE	Rogers and d'Arcangeli 2004
itl	ITELMEN	UPSID	Volodin 1976
ito	Itonama	SPA	Liccardi and Grimes 1968

ISO 639-3	Language Name	Source	Reference
ito	ITONAMA	UPSID	Liccardi and Grimes 1968
ium	Yao	SPA	Purnell 1965
ium	MIEN	UPSID	Purnell 1965; Mao et al. 1982
ivv	IVATAN	UPSID	Cottle and Cottle 1958; Heye and Heye 1967; Hidalgo and Hidalgo 1971
iwm	IWAM	UPSID	Laycock 1965
izi	ẸZaa	AA	Hartell 1993
izi	Izi	AA	Hartell 1993
izi	Ikwo	AA	Chanard 2006; Hartell 1993
izi	ẸZaa	AA	Chanard 2006
izi	Izi	AA	Chanard 2006
jaa	Jarawara	PHOIBLE	Vogel 2003
jac	JACALTEC	UPSID	Day 1973; Craig 1977
jam	Jamaican Creole	PHOIBLE	Harry 2006
jao	YANYUWA	UPSID	Kirton 1967; Kirton and Charlie 1978; Huttar and Kirton 1981
jar	Jarawa	PHOIBLE	Lukas and Willms 1961
jav	Javanese	SPA	Uhlenbeck 1949, 1963
jav	JAVANESE	UPSID	Horne 1961; Uhlenbeck 1963; Herrfurth 1964; Fagan 1988
jbj	Arandai	PHOIBLE	Voorhoeve 1985
jeb	JEBERO	UPSID	Bendor-Samuel 1961
jic	TOL	UPSID	Fleming and Dennis 1977
jiv	Jivaro	SPA	Beasley and Pike 1957
jiv	JIVARO	UPSID	Beasley and Pike 1957
jmc	Machame	PHOIBLE	Kagaya and Olomi 2006
jow	Jowulu	PHOIBLE	Carlson 1993
jpn	Japanese	SPA	Bloch 1950; Martin 1952; Jorden 1963

ISO 639-3	Language Name	Source	Reference
jpn	JAPANESE	UPSID	Bloch 1950; Martin 1952; Jorden 1963; Shibatani 1990
jqr	Aymara	SPA	Hardman 1966
jqr	JAQARU	UPSID	Hardman 1966, 1983
jru	JAPRERIA	UPSID	Durbin and Seijas 1972
jum	Jumjum	PHOIBLE	Andersen 2004
jun	Juang	PHOIBLE	Matson 1964
jup	Hup	PHOIBLE	Epps 2005
jya	Jiarong	PHOIBLE	Jacques 2004
kab	Kabyle	PHOIBLE	Hamouma 1987
kac	JINGPHO	UPSID	Liu 1964
kal	Inuit	SPA	Kleinschmidt 1851; Thalbitzer 1904; Rischel 1974
kal	INUIT	UPSID	Kleinschmidt 1851; Thalbitzer 1904; Rischel 1974
kas	Kashimiri	SPA	Kelkar and Trisal 1964
kas	KASHMIRI	UPSID	Kelkar and Trisal 1964; Zakhar'in and Edelman 1971; Zakhar'in 1974; Bhat 1987
kat	Georgian	SPA	Selmer 1935; Vogt 1938, 1939; Robins and Waterson 1952; Tschenkéli 1958; Vogt 1958, 1971
kat	GEORGIAN	UPSID	Selmer 1935; Vogt 1938; Robins and Waterson 1952; Neisser 1953; Tschenkéli 1958; Vogt 1958, 1971
kbc	Kadiweu	PHOIBLE	Sandalo 1995
kbd	Kabardian	SPA	Kuipers 1960
kbd	KABARDIAN	UPSID	Kuipers 1960

ISO 639-3	Language Name	Source	Reference
kbh	CAMSA	UPSID	Howard 1972, 1967; Mongui Sánchez 1981
kbk	KOIARI	UPSID	Dutton 1969
kbp	Kabiyɛ	AA	Hartell 1993; Chanard 2006
kbp	Kabiye	PHOIBLE	Padayodi 2008
kbr	KEFA	UPSID	Fleming 1976
kbv	DERA	UPSID	Voorhoeve 1971
kca	Ostyak	SPA	Steinitz 1950; Gulya 1966; Katz 1975a
kca	KHANTY	UPSID	Steinitz 1950; Gulya 1966; Katz 1975a
kca	Eastern Khanty	PHOIBLE	Filchenko 2007
kck	Ikalanga	PHOIBLE	Mathangwane 1999
kcl	Kela	PHOIBLE	Collier and Collier 1975
kcn	Nubi	PHOIBLE	Wellens 2003
kcv	Kete	PHOIBLE	Muzenga 1980
kde	Shimakonde	PHOIBLE	Liphola 2001
kdh	Tem	AA	Hartell 1993; Chanard 2006
kdt	Kuay	PHOIBLE	Oranuch 1984
kek	K'EKCHI	UPSID	Haeseriju 1966; Freeze 1975
ken	Kenyang	PHOIBLE	Mbuagbaw 2000
ker	KERA	UPSID	Ebert 1976, 1979
ket	Ket	SPA	Dul'zon 1968; Krejnovich 1968b
ket	KET	UPSID	Dul'zon 1968; Krejnovich 1968b
kfe	Kota	SPA	Emeneau 1944
kfe	KOTA	UPSID	Emeneau 1944
kff	KOYA	UPSID	Tyler 1969
kfk	Kinnauri	PHOIBLE	Nigam and Neethivanan 1971
kgg	Kusunda	PHOIBLE	Watters 2006

ISO 639-3	Language Name	Source	Reference
kgp	KAINGANG	UPSID	Henry 1935; Kindell 1972; Wiesemann 1972
kha	Khasi	SPA	Rabel 1961
kha	KHASI	UPSID	Rabel 1961
khb	LUE	UPSID	Li 1964
khc	Tukang Besi	PHOIBLE	Donohue 1994
khk	Khalkha	SPA	Street 1963; Luvsanvandan 1964; Hangin 1968
khk	KHALKHA	UPSID	Street 1963; Luvsanvandan 1964; Hangin 1968; Svantesson 1985
khl	Kaliai	SPA	Counts 1969
khl	KALIAI	UPSID	Counts 1969
khm	Cambodian	SPA	Jacob 1968; Huffman 1970b,a
khm	KHMER	UPSID	Jacob 1968; Huffman 1970b,a; Ehrman 1972
khq	Songhoy	AA	Hartell 1993; Chanard 2006
khq	Songhoy	PHOIBLE	Heath 2005b
khr	Kharia	SPA	Pinnow 1959; Biligiri 1965
khr	KHARIA	UPSID	Pinnow 1959; Biligiri 1965
khr	Kharia	PHOIBLE	Peterson 2006
khy	Iikile	PHOIBLE	Carrington 1977
kig	Khmu	PHOIBLE	Wongnoppharalert 1993
kik	Kikuyu	AA	Hartell 1993; Chanard 2006
kin	Kinyarwanda	PHOIBLE	Mpayimana 2003
kio	KIOWA	UPSID	Harrington 1928; Sivertsen 1956; Watkins and McKenzie 1984
kir	Kirghiz	SPA	Herbert and Poppe 1963

ISO 639-3	Language Name	Source	Reference
kir	KIRGHIZ	UPSID	Herbert and Poppe 1963; Junusaliev 1966
kjd	SOUTHERN KIWAI	UPSID	Wurm 1977
kjg	KHMU?	UPSID	Svantesson 1983
kjn	Kunjen	SPA	Capell 1967; Sommer 1969
kjq	ACOMA	UPSID	Miller 1966
kjs	KEWA	UPSID	Franklin and Franklin 1962
kkj	Kakɔ	AA	Hartell 1993; Chanard 2006
kkk	Kokota	PHOIBLE	Palmer 1999
kkw	TEKE	UPSID	Paulian 1975
kla	KLAMATH	UPSID	Barker 1964
kle	Kulung	PHOIBLE	Tolsma 1999
klu	KLAO	UPSID	Singler 1979
kma	Konni	PHOIBLE	Cahill 1999
kme	Kole	PHOIBLE	Asobo 1989
kmn	Awtuw	PHOIBLE	Feldman 1986
kmo	Washkuk	SPA	Kooyers et al. 1971
kmo	KWOMA	UPSID	Kooyers et al. 1971
kmr	KURDISH	UPSID	Abdulla and McCarus 1967
kms	Kamasau	PHOIBLE	Sanders and Sanders 1994
kmv	Karipuna Creole	PHOIBLE	Tobbler 1983
kmw	Komo	AA	Hartell 1993; Chanard 2006
kna	KANAKURU	UPSID	Newman 1974
knc	Kanuri	SPA	Lukas 1937
knc	KANURI	UPSID	Lukas 1937; Awobuluyi 1971; Hutchison 1981
knc	Kanuri	AA	Hartell 1993; Chanard 2006

ISO 639-3	Language Name	Source	Reference
knn	KONKANI	UPSID	Gajendragadkar 1970; Major 1979; Madtha 1984
kor	Korean	SPA	Martin 1951, 1954; Cho 1967; Kim 1968; Martin and Lee 1969; Kim 1972
kor	KOREAN	UPSID	Martin 1951, 1954; Cho 1967; Martin and Lee 1969; Kim 1972, 1986
kpk	KPAN	UPSID	Shimizu 1971
kpm	SRE	UPSID	Manley 1972
kpr	Korafe	PHOIBLE	Farr and Farr 1974
kpv	Komi	SPA	Bubrix 1949a; Lytkin 1966
kpv	KOMI	UPSID	Bubrix 1949b; Lytkin 1966
kpy	KORYAK	UPSID	Zhukova 1980
kpz	SEBEI	UPSID	Montgomery 1970
kqk	Kotafon	PHOIBLE	Capo 1991
kqs	Kisiei	AA	Hartell 1993; Chanard 2006
kri	Krio	AA	Hartell 1993; Chanard 2006
krm	Krim	PHOIBLE	Pichl 1973a
krs	Kresh	AA	Hartell 1993; Chanard 2006
kru	Kurukh	SPA	Grignard 1924; Pinnow 1964; Pfeiffer 1972
kru	KURUKH	UPSID	Pinnow 1964; Pfeiffer 1972
ksf	Kpaʔ	PHOIBLE	Guarisma 2006
ksi	Isaka	PHOIBLE	Donohue and Roque 2002
ksw	Karen	SPA	Jones Jr 1961
ksw	KAREN	UPSID	Jones Jr 1961
ktg	KALKATUNGU	UPSID	Blake 1979
ktn	Karitiana	PHOIBLE	Everett 2006
kto	Kuot	PHOIBLE	Lindstrom 2002

ISO 639-3	Language Name	Source	Reference
ktz	!XU	UPSID	Snyman 1970, 1975
kub	Kuteb	PHOIBLE	Koops 1990
kuj	Kuria	AA	Hartell 1993; Chanard 2006
kun	KUNAMA	UPSID	Tucker and Bryan 1966
kup	Kunimaipa	SPA	Pence 1966
kup	KUNIMAIPA	UPSID	Pence 1966
kus	Kusal	AA	Hartell 1993; Chanard 2006
kvn	CUNA	UPSID	Sherzer 1983
kwd	KWAIO	UPSID	Keesing 1985
kwi	Awa Pit	PHOIBLE	Curnow 1997
kwk	Kwakiutl	SPA	Boas 1947; Newman 1950
kwk	KWAKW'ALA	UPSID	Boas 1911, 1947; Newman 1950; Grubb 1977
kwl	Kofyar	PHOIBLE	Netting 1973
kwn	Kwangari	PHOIBLE	Sommer 2003
kws	Kwezo	PHOIBLE	Forges 1983
kxl	Dhangar	PHOIBLE	Yadava 2000
kxm	Northern Khmer	PHOIBLE	Phunsap 1984
kxo	Kanoe	PHOIBLE	Bacelar 2004
kxv	Kuvi	PHOIBLE	Israel 1979
kye	Krache	PHOIBLE	Cleal 1973c
kyh	Karok	SPA	Bright 1957
kyh	KAROK	UPSID	Bright 1957
kyz	Kaiabi	PHOIBLE	de Oliveira Borges E Souza 2004
kza	Karaboro	AA	Hartell 1993; Chanard 2006
kzr	Karaŋ	AA	Hartell 1993; Chanard 2006
lag	Langi	PHOIBLE	Dunham 2005
lao	Lao	PHOIBLE	Morev et al. 1979

ISO 639-3	Language Name	Source	Reference
las	Lama	AA	Hartell 1993; Chanard 2006
lbc	Lakkia	SPA	Haudricourt 1967
lbc	LAKKIA	UPSID	Haudricourt 1967
lbe	Lak	SPA	Zhirkov 1955; Khajdakov 1966; Murkelinskij 1967
lbe	LAK	UPSID	Zhirkov 1955; Khajdakov 1966; Murkelinskij 1967
lbj	Ladakhi	PHOIBLE	Koshal 1979
lch	Lucazi	PHOIBLE	Fleisch 2000
lea	Lega-Shabunda	PHOIBLE	Botne 2003
led	Lendu	AA	Hartell 1993; Chanard 2006
lee	LyéLé	AA	Hartell 1993; Chanard 2006
lef	LELEMI	UPSID	Höftmann 1971
lem	Nɔmaa (NɔmaáNdɛ́)	AA	Hartell 1993; Chanard 2006
lgg	Logbara	SPA	Crazzolara 1960; Tucker and Bryan 1966
lgg	LUGBARA	UPSID	Crazzolara 1960; Tucker and Bryan 1966; Anderson 1986
lgg	Lugbara	AA	Hartell 1993; Chanard 2006
lhu	Lahu	SPA	Matisoff 1973
lhu	LAHU	UPSID	Matisoff 1973
lia	Limba	AA	Hartell 1993; Chanard 2006
lig	Ligbi	AA	Hartell 1993; Chanard 2006
lik	Lika	PHOIBLE	Kutsch Lojenga 2008
lin	Lingala	AA	Hartell 1993; Chanard 2006
lip	Likpe	PHOIBLE	Allan 1974
lis	Lisu	PHOIBLE	Roop 1970

ISO 639-3	Language Name	Source	Reference
lit	Lithuanian	SPA	Augustitis 1964; Senn 1966; Ambrazas et al. 1966
lit	LITHUANIAN	UPSID	Augustitis 1964; Ambrazas et al. 1966; Senn 1966
lkt	Dakota	SPA	Boas and Deloria 1941
lkt	DAKOTA	UPSID	Boas and Deloria 1941
lln	Lele	AA	Hartell 1993; Chanard 2006
lln	Lele	PHOIBLE	Frajzyngier 2001
lme	LAME	UPSID	Sachnine 1982
lmp	Limbum	AA	Hartell 1993; Chanard 2006
lmp	Limbum (Southern)	PHOIBLE	Nforgwei 2004
lmp	Limbum (Central)	PHOIBLE	Nforgwei 2004
lmp	Limbum (Northern)	PHOIBLE	Nforgwei 2004
lmw	Lake Miwok	PHOIBLE	Callaghan 1963
lns	Nso’	AA	Hartell 1993; Chanard 2006
log	Logo	PHOIBLE	Tucker 1967
lok	Lɔkɔ	AA	Hartell 1993; Chanard 2006
lol	Mongo-Nkundu	PHOIBLE	Hulstaert 1961
lom	Lɔgɔmagooi	AA	Hartell 1993; Chanard 2006
lor	TééN	AA	Hartell 1993; Chanard 2006
los	Loniu	PHOIBLE	Hamel 1985
lue	Luvale	SPA	Horton 1949
lug	Luganda	AA	Hartell 1993; Chanard 2006
lui	Luiseno	SPA	Kroeber and Grace 1960; Malecot 1963; Bright 1965, 1968
lui	LUISENO	UPSID	Kroeber and Grace 1960; Malecot 1963; Bright 1965, 1968; Hyde 1971
lul	Lulubo	PHOIBLE	Andersen 1987a

ISO 639-3	Language Name	Source	Reference
lun	Lunda	PHOIBLE	Kawasha 2003
luo	Luo	SPA	Gregersen 1961
luo	LUO	UPSID	Gregersen 1961
luo	Dholuo	AA	Hartell 1993; Chanard 2006
lut	LUSHOOTSEED	UPSID	Snyder 1968
lvk	Lavukaleve	PHOIBLE	Terrill 1999
lyn	Louyi	PHOIBLE	Jacottet 1896
lzz	Laz	PHOIBLE	Anderson 1963
mam	Western Mam	PHOIBLE	Godfrey 1981
maq	Mazateco	SPA	Pike and Pike 1947
maq	MAZATEC	UPSID	Pike and Pike 1947; Jamieson 1976a,b
mas	Maasai	SPA	Tucker and Mpaayei 1955; Tucker and Bryan 1966
mas	MAASAI	UPSID	Tucker and Mpaayei 1955; Tucker and Bryan 1966
maw	Mampruli	AA	Hartell 1993; Chanard 2006
maw	Mampruli	PHOIBLE	Osbiston 1975
maz	Mazahua	SPA	Pike 1951; Spotts 1953
maz	MAZAHUA	UPSID	Pike 1951; Spotts 1953
mbe	Molalla	PHOIBLE	Pharris 2006
mbl	MAXAKALI	UPSID	Gudschinsky et al. 1970
mbo	Mboó	AA	Hartell 1993; Chanard 2006
mcf	Matses	PHOIBLE	Fleck 2003
mcp	Mekaa	AA	Hartell 1993; Chanard 2006
mcs	Mambay	PHOIBLE	Anonby 2006
mcu	MAMBILA	UPSID	Perrin and Hill 1969
mcu	Mambila (Cameroon)	AA	Hartell 1993; Chanard 2006

ISO 639-3	Language Name	Source	Reference
mda	Mada	PHOIBLE	Price 1989
mdd	MBUM	UPSID	Hagège 1970
mde	MABA	UPSID	Tucker and Bryan 1966
mdj	Mangbetu (Meje)	AA	Hartell 1993; Chanard 2006
mdx	DIZI	UPSID	Allan 1976a
mec	Mara	PHOIBLE	Heath 1981
men	Mende	AA	Hartell 1993; Chanard 2006
mfc	MBA-NE	UPSID	Pasch 1986
mff	Naki	PHOIBLE	Kum Nang 2002
mfn	Mbembe	AA	Hartell 1993; Chanard 2006
mfo	Mbe	PHOIBLE	Bamgbose 1967
mfz	Mabaan	PHOIBLE	Andersen 1992
mgd	Moru	PHOIBLE	Tucker 1967
mgg	Mpumpun	PHOIBLE	Djiafeua 1989
mgi	Jili	PHOIBLE	Blench 2006b
mgo	Metta	AA	Hartell 1993; Chanard 2006
mgr	Cilungu	PHOIBLE	Bickmore 2007
mgw	Matuumbi	PHOIBLE	Odden 2003
mhi	Pandikeri	AA	Hartell 1993; Chanard 2006
mhj	MOGHOL	UPSID	Weiers 1971
mhk	Mungaka	PHOIBLE	Awah 1997
mhr	Cheremis	SPA	Ristinen 1960
mhr	MARI	UPSID	Ristinen 1960
mhr	Cheremis	PHOIBLE	Sebeok and Ingemann 1961
mhw	Mbukushu	PHOIBLE	Sommer 2003
mhz	MOR	UPSID	Laycock 1978
mif	Mofu-Gudur	AA	Hartell 1993; Chanard 2006
mig	Mixtec	SPA	Hunter and Pike 1969

ISO 639-3	Language Name	Source	Reference
mig	MIXTEC	UPSID	Hunter and Pike 1969
mjg	MONGUOR	UPSID	Todaeva 1973
mkd	Macedonian	PHOIBLE	Friedman 2002
mkw	Munukutuba	AA	Hartell 1993; Chanard 2006
mky	Taba	PHOIBLE	Bowden and Hajek 1996
mky	East Makian	PHOIBLE	Bowden 1997b
mla	Tamambo	PHOIBLE	Riehl and Jauncey 2005
mlf	Mal	PHOIBLE	Singnoi 1988a
mlq	Mande	AA	Hartell 1993; Chanard 2006
mlt	Maltese	SPA	Borg 1973
mlv	Mwotlap	PHOIBLE	Francois 2001
mnb	Muna	PHOIBLE	van den Berg 1989
mnc	MANCHU	UPSID	Austin 1962
mnf	Mundani	AA	Hartell 1993; Chanard 2006
mnh	Mono	PHOIBLE	Olson 2004
mni	Manipuri	PHOIBLE	Chelliah 1992
mnk	Mandingo	PHOIBLE	Drame 1981
moa	Muan	AA	Hartell 1993; Chanard 2006
moc	Mocovi	PHOIBLE	Grondona 1998
mor	MORO	UPSID	Black and Black 1971; Schadeberg 1981a
mos	Moore	AA	Hartell 1993; Chanard 2006
mpb	MALAKMALAK	UPSID	Tryon 1974; Birk 1975
mph	Maung	SPA	Capell and Hinch 1970
mph	MAUNG	UPSID	Capell and Hinch 1970
mps	DADIBI	UPSID	MacDonald 1973
mpt	Mianmin	PHOIBLE	Smith and Weston 1974
mqs	WEST MAKIAN	UPSID	Watuseke 1976; Voorhoeve 1982

ISO 639-3	Language Name	Source	Reference
mri	Maori	SPA	Williams and Williams 1965; Hohepa 1967
mrq	North Marquesan	PHOIBLE	Zewen 1987
mrt	Margi	SPA	Hoffmann 1963
mrt	MARGI	UPSID	Hoffmann 1963; Maddieson 1987
mrw	MARANAO	UPSID	McKaughan 1958; McKaughan and Macaraya 1967
mtb	Agni Sanvi	AA	Hartell 1993; Chanard 2006
mtb	Agni Morofo	PHOIBLE	Quaireau 1987
mto	MIXE	UPSID	Crawford 1963; Schoenhals and Schoenhals 1965
muh	MüNdü	AA	Hartell 1993; Chanard 2006
muo	Bali-Kumbat	PHOIBLE	Kouonang 1983
mur	MURSI	UPSID	Turton and Bender 1976; Arensen 1982
mur	Murle	AA	Hartell 1993; Chanard 2006
mva	Manam	PHOIBLE	Turner 1986
mwf	MURINHPATHA	UPSID	Street and Mollinjin 1981
mwk	Maninka-Kan	AA	Hartell 1993; Chanard 2006
mwp	KALA LAGAW YA	UPSID	Wurm 1972b; Kennedy 1981
mwt	Moken	PHOIBLE	Veena 1980
mxu	Mada	PHOIBLE	Blench 2006a
mya	Burmese	SPA	Okell 1969
mya	BURMESE	UPSID	Okell 1969
mye	Myene	PHOIBLE	Jacquot et al. 1976
myk	Minyanka	AA	Hartell 1993; Chanard 2006
myp	PIRAHA	UPSID	Sheldon 1974; Rodrigues 1980; Everett 1982
myu	Mundurukú	PHOIBLE	Picanço 2005

ISO 639-3	Language Name	Source	Reference
mzk	Mambila (Nigeria)	AA	Hartell 1993; Chanard 2006
mzm	MUMUYE	UPSID	Shimizu 1983
mzn	Mazanderani	PHOIBLE	Mokhtarian 2004
mzp	MOVIMA	UPSID	Judy and Judy 1962
mzw	Mo (Deg)	AA	Hartell 1993; Chanard 2006
nab	SOUTHERN NAMBIQUARA	UPSID	Price 1976
nag	Nagamese	PHOIBLE	Bhattacharjya 2001
nam	Nganikurungkurr	PHOIBLE	Hoddinott and Kofod 1988
nan	XIAMEN	UPSID	Jiahua 1960
naq	Nama	SPA	Beach 1938
naq	NAMA	UPSID	Beach 1938; Ladefoged and Traill 1980
nas	Nasioi	SPA	Hurd and Hurd 1966
nas	NASIOI	UPSID	Hurd and Hurd 1966
nav	Navaho	SPA	Sapir and Hoijer 1967
nav	NAVAJO	UPSID	Sapir and Hoijer 1967
nbf	NAXI	UPSID	Bradley 1975; Jiang 1980
nbj	Bilinara	PHOIBLE	Nordlinger 1990
ncg	Nishgha	PHOIBLE	Tarpent 1987
ncj	NAHUATL	UPSID	Law 1955; Brockway 1963; Schumann and Garcia de Leon 1966
ncl	Michoacan Nahual	PHOIBLE	Sischo 1979
ncu	Chumburung	AA	Hartell 1993; Chanard 2006
ndb	Kensei Nsei	PHOIBLE	Akeriweh 2000
ndi	Samba Leko	PHOIBLE	Kong 2004
ndo	Ndonga	PHOIBLE	Sommer 2003
nds	Low German	PHOIBLE	Mierau 1965
ndv	NDUT	UPSID	Gueye 1986

ISO 639-3	Language Name	Source	Reference
neb	Toura	AA	Hartell 1993; Chanard 2006
nep	NEPALI	UPSID	Bandhu et al. 1971
new	NEWARI	UPSID	Hale and Hale 1969; Manandhar 1986
nez	Nez Perce	SPA	Aoki 1966, 1970b,a
nez	NEZ PERCE	UPSID	Aoki 1966, 1970b,a
nfr	Nafaanra	AA	Hartell 1993; Chanard 2006
nga	Ngbaka	AA	Hartell 1993; Chanard 2006
ngb	Ngbandi	AA	Hartell 1993; Chanard 2006
nge	Ngemba	PHOIBLE	Swiri 1998
ngi	NGIZIM	UPSID	Schuh 1972
nhb	Bɛ̀ŋ (Ngain)	AA	Hartell 1993; Chanard 2006
nhu	NONI	UPSID	Hyman 1981
nie	LUA	UPSID	Boyeldieu 1985
nig	Ngalakan	PHOIBLE	Baker 1999
nio	NGANASAN	UPSID	Castren 1966; Tereshchenko 1966b, 1979
niq	Nandi	PHOIBLE	Creider and Creider 1989
nir	NIMBORAN	UPSID	Anceaux 1965
niv	Gilyak	SPA	Panfilov 1962, 1968
niv	NIVKH	UPSID	Krejnovich 1937; Zinder and Matusevich 1937; Austerlitz 1956; Panfilov 1962, 1968
niz	Ningil	PHOIBLE	Manning and Saggers 1977
njo	AO	UPSID	Gurubasave Gowda 1972, 1975
nla	Ngombale	PHOIBLE	Voutsa 2003
nld	Dutch	PHOIBLE	Verhoeven 2005
nmg	Mvumbo	PHOIBLE	Ngue um 2002
nml	Ndemli	PHOIBLE	Ngoran 1999

ISO 639-3	Language Name	Source	Reference
nmm	Manange	PHOIBLE	Hildebrandt 2004
nmu	Maidu	SPA	Shipley 1956, 1964
nmu	MAIDU	UPSID	Shipley 1956, 1964
nmz	Nawdm	AA	Hartell 1993; Chanard 2006
nna	Nyangumata	SPA	O'Grady 1964
nnh	Ngyembɔɔn	AA	Hartell 1993; Chanard 2006
nnk	Nankina	PHOIBLE	Spaulding and Spaulding 1994
nnm	Namia	PHOIBLE	Feldpausch and Feldpausch 1992
nob	Norwegian	SPA	Vanvik 1972a,b
nob	NORWEGIAN	UPSID	Vanvik 1972a
noo	Nootka	SPA	Sapir and Swadesh 1939, 1955; Jacobsen 1969
noo	TSESHAHT	UPSID	Sapir and Swadesh 1939, 1955; Jacobsen 1969
noo	Nootka	PHOIBLE	Davidson 2002
nrb	NERA	UPSID	Bender 1968; Thompson 1976
nsh	Ngishe	PHOIBLE	Bolima 1998
nup	Nupe	AA	Hartell 1993; Chanard 2006
nus	Nuer	PHOIBLE	Frank 1999
nut	LUNGCHOW	UPSID	Li 1977a
nuv	Nuni	AA	Hartell 1993; Chanard 2006
nuy	Nunggubuyu	SPA	Hughes and Leeding 1971
nuy	NUNGGUBUYU	UPSID	Hughes and Leeding 1971
nwb	Niaboua	AA	Hartell 1993; Chanard 2006
nwe	Ngwe	PHOIBLE	Dunstan 1964
nxg	Ngad'a	PHOIBLE	Djawanai 1983
nxl	Southern Nuautl	PHOIBLE	Bolton 1990
nyi	NYIMANG	UPSID	Stevenson 1957; Tucker and Bryan 1966

ISO 639-3	Language Name	Source	Reference
nyl	Nyeu	PHOIBLE	Taweeporn 1998
nym	Kinyamwezi	PHOIBLE	Maganga and Schadeberg 1992
nyn	Runyankore	PHOIBLE	Poletto 1998
nyp	NYANGI	UPSID	Heine 1975b
oca	Ocaina	SPA	Agnew and Pike 1957
oca	OCAINA	UPSID	Agnew and Pike 1957
ogb	OGBIA	UPSID	Williamson 1970, 1972
ojg	Ojibwa	SPA	Bloomfield 1957
ojg	OJIBWA	UPSID	Bloomfield 1957
okr	Kirike	PHOIBLE	Blench 2005b
oku	Oku	PHOIBLE	Yensi 1996
okv	Orokaiva	PHOIBLE	Larsen and Larsen 1977
one	Oneida	SPA	Lounsbury 1953
ood	Pima	SPA	Hale 1959; Saxton 1963
ood	PAPAGO	UPSID	Hale 1959; Saxton 1963
oon	Öñge	PHOIBLE	Dasgupta and Sharma 1982
oru	ORMURI	UPSID	Efimov 1986
oss	Ossetian	PHOIBLE	Hettich 1997
ote	Otomi	SPA	Blight and Pike 1976
ozm	Kɔɔzime	AA	Hartell 1993; Chanard 2006
pac	PACOH	UPSID	Watson 1964
pae	Pagibete	PHOIBLE	Reeder 1998
pan	Punjabi	SPA	Gill and Gleason 1969
pao	Northern Paiute	PHOIBLE	Thornes 2003
par	Panamint	PHOIBLE	McLaughlin 1987
pau	Palauan	PHOIBLE	Josephs 1975
pav	Wari	PHOIBLE	Everett and Kern 1997
pay	PAYA	UPSID	Holt 1986

ISO 639-3	Language Name	Source	Reference
pbb	Paez	SPA	Gerdel 1973
pbb	PAEZ	UPSID	Gerdel 1973
pbh	PANARE	UPSID	Cauty 1974a,b, 1978
pbi	Podoko	AA	Hartell 1993; Chanard 2006
pbp	Pajade	PHOIBLE	Ducos 1974
pcc	Yay	SPA	Gedney 1965
pcc	YAY	UPSID	Gedney 1965
pcm	Nigerian Pidgin	PHOIBLE	Faraclas 1989
pej	Northern Pomo	PHOIBLE	O'Connor 1987
pes	Persian	SPA	Obolensky et al. 1963
pes	FARSI	UPSID	Obolensky et al. 1963
pex	Petats	PHOIBLE	Allen and Beason 1975
phl	Palula	PHOIBLE	Liljegren 2008
pib	Yine	PHOIBLE	Urquía Sebastián and Marlett 2008
pil	Yom	AA	Hartell 1993; Chanard 2006
plg	Pilagá	PHOIBLE	Vidal 2001b
plo	Oluta Popoluca	PHOIBLE	Zavala 2000
plt	Malagasy	SPA	Dahl 1952; Dyen 1971
plt	MALAGASY	UPSID	Dahl 1952; Dyen 1971
pmq	Northern Pame	PHOIBLE	Berthiaume 2003
pol	Polish	PHOIBLE	Jassem 2003
pom	Pomo	SPA	Moshinsky 1974
pom	POMO	UPSID	Moshinsky 1974
pon	POHNPEIAN	UPSID	Rehg 1981, 1984a,b
poq	Texistepec Popoluca	PHOIBLE	Reilly 2002
por	Portuguese	SPA	Head 1964; Camara 1972
pos	Popoluca de Sayula	PHOIBLE	Clark 1995

ISO 639-3	Language Name	Source	Reference
prk	PARAUK	UPSID	Diffloth 1980; Qiu Efeng 1980; Maddieson and Ladefoged 1985
prt	Pray	PHOIBLE	Singnoi 1988b
pst	Pashto	SPA	Shafeev 1964
pst	PASHTO	UPSID	Penzl 1955; Shafeev 1964; Grjunberg 1987
ptp	Patep	PHOIBLE	Adams and Lauck 1975
pwn	PAIWAN	UPSID	Ho 1977; Ferrel 1982
pww	PHLONG	UPSID	Cooke et al. 1976
pww	Pwo Karen	PHOIBLE	Naruemon 1995
quc	Quiche	PHOIBLE	Larsen 1988
qug	Chimborazo Quichua	PHOIBLE	Beukema 1975
quh	Quechua	SPA	Lastra 1968; Bills et al. 1969; Parker 1977
quh	QUECHUA	UPSID	Lastra 1968; Bills et al. 1969; Parker 1977
qui	QUILEUTE	UPSID	Powell 1975
qum	Sipakapense Maya	PHOIBLE	Barrett 1999
qvh	Huallaga (Huanuco) Quechua	PHOIBLE	Weber 1983
qwh	Huaylas	PHOIBLE	Levengood de Estrello and Larsen 1982
qxl	Salasaca Quichua	PHOIBLE	Masaquiza and Marlett 2008
qxw	Huanca	PHOIBLE	Wroughton 1996
rel	Rendille	AA	Hartell 1993; Chanard 2006
rgr	RESIGARO	UPSID	Allin 1976
rif	Shilha	SPA	Applegate 1958
rma	Rama	PHOIBLE	Grinevald 1990
ron	Rumanian	SPA	Agard 1958; Ruhlen 1973

ISO 639-3	Language Name	Source	Reference
ron	ROMANIAN	UPSID	Agard 1958; Ruhlen 1973; Tataru 1978
roo	ROTOKAS	UPSID	Firchow and Firchow 1969b
rro	RORO	UPSID	Bluhme 1970; Davis 1974
ruk	Che	PHOIBLE	Wilson 1996
run	Rundi	PHOIBLE	Rodegem 1967
rus	Russian	SPA	Halle 1959; Jones and Ward 1969
rus	RUSSIAN	UPSID	Halle 1959; Jones and Ward 1969
rut	RUTUL	UPSID	Dzhejranishvili 1967; Ibragimov 1978
rwr	Marwari	PHOIBLE	Magier 1983
sad	SANDAWE	UPSID	Dempwolff 1916; Tucker et al. 1977; Elderkin 1982
sae	Sabane	PHOIBLE	Antunes 2004
sag	SANGO	UPSID	Samarin 1967b,a
sag	Sango (CAF)	AA	Hartell 1993; Chanard 2006
sag	Sango (COD)	AA	Hartell 1993; Chanard 2006
sah	Yakut	SPA	Krueger 1962; Bohtlingk 1964
sah	YAKUT	UPSID	Krueger 1962; Bohtlingk 1964; Ubrjatova 1966
sas	Sasak	PHOIBLE	Jacq 1998
sba	Ngambai	AA	Hartell 1993; Chanard 2006
sbd	Samo de Toma	PHOIBLE	Platiel 1979
sbf	Shabo	PHOIBLE	Teferra 1991
sbs	Subiya	PHOIBLE	Baumbach 1997b
sed	Sedang	SPA	Smith 1968
sed	SEDANG	UPSID	Smith 1968
see	Seneca	SPA	Chafe 1967
see	SENECA	UPSID	Chafe 1967
sef	SENADI	UPSID	Welmers 1950

ISO 639-3	Language Name	Source	Reference
sef	Senoufo-Cebaara	PHOIBLE	Herington et al. 2009
sei	Seri	PHOIBLE	Marlett 2005
sel	SELKUP	UPSID	Katz 1975b
seq	Senoufo	AA	Hartell 1993; Chanard 2006
ser	Serrano	PHOIBLE	Hill 1967
ses	Songhai	SPA	Prost 1956
ses	SONGHAI	UPSID	Prost 1956; Williamson 1967
ses	Songhay, Koyraboro Senni	PHOIBLE	Heath 1999
set	Sentani	SPA	Cowan 1965
set	SENTANI	UPSID	Cowan 1965
sey	Secoya	PHOIBLE	Johnson and Levinsohn 1990
sgi	Nizaa	PHOIBLE	Endresen 1991
sgz	Sursurunga	PHOIBLE	Hutchisson and Hutchisson 1975
shb	SHIRIANA	UPSID	Migliazza and Grimes 1961
shi	SHILHA	UPSID	Applegate 1958
shs	SHUSWAP	UPSID	Kuipers 1974
sht	SHASTA	UPSID	Silver 1964
sid	Sidaama	PHOIBLE	Kawachi 2007
sin	Sinhalese	SPA	Coates and de Silva 1960
sin	SINHALESE	UPSID	Coates and de Silva 1960
sja	EPENA PEDEE	UPSID	Harms 1984, 1985
sjr	Siar-Lak	PHOIBLE	Rowe 2005
sjw	Shawnee	PHOIBLE	Andrews 1994
skd	SIERRA MIWOK	UPSID	Freeland 1951; Broadbent 1964
skf	Sakirabiá	PHOIBLE	Galucio 2001
skr	Siraiki	PHOIBLE	Shackle 1976
skv	Skou	PHOIBLE	Donohue 2004

ISO 639-3	Language Name	Source	Reference
slc	SALIBA	UPSID	Benaissa 1979
sld	Sissala (Burkina)	AA	Hartell 1993; Chanard 2006
sld	Sissala (Ghana)	AA	Hartell 1993; Chanard 2006
slu	Selaru	PHOIBLE	Coward 1990
slv	Slovene	PHOIBLE	Ŝuŝtarŝiĉ et al. 1995
sma	SAAMI	UPSID	Hasselbrink 1965; Kert 1971
smq	Samo	PHOIBLE	Daniel and Shaw 1977
sna	Shona	PHOIBLE	Fortune 1955
snd	Sindhi	PHOIBLE	Nihalani 1995
snk	Soninke (Mali)	AA	Hartell 1993; Chanard 2006
snk	Sooninke (Senegal)	AA	Hartell 1993; Chanard 2006
snk	Soninke (Kaedi)	PHOIBLE	Diagana 1995
snm	Madi	PHOIBLE	Blackings and Fabb 2003b
snn	SIONA	UPSID	Wheeler and Wheeler 1962
snv	Sa'ban	SPA	Clayre 1973
snv	SA'BAN	UPSID	Clayre 1973
snw	Sele	PHOIBLE	Allen 1973
som	Somali	SPA	Andrzejewsky 1955, 1956; Armstrong 1964
som	SOMALI	UPSID	Andrzejewsky 1955, 1956; Armstrong 1964; Cardona 1981; Farnetani 1981
spa	Spanish	SPA	Navarro 1961; Saporta and Contreras 1962; Harris 1969
spa	SPANISH	UPSID	Navarro 1961; Saporta and Contreras 1962; Harris 1969
spl	Selepet	SPA	McElhanon 1970a
spl	SELEPET	UPSID	McElhanon 1970a,b
spo	Spokan	PHOIBLE	Carlson 1972

ISO 639-3	Language Name	Source	Reference
spy	Sabaot	AA	Hartell 1993; Chanard 2006
sqt	SOCOTRI	UPSID	Leslau 1938; Johnstone 1975
squ	Squamish	SPA	Kuipers 1967
srq	Siriono	SPA	Priest 1968
srq	SIRIONO	UPSID	Priest 1968
srr	Sereer	AA	Hartell 1993; Chanard 2006
ssg	Seimat	PHOIBLE	Wozna and Wilson 2005
sso	Sesotho	PHOIBLE	Demuth 1992
stc	NAMBAKAENGO	UPSID	Wurm 1972a
str	Salish	SPA	Snyder 1968
str	Saanich	PHOIBLE	Montler 2005
stw	Satawalese	PHOIBLE	Roddy 2007
sue	SUENA	UPSID	Wilson 1969
sun	Sundanese	SPA	Robins 1953, 1957; Van Syoc 1959; Anderson 1972
suq	Suri	PHOIBLE	Bryant 1999
sur	Mwaghavul	AA	Hartell 1993; Chanard 2006
sus	Soso	AA	Hartell 1993; Chanard 2006
svr	Savara	PHOIBLE	Anonymous 1927
svs	SAVOSAVO	UPSID	Todd 1975
swe	Swedish	PHOIBLE	Engstrand 1990
swh	Swahili	SPA	Polome 1967
swh	Swahili	AA	Hartell 1993; Chanard 2006
swi	SUI	UPSID	Li 1948
sxm	Samre	PHOIBLE	Ploykaew 2001
sza	Semelai	PHOIBLE	Kruspe 1999
taj	TAMANG	UPSID	Mazaudon 1973
tam	Tamil	PHOIBLE	Keane 2004

ISO 639-3	Language Name	Source	Reference
taq	Tamasheq	AA	Hartell 1993; Chanard 2006
tav	Tatuyo	PHOIBLE	Bostrom 1998
tay	Atayal	SPA	Egerod 1966
tay	ATAYAL	UPSID	Egerod 1966
tba	HUARI	UPSID	Hanke 1956
tbi	TABI	UPSID	Tucker and Bryan 1966
tbz	Ditammari	AA	Hartell 1993; Chanard 2006
tca	Ticuna	SPA	Anderson 1959, 1962
tca	TICUNA	UPSID	Anderson 1959, 1962
tcb	Tanacross	PHOIBLE	Holton 2000b
tcy	TULU	UPSID	Bhat 1967
tdh	Thulung	PHOIBLE	Lahaussois 2002
ted	Kroumen TéPo	AA	Hartell 1993; Chanard 2006
tee	Tepehua, Huehuetla	PHOIBLE	Kung 2007
teh	TEHUELCHE	UPSID	Gerzenstein 1968
tel	Telugu	SPA	Krishnamurti 1961; Lisker 1963; Kelley 1963
tel	TELUGU	UPSID	Krishnamurti 1961; Kelley 1963; Lisker 1963; Sastry 1972; Kostic et al. 1977
tem	TEMNE	UPSID	Wilson 1961; Dalby 1966
tem	Themne	AA	Hartell 1993; Chanard 2006
teq	TEMEIN	UPSID	Tucker and Bryan 1966
tet	TETUN	UPSID	Morris 1984
tew	Tewa	SPA	Hoijer and Dozier 1949
tfi	Tɔfin	AA	Hartell 1993; Chanard 2006
tft	Ternate	PHOIBLE	Hayami-Allen 2001
tgc	TIGAK	UPSID	Beaumont 1979

ISO 639-3	Language Name	Source	Reference
tgl	Tagalog	SPA	Bloomfield 1917; Schachter and Otanes 1972
tgl	TAGALOG	UPSID	Bloomfield 1917; Schachter and Otanes 1972
tgw	Tagwana	PHOIBLE	Casimir 1988
tha	Thai	SPA	Noss 1954; Kruatrachue 1960; Abramson 1962; Noss 1964
tha	THAI	UPSID	Noss 1954; Haas 1956; Kruatrachue 1960; Abramson 1962; Haas 1964; Noss 1964
thm	So	PHOIBLE	Migliazza 1998b
thm	Thavung	PHOIBLE	Nuchanart 1998b
thv	TAMASHEQ	UPSID	Prasse 1972
tig	Tigre	SPA	Palmer 1962
tig	TIGRE	UPSID	Palmer 1962; Klingenheben 1966
tik	Tikar	AA	Hartell 1993; Chanard 2006
tiv	Tiv	PHOIBLE	Arnott 1968b
tiw	TIWI	UPSID	Osborne 1974a; Lee 1983, 1984
tiw	Tiwi	PHOIBLE	Osborne 1974b
tix	Tiwa	SPA	Trager 1971
tiy	TIRURAY	UPSID	Post 1966; Schlegel 1971
tlf	Telefol	SPA	Healey 1964
tli	TLINGIT	UPSID	Swanton 1909, 1911; Story and Naish 1973
tlo	JOMANG	UPSID	Schadeberg 1981b
tma	TAMA	UPSID	Tucker and Bryan 1966
tml	Asmat	SPA	Voorhoeve 1965
tml	ASMAT	UPSID	Voorhoeve 1965

ISO 639-3	Language Name	Source	Reference
tna	TACANA	UPSID	Van Wynen and de Van Wynen 1962; Key 1968
tnl	LENAKEL	UPSID	Lynch 1978
tob	Toba	PHOIBLE	Klein 1973
toi	Shanjo	PHOIBLE	Bosteon 2009
tol	Chasta Costa	SPA	Bright 1964
tol	Tolowa	PHOIBLE	Bommelyn 1997
ton	Tongan	PHOIBLE	Feldman 1978
top	Totonac	SPA	Aschmann 1946
top	TOTONAC	UPSID	Aschmann 1946
toq	Toposa	AA	Hartell 1993; Chanard 2006
tow	Jemez	PHOIBLE	Yumitani 1998
tpi	Tapiete	PHOIBLE	González 2005
tpm	TAMPULMA	UPSID	Bergman et al. 1969
tpt	Tepehua	PHOIBLE	Watters 1988
tpx	TLAPANEC	UPSID	Suárez 1983
tpy	TRUMAI	UPSID	Monod-Becquelin 1975
tpz	Tinputz	PHOIBLE	Hostetler and Hostetler 1975
tqo	TAORIPI	UPSID	Brown 1973
tqw	TONKAWA	UPSID	Hoijer 1946, 1949, 1972
trg	NEO-ARAMAIC	UPSID	Garbell 1965
trv	Sedik	PHOIBLE	Asal 1969
trw	Torwali	PHOIBLE	Lunsford 2001
tsi	TSIMSHIAN	UPSID	Dunn 1978; Hoard 1978; Dunn 1979; Mulder 1988
tsj	Tshangla	PHOIBLE	Andvik 1999
tsp	Toussian	AA	Hartell 1993; Chanard 2006
tsu	TSOU	UPSID	Tung 1964

ISO 639-3	Language Name	Source	Reference
tsz	Tarascan	SPA	Foster 1969
tsz	TARASCAN	UPSID	Foster 1969; Friedrich 1975
ttl	Totela	PHOIBLE	Baumbach 1997c
ttq	Tamajaq	AA	Hartell 1993; Chanard 2006
ttr	TERA	UPSID	Newman 1970
tun	Tunica	SPA	Haas 1941
tun	TUNICA	UPSID	Haas 1941
tuq	Teda	AA	Hartell 1993; Chanard 2006
tur	Turkish	SPA	Lees 1961; Swift 1963; Underhill 1976
tur	TURKISH	UPSID	Lees 1961; Swift 1963
tvd	Tsuvadi	PHOIBLE	Lovelace 1992
twf	PICURIS	UPSID	Hoijer and Dozier 1949; Trager 1971
txx	Tatana’	PHOIBLE	Dillon 1994
tyv	TUVA	UPSID	Sat 1966; Seglenmej 1979; Song 1982
tyv	Tuva	PHOIBLE	Harrison 2000b
tzh	Tzeltal	SPA	Kaufman 1971
tzh	TZELTAL	UPSID	Kaufman 1971
tzj	Tzutujil	PHOIBLE	Dayley 1985
tzo	Tzotzil, Chamula	PHOIBLE	Shklovsky 2005
ukr	Ukrainian	PHOIBLE	Pugh and Press 1999
ulw	Sumo	PHOIBLE	Green 1999
umb	Umbundu	PHOIBLE	Sommer 2003
ung	NGARINJIN	UPSID	Coate and Elkin 1974
unm	Delaware	SPA	Voegelin 1946
unr	Mundari	SPA	Gumperz and Bilibiri 1957
unr	MUNDARI	UPSID	Gumperz and Bilibiri 1957; Pinnow 1959
usa	Usarufa	PHOIBLE	Bee 1965a

ISO 639-3	Language Name	Source	Reference
ute	Southern Ute	PHOIBLE	Oberly 2008
uvh	Urii	PHOIBLE	Webb 1974
uzn	UZBEK	UPSID	Sjoberg 1962, 1963
vag	Vagala	AA	Hartell 1993; Chanard 2006
vai	Vai	PHOIBLE	Welmers 1976
vam	VANIMO	UPSID	Ross 1980
var	Warihio	PHOIBLE	Armendáriz 2005
vie	Vietnamese	SPA	Thompson 1965
vie	VIETNAMESE	UPSID	Thompson 1965; Nguyen 1974
vmb	MBABARAM	UPSID	Dixon 1966a,b
vut	Vute	AA	Hartell 1993; Chanard 2006
wan	Wan	AA	Hartell 1993; Chanard 2006
wao	WAPPO	UPSID	Sawyer 1965
wap	Wapishana	SPA	Tracy 1972
wap	WAPISHANA	UPSID	Tracy 1972
was	Washo	PHOIBLE	Jacobsen 1964
way	Wayana	PHOIBLE	da Silva Tavares 2005
wba	WARAO	UPSID	Osborn 1966
wbm	Wa	PHOIBLE	Tantiwithipakorn 1998
wgi	WAHGI	UPSID	Phillips 1976
wic	Wichita	SPA	Garvin 1950
wic	WICHITA	UPSID	Garvin 1950; Rood 1975
wim	Wik-Munkan	SPA	Sayers and Godfrey 1964
wim	WIK-MUNKAN	UPSID	McConnel 1945; Sayers and Godfrey 1964
wit	WINTU	UPSID	Broadbent and Pitkin 1964
wiy	WIYOT	UPSID	Teeter 1964
wms	Wambon	PHOIBLE	Vries 1992

ISO 639-3	Language Name	Source	Reference
wnc	WANTOAT	UPSID	Davis 1969
wnu	USAN	UPSID	Reesink 1987
wob	Wobé	AA	Hartell 1993; Chanard 2006
woc	Wogeo	PHOIBLE	Exter 2003
woi	WOISIKA	UPSID	Stokhof 1979
wok	Longto	PHOIBLE	Kuperus 1985
wol	Wolof	SPA	Manessy and Sauvageot 1963; Ward 1963; Sauvageot 1965
wol	WOLOF	UPSID	Manessy and Sauvageot 1963; Ward 1963; Sauvageot 1965
wol	Wolof	AA	Hartell 1993; Chanard 2006
wos	Hanga Hundi	PHOIBLE	Wendel 1993
wrs	WARIS	UPSID	Brown 1988
wrz	WARAY	UPSID	Harvey 1986
wti	BERTA	UPSID	Triulzi et al. 1976
wtm	Mewati	PHOIBLE	Gusain 2003
wuu	Wu	SPA	Chao 1970
wuu	CHANGZHOU	UPSID	Chao 1970
wwa	Waama	AA	Hartell 1993; Chanard 2006
wya	Huron	PHOIBLE	Lagarde 1980
wyb	NGIYAMBAA	UPSID	Donaldson 1980
xan	Xamtanga	PHOIBLE	Fallon 2009
xaw	KAWAIISU	UPSID	Zigmond et al. 1988
xho	Xhosa	PHOIBLE	Gowlett 2003
xmf	Mingrelian	PHOIBLE	Harris 1991
xmt	Matbat	PHOIBLE	Remijsen 2002
xom	KOMA	UPSID	Tucker and Bryan 1966
xon	Konkomba	AA	Hartell 1993; Chanard 2006

ISO 639-3	Language Name	Source	Reference
xpe	Kpelle	SPA	Welmers 1962, 1973
xpe	KPELLE	UPSID	Welmers 1962; Hyman 1973; Welmers 1973
xpe	Kpelle	AA	Hartell 1993; Chanard 2006
xrb	Kar	PHOIBLE	Wichser 1994
xsm	Kasem	AA	Hartell 1993; Chanard 2006
xsm	Kasim	AA	Hartell 1993; Chanard 2006
xsu	Sanumá	PHOIBLE	Borgman 1990
xtc	Katcha	SPA	Stevenson 1957; Tucker and Bryan 1966
xtc	KADUGLI	UPSID	Abdalla 1973
xub	Betta Kurumba	PHOIBLE	Coelho 2003
xwa	Kwaza	PHOIBLE	van der Voort 2004
xwe	Xwela	PHOIBLE	Capo 1991
xwl	Western Xwla	PHOIBLE	Capo 1991
xxk	Kéo	PHOIBLE	Baird 2002
yad	YAGUA	UPSID	Payne 1985
yal	Jalonke	PHOIBLE	Lüpke 2005
yam	Yamba	AA	Hartell 1993; Chanard 2006
yao	Yao	PHOIBLE	Odden 2003
yap	Yapese	PHOIBLE	Ballantyne 2005
yaq	YAQUI	UPSID	Crumrine 1961; Johnson 1962
yas	Nugunu	AA	Hartell 1993; Chanard 2006
yat	Yambɛta	AA	Hartell 1993; Chanard 2006
yba	Yala	PHOIBLE	Armstrong 1968
ybb	Yemba	AA	Hartell 1993; Chanard 2006
ycn	YUCUNA	UPSID	Schauer and Schauer 1967
ydd	Standard Yiddish	PHOIBLE	Kleine 2003
yer	TAROK	UPSID	Robinson 1976

ISO 639-3	Language Name	Source	Reference
yer	Tarok	AA	Hartell 1993; Chanard 2006
yey	Yeyi	PHOIBLE	Baumbach 1997d
ygr	YAGARIA	UPSID	Renck 1967, 1975; Haiman 1980
yii	YIDINY	UPSID	Dixon 1977
ykg	YUKAGHIR	UPSID	Krejnovich 1958, 1968a
yll	Yil	PHOIBLE	Martens and Tuominen 1977
ymm	Maay	PHOIBLE	Paster 2006
ynn	YANA	UPSID	Sapir and Swadesh 1960
yns	Yanzi	PHOIBLE	Rottland 1977
yor	YORUBA	UPSID	Bamgbose 1966
yor	Yorouba (Benin)	AA	Hartell 1993; Chanard 2006
yor	Yorouba (Nigeria)	AA	Hartell 1993; Chanard 2006
yrb	YAREBA	UPSID	Weimer and Weimer 1972
yre	Yaouré	AA	Hartell 1993; Chanard 2006
yrk	Yurak	SPA	Hajdú 1963; Ristinen 1965; Décsy 1966; Ristinen 1968; Katz 1975a
yrk	NENETS	UPSID	Hajdú 1963; Ristinen 1965; Décsy 1966; Tereshchenko 1966a; Ristinen 1968; Katz 1975a
yss	YESSAN-MAYO	UPSID	Foreman and Marten 1973
yua	YUCATEC	UPSID	Straight 1976
yuc	Yuchi	SPA	Crawford 1973; Ballard 1975
yuc	YUCHI	UPSID	Crawford 1973; Ballard 1975
yuc	Yuchi	PHOIBLE	Linn 2001
yue	Cantonese	SPA	Cheng 1973b
yue	TAISHAN	UPSID	Chao 1947, 1951; Cheng 1973b; Chan 1980
yuk	Yuki	PHOIBLE	Schlicter 1985

ISO 639-3	Language Name	Source	Reference
yul	YULU	UPSID	Boyeldieu 1987
yum	Yuma	PHOIBLE	Halpern 1944
yur	Yurok	PHOIBLE	Robins 1958
yux	Yukaghir	SPA	Krejnovich 1958, 1968a
yux	Yukaghir (Kolyma)	PHOIBLE	Maslova 2003b
yuz	Yuracure	PHOIBLE	van Gijn 2006
yva	YAWA	UPSID	Jones 1986
ywn	Shanenawa	PHOIBLE	Cândido 2004
zab	Tlacolula Valley Zapotec	PHOIBLE	Lillehaugen 2006
zmr	Maranungku	SPA	Tryon 1970
zne	AZANDE	UPSID	Tucker and Hackett 1959
zne	Zande	AA	Hartell 1993; Chanard 2006
zoc	ZOQUE	UPSID	Wonderly 1951
zoh	Zoque	SPA	Wonderly 1951
zoh	San Miguel Chimalapa Zoque	PHOIBLE	Johnson 2000
zpq	San Bartolomé Zoogocho Zapotec	PHOIBLE	Sonnenschein 2004
zsm	Malay	SPA	Verguin 1967; Macdonald and Soenyono 1967
ztp	Zapotec	PHOIBLE	Beam de Azcona 2004
zts	Tilquiapan Zapotec	PHOIBLE	Merrill 2008
zul	Zulu	SPA	Doke 1926, 1961
zul	ZULU	UPSID	Doke 1926, 1961; Rycroft and Ngcobe 1979
zun	Zuni	SPA	Newman 1965
zun	ZUNI	UPSID	Newman 1965; Walker 1972

Appendix C

PHOIBLE SEGMENT CONVENTIONS

In this appendix I describe the general conventions that were used to encode segments in inventories that were added to the PHOIBLE data set. I begin by explaining the segment and diacritic ordering that was used. I then address general consonant- and vowel-specific decisions, including which symbols were used to indicate sounds not officially in the International Phonetic Alphabet (IPA; International Phonetic Association 2005).[1] Lastly, I briefly discuss marginal sounds and how they are marked in PHOIBLE.

C.1 Diacritic ordering

Each segment type that is composed of more than one character is first normalized into a canonical decomposition form that adheres to the Unicode Normalization Form D (NFD; The Unicode Consortium 2007).[2] The diacritic ordering conventions I describe below deal with Unicode characters that are not in the "Combining Diacritical Marks" block. The logical ordering of Combining Diacritical Marks is handled by normalization into NFD. Characters sequences that are not handled by NFD must be explicitly ordered, including characters from the "Spacing Modifier Letters" block, which may appear as diacritics to the user. The ordering is influenced by the linguistic literature and to my knowledge the IPA does not explicitly state in which order diacritics should appear in segments.

If a segment type contains more than one rightward diacritic, I use this order:

- unreleased/lateral release/nasal release → palatalized → labialized → velarized → pharyngealized → aspirated/ejective → long

[1] See also Appendices E and F for SPA and UPSID$_{451}$ specific notes. Appendix D provides a list of the Unicode IPA characters used in segments in inventories in PHOIBLE.

[2] See discussion in Section 2.1.4.

For example, a labialized aspirated long alveolar plosive: < tʷʰː >. If a segment type contains more than one diacritic below the base segment:

- the place feature is applied first (dental, laminal, apical, fronted, backed, lowered, raised), then the laryngeal setting (voiced, voiceless, creaky voice, breathy voice), and finally the syllabic or non-syllabic marker (for vowels, ATR gets put on between place and laryngeal setting)

For example, a creaky voiced syllabic dental nasal: < n̪̰̩ >.

C.2 Consonants

There are some common encoding errors that occur when linguists use the (Latin-based) keyboard to input certain IPA symbols that Unicode has assigned to different code points. These include:

- the IPA symbol <ɡ> latin small letter script g (U+0261) is not the same code point as keyboarded <g> latin small letter g (U+0067)

- the IPA symbol <ǃ> latin letter retroflex click[3] (U+01C3) is not the same code point as keyboarded <!> exclamation mark (U+0021)

- the IPA symbol <ǀ> latin letter dental click (U+01C0) is not the same code point as keyboarded <|> vertical line (U+007C)

- the IPA symbol <ʼ> modifier letter apostrophe (U+02BC) is not the same code point as keyboarded <'> apostrophe (U+0027)

- the IPA symbol <ː> modifier letter triangular colon (U+02D0) is not the same code point as keyboarded <:> colon (U+003A)

Other segment conventions relevant to consonants are given below by subsection.

[3] In the IPA, the <ǃ> is an alveolar or postalveolar click, not a retroflex click as stated in the Unicode Standard.

C.2.1 Aspiration

For aspiration, the conventions include:

- Aspirated: pʰ
- Preaspirated: ʰt
- Breathy release: tɦ

C.2.2 Double articulations

I do not currently use a "tie bar", i.e. combining double inverted breve (U+0361) or combining double breve below (U+035C), to signal double articulations (e.g. affricates, clicks and diphthongs). So for example, <k͡p> and <t͜s> appear as <kp> and <ts> in inventories in PHOIBLE.

Affricates are marked for homorganic place of articulation. For example, in SPA the "t/s-hacek-prenasalized" is indicated by the symbol <n̠t̠ʃ> and the "voiceless retroflex sibilant affricate" in UPSID_{451} is signaled by <ʈʂ>.

C.2.3 Fricatives

I use a lowered diacritic, the <o̞> combining down tack below (U+031E), with a fricative to make an approximant, e.g. SPA's "beta-approximant" looks like <β̞>. The raised diacritic is also used with the pharyngeal fricative to indicate a voiced pharyngeal plosive <ʕ̝>.

All "affricated" trills and clicks are marked with the non-IPA diacritic <o͓> combining x below (U+0353), which I use to indicate "frictionalized". For example "r-flap-fricative" in SPA and "voiced alveolar fricative flap" in UPSID_{451} are both indicated as <ɾ͓>.

UPSID_{451} forces the distinction between sibilant and non-sibilant fricatives, so another non-IPA diacritic was selected. To mark "non-sibilant" fricatives, I use the < o͇ > combining equals sign below (U+0347), e.g. "r-fricative" is <z͇>.

C.2.4 Glottalization

Glottalization conventions include:

- Preglottalized: ˀd
- Glottalized / postglottalized: dˀ
- Creaky voiced / laryngealized: d̰

C.2.5 Nasalization

For prenasalized consonants, i.e. homorganic nasals, I use <NC> where <N> is a nasal that agrees in place of articulation with the following consonant, e.g. <mb>, <nd>, <ŋg>, etc. The character <ⁿ> SUPERSCRIPT LATIN SMALL LETTER N (U+8319) is used to indicate nasal release, e.g. the “d-nasal-release” in UPSID$_{451}$ is given as <dⁿ>.

C.2.6 Clicks

Clicks are ordered with the voice setting first:

- <k> indicates voiceless
- <g> indicates voiced
- <ŋ> indicates nasal

Following the voice setting, the place/manner of the click is indicated, e.g. a voiceless alveolar click is encoded as <k!>. Laryngeal modifiers are placed on the voice setting and diacritics for place are placed on the symbol for the click. For example, a “voiceless nasal palatoalveolar click”: <ŋ̊!̠>.

C.2.7 Labialized

Labialized segments are represented with the <ʷ> modifier letter small w (U+02B7), e.g. the "labialized voiceless labio-velar plosive" in UPSID$_{451}$ is <k͡pʷ>. For velarized segments I use the <ˠ> modifier letter small gamma (U+02E0), e.g. SPA's "d-velarized" is <dˠ>. Labiovelarized segments use the combination of both space modifying characters in this order: <ʷˠ>.

C.3 Vowels

When a low back unrounded vowel appears in a phonological description, I use the character <ɑ> latin small letter alpha (U+0251), even if the author used the keyboard <a> in his or her phoneme inventory chart (which seems to be the case more often than not).

For diphthongs I use <i> or <u> and not <j> or <w> to indicate the glide component of the diphthong. In cases in which this leads to a sequence of two identical vowels, I use the non-syllabic diacritic marker <o̯> combining inverted breve below (U+032F), e.g. SPA's "i/yod" is marked with <ii̯>. Long vowels are marked with the length diacritic <ː>, e.g. SPA's "iota-creaky voice-long" is <ɪ̰ː>.

C.4 Marginal phonemes

Marginal phonemes are those that behave notably different phonologically than the majority of segments found in a particular language. Language contact factors contribute to marginal phonemes. For example, loanwords containing non-native sounds can introduce maringal phonemes into the borrowing language. There are varying degrees of marginalism; see discussion in Jelaska and Machata (2005). For PHOIBLE it would be ideal to create a ranking or vocabulary for varying degrees of marginal status.[4] To do so, I have collected any remarks about the marginality of segments as described in the resources from which I extracted inventories. However, since different authors use different descriptions of marginality, these have to be fit into some type of ranking. I propose adding this information in a future release of PHOIBLE. Currently I simply mark any type of phoneme

[4] Perhaps along the line of "anomalous" segments in UPSID (Maddieson, 1984, 170).

described as marginal or loan by an author of a language description by enclosing those segments in less-than and greater-than symbols < >.

Appendix D

UNICODE IPA DESCRIPTION TABLE

The table below provides a complete and unique list of the Unicode characters that appear in the PHOIBLE data set. The table also contains some characters that appear in IPA but that do not appear in inventories in PHOIBLE and it contains any additional characters that appear in the Hayes 2009 extended feature set. The "Glyph" column provides a visual representation of each Unicode character and in the "Visual" column I have added a base character in cases of diacritics. The "Decimal" and "Hex" columns provide the Unicode code point of each character. The "Class" column is the class of segment that I have manually assigned to each character. Note that a character like ʰ that marks aspiration is assigned the class consonant so that my algorithm that automatically assigned a segment class to each segment type in PHOIBLE will tag pre-aspirated consonants as "consonant". Lastly in the "Notes" column I provide any clarifications that I thought would be helpful.

Glyph	Visual	Decimal	Hex	Class	Notes
|	|	124	007C	NULL	UPSID ``or" marker, e.g. t|t̪ (t or dental t)
*	*	42	002A	consonant	archi-phoneme marker
L	L	76	004C	consonant	archi-phoneme
N	N	78	004E	consonant	archi-phoneme
R	R	82	0052	consonant	archi-phoneme
ˈ	ˈ	712	02C8	diacritic	(primary) stress mark
ˌ	ˌ	716	02CC	diacritic	secondary stress
˞	˞	734	02DE	diacritic	rhotacized
˥	˥	741	02E5	tone	extra high tone
˦	˦	742	02E6	tone	high tone

Glyph	Visual	Decimal	Hex	Class	Notes
˧	˧	743	02E7	tone	mid tone
˨	˨	744	02E8	tone	low tone
˩	˩	745	02E9	tone	extra low tone
↑	↑	8593	2191	tone	
↓	↓	8595	2193	tone	
ː	ː	720	02D0	diacritic	length mark
ˑ	ˑ	721	02D1	diacritic	half-length
a	a	97	0061	vowel	
æ	æ	230	00E6	vowel	
ɐ	ɐ	592	0250	vowel	
ɑ	ɑ	593	0251	vowel	
ɒ	ɒ	594	0252	vowel	
b	b	98	0062	consonant	
ʙ	ʙ	665	0299	consonant	
ɓ	ɓ	595	0253	consonant	
c	c	99	0063	consonant	
ç	ç	231	00E7	consonant	
ɕ	ɕ	597	0255	consonant	
d	d	100	0064	consonant	
ð	ð	240	00F0	consonant	
ɖ	ɖ	598	0256	consonant	
ɗ	ɗ	599	0257	consonant	
ᶑ	ᶑ	7569	1D91	consonant	
e	e	101	0065	vowel	
ə	ə	601	0259	vowel	
ɛ	ɛ	603	025B	vowel	
ɘ	ɘ	600	0258	vowel	

Glyph	Visual	Decimal	Hex	Class	Notes
ɚ	ɚ	602	025A	vowel	
ɜ	ɜ	604	025C	vowel	
ɝ	ɝ	605	025D	vowel	
ɞ	ɞ	606	025E	vowel	
ɤ	ɤ	612	0264	vowel	
f	f	102	0066	consonant	
ɡ	ɡ	609	0261	consonant	
ɢ	ɢ	610	0262	consonant	
ɠ	ɠ	608	0260	consonant	
ʛ	ʛ	667	029B	consonant	
ɣ	ɣ	611	0263	consonant	
ˠ	ˠ	736	02E0	diacritic	velarized
h	h	104	0068	consonant	
ʰ	ʰ	688	02B0	consonant	
ħ	ħ	295	0127	consonant	
ʜ	ʜ	668	029C	consonant	
ɦ	ɦ	614	0266	consonant	
ʱ	ʱ	689	02B1	diacritic	breathy-voice-aspirated
ɧ	ɧ	615	0267	consonant	
i	i	105	0069	vowel	
ɪ	ɪ	618	026A	vowel	
ɨ	ɨ	616	0268	vowel	
j	j	106	006A	consonant	
ʲ	ʲ	690	02B2	diacritic	palatalized
ʝ	ʝ	669	029D	consonant	
ɟ	ɟ	607	025F	consonant	
ʄ	ʄ	644	0284	consonant	

Glyph	Visual	Decimal	Hex	Class	Notes
k	k	107	006B	consonant	
l	l	108	006C	consonant	
ˡ	ˡ	737	02E1	diacritic	
ʟ	ʟ	671	029F	consonant	
ɫ	ɫ	619	026B	consonant	
ɬ	ɬ	620	026C	consonant	
ɭ	ɭ	621	026D	consonant	
ɮ	ɮ	622	026E	consonant	
ʎ	ʎ	654	028E	consonant	
m	m	109	006D	consonant	
ɱ	ɱ	625	0271	consonant	
n	n	110	006E	consonant	
ⁿ	ⁿ	8319	207F	diacritic	
ɴ	ɴ	628	0274	consonant	
ɲ	ɲ	626	0272	consonant	
ɳ	ɳ	627	0273	consonant	
ŋ	ŋ	331	014B	consonant	
o	o	111	006F	vowel	
ø	ø	248	00F8	vowel	
œ	œ	339	0153	vowel	
ɶ	ɶ	630	0276	vowel	
ɔ	ɔ	596	0254	vowel	
ɵ	ɵ	629	0275	vowel	
p	p	112	0070	consonant	
ɸ	ɸ	632	0278	consonant	
q	q	113	0071	consonant	
r	r	114	0072	consonant	

Glyph	Visual	Decimal	Hex	Class	Notes
ʀ	ʀ	640	0280	consonant	
ɹ	ɹ	633	0279	consonant	
ʴ	ʴ	692	02B4	diacritic	rhotacized
ɺ	ɺ	634	027A	consonant	
̢	ɺ̢	802	0322	diacritic	
ɻ	ɻ	635	027B	consonant	
ɽ	ɽ	637	027D	consonant	
ɾ	ɾ	638	027E	consonant	
ʁ	ʁ	641	0281	consonant	
s	s	115	0073	consonant	
ʂ	ʂ	642	0282	consonant	
ʃ	ʃ	643	0283	consonant	
t	t	116	0074	consonant	
ʈ	ʈ	648	0288	consonant	
u	u	117	0075	vowel	
ʉ	ʉ	649	0289	vowel	
ɥ	ɥ	613	0265	consonant	
ɯ	ɯ	623	026F	vowel	
ɰ	ɰ	624	0270	consonant	
ʊ	ʊ	650	028A	vowel	
v	v	118	0076	consonant	
ʋ	ʋ	651	028B	consonant	
ⱱ	ⱱ	11377	2C71	consonant	
ʌ	ʌ	652	028C	vowel	
w	w	119	0077	consonant	
ʷ	ʷ	695	02B7	diacritic	labialized
ʍ	ʍ	653	028D	consonant	

Glyph	Visual	Decimal	Hex	Class	Notes
x	x	120	0078	consonant	
̆	x̆	774	0306	diacritic	extra-short
̊	x̊	778	030A	diacritic	voiceless (use if character has descender)
̈	ẍ	776	0308	diacritic	centralized
̃	x̃	771	0303	consonant	
̚	x̚	794	031A	diacritic	not audibly released
̽	x̽	829	033D	diacritic	mid-centralized
̘	x̘	792	0318	diacritic	advanced tongue root
̙	x̙	793	0319	diacritic	retracted tongue root
̜	x̜	796	031C	diacritic	less rounded
̝	x̝	797	031D	diacritic	raised
̞	x̞	798	031E	diacritic	lowered
̟	x̟	799	031F	diacritic	advanced
̠	x̠	800	0320	diacritic	retracted
̩	x̩	809	0329	diacritic	syllabic
̪	x̪	810	032A	diacritic	dental
̬	x̬	812	032C	diacritic	voiced
̯	x̯	815	032F	diacritic	non-syllabic
̺	x̺	826	033A	diacritic	apical
̻	x̻	827	033B	diacritic	laminal
̼	x̼	828	033C	diacritic	linguolabial
̤	x̤	804	0324	diacritic	breathy voiced
̥	x̥	805	0325	diacritic	voiceless
̰	x̰	816	0330	diacritic	creaky voiced
̴	x̴	820	0334	diacritic	velarized or pharyngealized
̹	x̹	825	0339	diacritic	more rounded
͜	x͜x	860	035C	diacritic	tie bar below

Glyph	Visual	Decimal	Hex	Class	Notes
͡	x͡x	865	0361	diacritic	tie bar above
y	y	121	0079	vowel	
ʏ	ʏ	655	028F	vowel	
z	z	122	007A	consonant	
ʐ	ʐ	656	0290	consonant	
ʑ	ʑ	657	0291	consonant	
ʒ	ʒ	658	0292	consonant	
ʔ	ʔ	660	0294	consonant	
ʼ	ʼ	700	02BC	consonant	
ʕ	ʕ	661	0295	consonant	
ˤ	ˤ	740	02E4	consonant	
ʡ	ʡ	673	02A1	consonant	
ʢ	ʢ	674	02A2	consonant	
ǀ	ǀ	448	01C0	consonant	
ǁ	ǁ	449	01C1	consonant	
ǂ	ǂ	450	01C2	consonant	
ǃ	ǃ	451	01C3	consonant	
ʘ	ʘ	664	0298	consonant	
β	β	946	03B2	consonant	
θ	θ	952	03B8	consonant	
χ	χ	967	03C7	consonant	
ᶾ	ᶾ	7614	1DBE	diacritic	
ᶣ	ᶣ	7587	1DA3	diacritic	
͓	o͓	851	0353	diacritic	fricated marker
͈	x͈	840	0348	diacritic	used in SPA to represent "tense" consonants
͉	x͉	841	0349	diacritic	used in SPA to represent "lax" consonants

Glyph	Visual	Decimal	Hex	Class	Notes
ᴅ	ᴅ	7429	1D05	consonant	used to represent a tap as distinguished from flap in UPSID
̧	ç	807	0327	diacritic	Unicode decomposition decomposes c-cedilla into a <c> and a cedilla
͇	x͇	839	0347	diacritic	non-sibilant marker on obstruents in UPSID
ᴴ	xᴴ	7476	1D34	diacritic	epiglottal
ȵ	ȵ	565	0235	consonant	not in an inventory; in extended Hayes
ˀ	ˀx	704	02C0	consonant	pre-glottalized

Appendix E

SPA AND IPA SEGMENT CORRESPONDENCES

For the mapping of SPA segment descriptions to Unicode IPA segments, the following points should be taken into consideration (the symbol <o> is used as a place holder for diacritics regardless if they apply to consonants, vowels or both):

- “aspirated-weak” is not distinguished from “aspirated”
- “half-voiced” is not distinguished from “voiced”
- “nasalized-weak” is not distinguished from “nasalized”
- “backed” is mapped to combining minus below (U+0320) <o̠>
- “retracted” is mapped to combining right tack below (U+0319) <o̙>
- “glottalized” and “postlottalized” is mapped to modifier letter glottal stop (U+02C0) <ˀ>
- “preglottalized” is mapped to the same character but it precedes the segment that it modifies
- “creaky” is mapped to combining tilde below (U+0330) <o̰>
- “lax” is mapped to combining left angle below (U+0349) <o͉>
- “tense” is mapped to combining double vertical line below (U+0348) <o͈>
- “uvularized” is mapped to combining tilde overlay (U+0334) <o̴>, which technically represents “velarization” or “pharyngealization” in the IPA

- voiceless implosives are represented with voiced implosive glyphs with a devoicing diacritic (consistent with IPA usage), e.g. <ɓ̥>

- non-strident coronal fricatives are represented as their strident counterparts with COMBINING EQUALS SIGN BELOW (U+0347) <o͇>

- affricates are homorganic for place of articulation, e.g. [ts] and [t̠ʃ]

- diphthongs use [i] or [u] and not [j] or [w], e.g. [ai]; the non-syllabic diacritic is used for the glide portion of the diphthong, e.g. [ii̯]

The full list of SPA segment descriptions and IPA interpretations is given below.

SPA code	Unicode IPA
a	a
a-backed	a̠
a-breathy voice	a̤
a-breathy voice-long	a̤ː
a-creaky voice	a̰
a-creaky voice-long	a̰ː
a-front	a̟
a-front-half-voice-long	a̟ː
a-front-long	a̟ː
a-front-long-nasalized	ã̟ː
a-front-long-retracted	a̟̠ː
a-front-nasalized	ã̟
a-front-nasalized-weak	ã̟
a-front-over-short	ă̟
a-front-retroflexed	a̟˞
a-fronted	a̟
a-glide/e	a̯e

SPA code	Unicode IPA
a-glide/schwa	a̯ə
a-half-long	aˑ
a-half-voice	a
a-half-voice-half-long	aˑ
a-half-voice-long	aː
a-long	aː
a-long-nasalized	ãː
a-long-nasalized-weak	ãː
a-long/yod	aːi
a-nasalized	ã
a-nasalized-weak	ã
a-over-short	ă
a-over-short-nasalized	ã̆
a-retroflexed	a˞
a-voiceless	ḁ
a-voiceless-long	ḁː
a/yod	ai
alpha	ɒ
alpha-long	ɒː
alpha-long-nasalized	ɒ̃ː
alpha-nasalized	ɒ̃
alpha-over-short	ɒ̆
alpha-unrounded	ɑ
alpha-unrounded-half-long-nasalized	ɑ̃ˑ
alpha-unrounded-long	ɑː
alpha-unrounded-long-nasalized	ɑ̃ː
alpha-unrounded-long-uvularized	ꬰː

SPA code	Unicode IPA
alpha-unrounded-nasalized	ɑ̃
alpha-unrounded-nasalized-retroflexed	ɑ̃˞
alpha-unrounded-over-short	ɑ̆
alpha-unrounded-uvularized	ꬰ
alpha-unrounded-voiceless	ɑ̥
ash	æ
ash-breathy voice	æ̤
ash-dot	ɐ
ash-dot-creaky voice	ɐ̰
ash-dot-long	ɐː
ash-dot-nasalized	ɐ̃
ash-dot-nasalized-retroflexed	ɐ̃˞
ash-dot-over-short	ɐ̆
ash-dot-retroflexed	ɐ˞
ash-dot-voiceless	ɐ̥
ash-dot/yod	ɐi
ash-half-voice-long	æː
ash-long	æː
ash-long-nasalized	æ̃ː
ash-nasalized	æ̃
ash-over-short	æ̆
ash-pharyngealized	æˤ
ash-trema	ɶ̞
ash-trema-long	ɶ̞ː
ash/e-glide	æe̯
ash/e-glide-breathy voice	æe̤̯
ash/e-glide-nasalized	æẽ̯

SPA code	Unicode IPA
ash/schwa-glide	æə̯
b	b
b-aspirated-half-voice	p̬ʰ
b-breathy voice	b̤
b-breathy voice-long	b̤ː
b-breathy voice-palatalized	b̤ʲ
b-creaky voice	b̰
b-glottalized	bˀ
b-half-voice	b
b-implosive	ɓ
b-implosive-labialized	ɓʷ
b-labialized	bʷ
b-labiodental	b̪
b-labiovelarized	bʷˠ
b-lateral-release	bˡ
b-lax	b̙
b-long	bː
b-long-labialized	bʷː
b-long-labialized-pharyngealized	bʷˤː
b-long-pharyngealized	bˤː
b-nasal-release	bⁿ
b-palatalized	bʲ
b-pharyngealized	bˤ
b-postglottalized	bˀ
b-preglottalized	ˀb
b-preglottalized-labialized	ˀbʷ
b-prenasalized	mb

SPA code	Unicode IPA
b-prenasalized-breathy voice	mb̤
b-prenasalized-labialized	mbʷ
b-prenasalized-palatalized	mbʲ
b-syllabic	b̩
b-tense	b͈
b-tense-long	b͈ː
b-unreleased	b̚
b-unreleased-half-voice	b̚
b-unreleased-labiovelarized	b̚ʷˠ
b-unreleased-palatalized	b̚ʲ
b-unreleased-postglottalized	b̚ˀ
b-velarized	bˠ
b/beta	bβ
b/m	bm
b/v	bv
beta	β
beta-approximant	β̞
beta-approximant-breathy voice-nasalized	β̞̤̃
beta-approximant-long	β̞ː
beta-approximant-nasalized	β̞̃
beta-half-voice	β
beta-half-voice-long	βː
beta-labiovelarized	βʷˠ
beta-long	βː
beta-nasalized-palatalized	β̃ʲ
beta-palatalized	βʲ
beta-velarized	βˠ

SPA code	Unicode IPA
c	c
c-aspirated	cʰ
c-aspirated-weak	cʰ
c-breathy voice	c̤
c-click	kǂ
c-ejective	c’
c-fricative	ç
c-fricative-labialized	çʷ
c-fricative-labialized-nasalized	ç̃ʷ
c-fricative-long	çː
c-fricative-palatalized	çʲ
c-fricative-palatoalveolar	ç̟
c-palatalized	cʲ
c-palatoalveolar	c̟
c-palatoalveolar-aspirated	c̟ʰ
c-palatoalveolar-click	k!
c-palatoalveolar-unreleased	c̟̚
c-unreleased	c̚
caret	ʌ
caret-glide	ʌ̯
caret-long	ʌː
caret-long-nasalized	ʌ̃ː
caret-nasalized	ʌ̃
caret-over-short	ʌ̆
caret-voiceless	ʌ̥
d	d
d-aspirated-half-voice	t̬ʰ

SPA code	Unicode IPA
d-breathy voice	d̤
d-breathy voice-long	d̤ː
d-creaky voice	d̰
d-dental	d̪
d-dental-breathy voice	d̪̤
d-dental-breathy voice-long	d̪̤ː
d-dental-lateral-release	d̪ˡ
d-dental-long	d̪ː
d-dental-nasal-release	d̪ⁿ
d-dental-palatalized	d̪ʲ
d-dental-preglottalized	ˀd̪
d-dental-prenasalized	n̪d̪
d-dental-prenasalized-breathy voice	n̪d̪̤
d-dental-unreleased	d̪̚
d-glottalized	dˀ
d-half-voice	d
d-implosive	ɗ
d-interdental	d̪
d-interdental-unreleased	d̪̚
d-labiovelarized	dʷˠ
d-laminal	d̻
d-laminal-lateral-release-palatalized	d̻ˡʲ
d-laminal-long	d̻ː
d-laminal-nasal-release-palatalized	d̻ⁿʲ
d-laminal-palatalized	d̻ʲ
d-lateral-release	dˡ
d-lax	d̙

SPA code	Unicode IPA
d-long	dː
d-long-pharyngealized	dˤː
d-nasal-release	dⁿ
d-palatalized	dʲ
d-pharyngealized	dˤ
d-postglottalized	dˀ
d-preglottalized	ˀd
d-prenasalized	nd
d-prenasalized-palatalized	ndʲ
d-prenasalized/r-trill-retroflex	ɳɖr̠
d-retroflex	ɖ
d-retroflex-breathy voice	ɖ̤
d-retroflex-breathy voice-long	ɖ̤ː
d-retroflex-implosive	ᶑ
d-retroflex-implosive-long	ᶑː
d-retroflex-labiovelarized	ɖʷˠ
d-retroflex-lateral-release	ɖˡ
d-retroflex-long	ɖː
d-retroflex-nasal-release	ɖⁿ
d-retroflex-palatalized	ɖʲ
d-retroflex-preglottalized	ˀɖ
d-retroflex-prenasalized	ɳɖ
d-retroflex-prenasalized-breathy voice	ɳɖ̤
d-retroflex-unreleased	ɖ̚
d-retroflex-unreleased-postglottalized	ɖ̚ˀ
d-syllabic	d̩
d-tense	d͈

SPA code	Unicode IPA
d-tense-long	d͈ː
d-unreleased	d̚
d-unreleased-half-voice	d̚
d-unreleased-postglottalized	d̚ˀ
d-velarized	dˠ
d/b	db
d/eth	d̪ð
d/j-fricative	ɟʝ
d/j-fricative-half-voice	ɟʝ
d/j-fricative-labialized	ɟʝʷ
d/j-fricative-long	ɟʝː
d/j-fricative-prenasalized	ɲɟʝ
d/l	dl
d/n	dn
d/r-trill-retroflex	ɖr̠
d/z	dz
d/z-aspirated-half-voice	t̬sʰ
d/z-creaky voice	dz̰
d/z-hacek	d̠ʒ
d/z-hacek-aspirated-half-voice	t̠̬ʃʰ
d/z-hacek-breathy voice	d̠ʒ̤
d/z-hacek-breathy voice-long	d̠ʒ̤ː
d/z-hacek-creaky voice	d̠ʒ̰
d/z-hacek-half-voice	d̠ʒ
d/z-hacek-labialized	d̠ʒʷ
d/z-hacek-labiovelarized	d̠ʒʷˠ
d/z-hacek-lax	d̠ʒ͉

SPA code	Unicode IPA
d/z-hacek-long	d̠ʒː
d/z-hacek-palatalized	d̠ʒʲ
d/z-hacek-postglottalized	d̠ʒˀ
d/z-hacek-preglottalized	ˀd̠ʒ
d/z-hacek-prenasalized	n̠d̠ʒ
d/z-hacek-prenasalized-breathy voice	n̠d̠ʒ̤
d/z-hacek-retroflex	ɖʐ
d/z-hacek-retroflex-prenasalized	ɳɖʐ
d/z-hacek-tense	d̠͈ʒ
d/z-half-voice	dz
d/z-labiovelarized	dzʷˠ
d/z-laminal	d̻z̻
d/z-lax	d͉z
d/z-long	dzː
d/z-palatalized	dzʲ
d/z-postglottalized	dzˀ
d/z-prenasalized	ndz
d/z-retroflex	ɖʐ
e	e
e-backed	e̠
e-breathy voice-long	e̤ː
e-creaky voice	ḛ
e-dot	ɘ
e-dot-fronted	ɘ̟
e-dot-glide	ɘ̯
e-dot-long	ɘː
e-glide	e̯

SPA code	Unicode IPA
e-glide/iota	e̞̯ɪ
e-half-voice-long	eː
e-long	eː
e-long-advanced	e̟ː
e-long-backed	e̠ː
e-long-nasalized	ẽː
e-long-nasalized-weak	ẽː
e-long/schwa-glide	eːə̯
e-long/yod	eːi
e-mid	e̞
e-mid-backed	e̞̠
e-mid-breathy voice	e̞̤
e-mid-creaky voice	e̞̰
e-mid-creaky voice-long	e̞̰ː
e-mid-creaky voice-nasalized	ẽ̞̰
e-mid-glide	e̞̯
e-mid-half-voice-half-long	e̞ˑ
e-mid-long	e̞ː
e-mid-long-nasalized	ẽ̞ː
e-mid-nasalized	ẽ̞
e-mid-nasalized-weak	ẽ̞
e-mid-over-short	ĕ̞
e-mid-pharyngealized	e̞ˤ
e-mid-retroflexed	e̞˞
e-mid-trema	ɤ̞
e-mid-trema-long	ɤ̞ː
e-mid-trema-long-nasalized	ɤ̞̃ː

SPA code	Unicode IPA
e-mid-trema-nasalized	ɤ̞̃
e-mid-trema-over-short	ɤ̞̆
e-mid-trema-voiceless	ɤ̞̥
e-mid-voiceless	e̞̥
e-nasalized	ẽ
e-nasalized-weak	ẽ
e-over-short	ĕ
e-retracted	e̠
e-retroflexed	e˞
e-trema	ɤ
e-trema-glide	ɤ̯
e-trema-long	ɤː
e-trema-long-nasalized	ɤ̃ː
e-trema-nasalized	ɤ̃
e-trema-retroflexed	ɤ˞
e-trema-voiceless	ɤ̥
e-trema/e	ɤe
e-trema/w	ɤu
e-trema/yod-trema	ɤɰ
e-voiceless	e̥
e/e-mid	ee̞
e/e-mid-long	ee̞ː
e/epsilon-glide	eɛ̯
e/i	ei
e/i-nasalized	eĩ
e/i-retracted	ei̠
e/schwa-glide	eə̯

SPA code	Unicode IPA
e/yod	ei
eng	ŋ
eng-creaky voice	ŋ̰
eng-glottalized	ŋˀ
eng-half-long	ŋˑ
eng-half-voice	ŋ
eng-labialized	ŋʷ
eng-labialized-syllabic	ŋ̩ʷ
eng-long	ŋː
eng-palatalized	ŋʲ
eng-postglottalized	ŋˀ
eng-preglottalized	ˀŋ
eng-prevelar	ŋ̟
eng-prevelar-half-long	ŋ̟ˑ
eng-prevelar-palatalized	ŋ̟ʲ
eng-prevelar-palatalized-syllabic	ŋ̟̩ʲ
eng-prevelar-preglottalized	ˀŋ̟
eng-prevelar-voiceless	ŋ̟̥
eng-prevelar-voiceless-half-long	ŋ̟̥ˑ
eng-syllabic	ŋ̩
eng-uvular	ɴ
eng-voiceless	ŋ̥
eng-voiceless-half-long	ŋ̥ˑ
eng-voiceless-palatalized	ŋ̥ʲ
eng/m	ŋm
eng/m-syllabic	ŋm̩
epsilon	ɛ

SPA code	Unicode IPA
epsilon-backed	ɛ̠
epsilon-creaky voice	ɛ̰
epsilon-dot	ɜ
epsilon-dot-backed	ɜ̠
epsilon-dot-fronted	ɜ̟
epsilon-dot-glide	ɜ̯
epsilon-dot-nasalized	ɜ̃
epsilon-dot-over-short	ɜ̆
epsilon-dot-retroflexed	ɜ˞
epsilon-dot/e-glide	ɜe̯
epsilon-dot/iota-glide	ɜɪ̯
epsilon-dot/o-glide	ɜo̯
epsilon-glide	ɛ̯
epsilon-half-long	ɛˑ
epsilon-half-long-nasalized	ɛ̃ˑ
epsilon-half-voice-long	ɛː
epsilon-long	ɛː
epsilon-long-advanced	ɛ̟ː
epsilon-long-nasalized	ɛ̃ː
epsilon-long-nasalized-weak	ɛ̃ː
epsilon-nasalized	ɛ̃
epsilon-nasalized-weak	ɛ̃
epsilon-over-short	ɛ̆
epsilon-over-short-nasalized	ɛ̆̃
epsilon-retroflexed	ɛ˞
epsilon-voiceless	ɛ̥
epsilon-voiceless-long	ɛ̥ː

SPA code	Unicode IPA
epsilon/a	ɛa
epsilon/caret-glide	ɛʌ̯
epsilon/epsilon-dot-glide	ɛɜ̯
epsilon/schwa	ɛə
epsilon/yod	ɛi
eth	ð
eth-approximant	ð̞
eth-half-long	ðˑ
eth-half-voice	ð
eth-half-voice-lax	ð̙
eth-lax	ð̙
eth-palatalized	ðʲ
eth-pharyngealized	ðˤ
f	f
f-ejective	fʼ
f-half-long	fˑ
f-labialized	fʷ
f-labiovelarized	fʷˠ
f-lax	f̙
f-long	fː
f-long-labialized	fʷː
f-long-labialized-pharyngealized	fʷˤː
f-long-pharyngealized	fˤː
f-nasalized	f̃
f-palatalized	fʲ
f-pharyngealized	fˤ
f-syllabic	f̩

SPA code	Unicode IPA
f-tense-long	f͈ː
f-velarized	fˠ
falling	˥˩
g	ɡ
g-aspirated-half-voice	k̬ʰ
g-breathy voice	ɡ̤
g-breathy voice-labialized	ɡ̤ʷ
g-breathy voice-long	ɡ̤ː
g-creaky voice	ɡ̰
g-half-voice	ɡ
g-implosive	ɠ
g-labialized	ɡʷ
g-labialized-syllabic	ɡ̩ʷ
g-labiovelarized	ɡʷˠ
g-lax	ɡ̞
g-long	ɡː
g-long-pharyngealized	ɡˤː
g-nasal-release	ɡⁿ
g-palatalized	ɡʲ
g-pharyngealized	ɡˤ
g-postglottalized	ɡˀ
g-preglottalized	ˀɡ
g-prenasalized	ŋɡ
g-prenasalized-breathy voice	ŋɡ̤
g-prenasalized-labialized	ŋɡʷ
g-prenasalized-palatalized	ŋɡʲ
g-prevelar	ɡ̟

SPA code	Unicode IPA
g-prevelar-palatalized	g̟ʲ
g-prevelar-prenasalized	ŋg̟
g-prevelar-tense	g̟͈
g-prevelar-unreleased	g̟˺
g-syllabic	g̩
g-tense-long	g͈ː
g-tense-long-labialized	g͈ʷː
g-unreleased	g˺
g-unreleased-half-voice	g˺
g/b	gb
g/b-prenasalized	ŋmgb
g/b-syllabic	gb̩
g/eng	gŋ
g/gamma	gɣ
gamma	ɣ
gamma-half-long	ɣˑ
gamma-half-voice	ɣ
gamma-labialized	ɣʷ
gamma-labialized-nasalized	ɣ̃ʷ
gamma-nasalized	ɣ̃
gamma-palatalized	ɣʲ
gamma-prevelar	ɣ̟
gamma-prevelar-palatalized	ɣ̟ʲ
gamma-syllabic	ɣ̩
gamma-tense-long	ɣ͈ː
gamma-tense-long-labialized	ɣ͈ʷː
gamma-uvular	ʁ

SPA code	Unicode IPA
gamma-uvular-creaky voice	ʁ̰
gamma-uvular-half-voice	ʁ
gamma-uvular-labialized	ʁʷ
gamma-uvular-long	ʁː
gamma-uvular-long-pharyngealized	ʁˤː
gamma-uvular-palatalized	ʁʲ
gamma-uvular-pharyngealized	ʁˤ
glottal stop	ʔ
glottal stop-aspirated	ʔʰ
glottal stop-labialized	ʔʷ
glottal stop-long	ʔː
glottal stop-palatalized	ʔʲ
glottal stop-pharyngealized	ʔˤ
glottal stop-unreleased-labialized	ʔ̚ʷ
h	h
h-half-voice	ɦ
h-labialized	hʷ
h-labialized-nasalized	h̃ʷ
h-lax	h̞
h-long	hː
h-nasalized	h̃
h-nasalized-palatalized	h̃ʲ
h-palatalized	hʲ
h-voice	ɦ
h-voice-labiovelarized	ɦʷˠ
h-voice-nasalized	ɦ̃
h-voice-palatalized	ɦʲ

SPA code	Unicode IPA
h-voice-velarized	ɦˠ
high	˥
high-creaky voice	˥̰
high-falling	˥˦
high-falling-creaky voice	˥˩̰
high-falling-glottalized	˥˩ˀ
high-falling-rising	˥˧˥
high-over-short	˥̆
high-rising	˦˥
high-rising-creaky voice	˦˥̰
high-rising-over-short	˦˥̆
higher-high	˥
higher-mid	˦
higher-mid-falling-low	˦˩
higher-mid-falling-mid	˦˧
higher-mid-rising	˧˦
i	i
i-backed	i̠
i-bar	ɨ
i-bar-backed	ɨ̠
i-bar-creaky voice	ɨ̰
i-bar-fronted	ɨ̟
i-bar-half-voice	ɨ
i-bar-half-voice-long	ɨː
i-bar-long	ɨː
i-bar-long-nasalized	ɨ̃ː
i-bar-long-nasalized-weak	ɨ̃ː

SPA code	Unicode IPA
i-bar-long-retroflexed	ɨ˞ː
i-bar-nasalized	ɨ̃
i-bar-nasalized-weak	ɨ̃
i-bar-over-short	ɨ̆
i-bar-retroflexed	ɨ˞
i-bar-voiceless	ɨ̥
i-bar-voiceless-retroflexed	ɨ̥˞
i-breathy voice-long	i̤ː
i-creaky voice	ḭ
i-creaky voice-long	ḭː
i-creaky voice-nasalized	ḭ̃
i-half-long	iˑ
i-half-voice	i
i-half-voice-long	iː
i-lax	ɪ
i-long	iː
i-long-backed	i̠ː
i-long-backed-retracted	i̠̞ː
i-long-nasalized	ĩː
i-long-nasalized-weak	ĩː
i-long-retracted	i̞ː
i-nasalized	ĩ
i-nasalized-weak	ĩ
i-over-short	ĭ
i-over-short-nasalized	ĭ̃
i-retroflexed	i˞
i-trema	ɯ

SPA code	Unicode IPA
i-trema-creaky voice	ɯ̰
i-trema-long	ɯː
i-trema-long-nasalized	ɯ̃ː
i-trema-nasalized	ɯ̃
i-trema-over-short	ɯ̆
i-trema-voiceless	ɯ̥
i-trema-voiceless-nasalized	ɯ̥̃
i-voiceless	i̥
i-voiceless-long	i̥ː
i-voiceless-over-short	ĭ̥
i/a-glide	ia̯
i/schwa-glide	iə̯
i/yod	ii̯
iota	ɪ
iota-backed	ɪ̠
iota-backed-retracted	ɪ̠̞
iota-bar	ɪ̈
iota-bar-long	ɪ̈ː
iota-bar-nasalized	ɪ̈̃
iota-bar-over-short	ɪ̈̆
iota-breathy voice	ɪ̤
iota-creaky voice	ɪ̰
iota-creaky voice-long	ɪ̰ː
iota-glide	ɪ̯
iota-glide-voiceless	ɪ̯̥
iota-long	ɪː
iota-long-backed	ɪ̠ː

SPA code	Unicode IPA
iota-long-nasalized	ɪ̃ː
iota-nasalized	ɪ̃
iota-nasalized-weak	ɪ̃
iota-over-short	ɪ̆
iota-retracted	ɪ̙
iota-retroflexed	ɪ˞
iota-trema	ɯ̞
iota-trema-glide	ɯ̞̯
iota-trema-long-nasalized	ɯ̞̃ː
iota-trema-nasalized	ɯ̞̃
iota-trema-voiceless-over-short	ɯ̞̆̊
iota-trema/yod-trema	ɯ̞ɰ
iota-voiceless	ɪ̥
iota-voiceless-over-short	ɪ̥̆
iota/i	ɪi
iota/iota-glide-backed	ɪɪ̠̯
iota/schwa	ɪə
iota/schwa-glide	ɪə̯
iota/yod	ɪi
j	ɟ
j-aspirated-half-voice	ç̬^{h}
j-creaky voice	ɟ̰
j-fricative	ʝ
j-fricative-half-voice	ʝ
j-fricative-labialized	ʝ^{w}
j-fricative-nasalized	ʝ̃
j-fricative-palatoalveolar	ʝ̟

SPA code	Unicode IPA
j-implosive	ʄ
j-long	ɟː
j-palatalized	ɟʲ
j-palatoalveolar	ɟ̟
j-palatoalveolar-prenasalized	ɲ̟ɟ̟̊
j-palatoalveolar-unreleased	ɟ̟˺
j-prenasalized	ɲɟ
j-unreleased-half-voice	ɟ˺
j-unreleased-postglottalized	ɟ˺ˀ
j/n-palatal	ɟɲ
k	k
k-aspirated	kʰ
k-aspirated-labialized	kʷʰ
k-aspirated-labiovelarized	kʷˠʰ
k-aspirated-long	kʰː
k-aspirated-long-labialized	kʷʰː
k-aspirated-palatalized	kʲʰ
k-aspirated-weak	kʰ
k-aspirated-weak-labialized	kʷʰ
k-breathy voice	ɡ̤̊
k-ejective	kʼ
k-ejective-labialized	kʷʼ
k-ejective-long	kʼː
k-ejective-long-labialized	kʷʼː
k-ejective-palatalized	kʲʼ
k-glottalized	kˀ
k-half-long	kˑ

SPA code	Unicode IPA
k-labialized	kʷ
k-labialized-pharyngealized	kʷˤ
k-labiovelarized	kʷˠ
k-lax	k̞
k-lax-preglottalized	ˀk̞
k-long	kː
k-long-labialized	kʷː
k-long-pharyngealized	kˤː
k-nasal-release	kⁿ
k-palatalized	kʲ
k-pharyngealized	kˤ
k-preaspirated	ʰk
k-preaspirated-half-long	ʰkˑ
k-preaspirated-labialized	ʰkʷ
k-preaspirated-long	ʰkː
k-preglottalized	ˀk
k-prenasalized	ŋk
k-prenasalized-aspirated	ŋkʰ
k-prenasalized-labialized	ŋkʷ
k-prenasalized-palatalized	ŋkʲ
k-prevelar	k̟
k-prevelar-aspirated	k̟ʰ
k-prevelar-aspirated-palatalized	k̟ʲʰ
k-prevelar-aspirated-weak	k̟ʰ
k-prevelar-aspirated-weak-palatalized	k̟ʲʰ
k-prevelar-ejective-palatalized	k̟ʲ’
k-prevelar-lax	k̟̞

SPA code	Unicode IPA
k-prevelar-long	k̟ː
k-prevelar-long-palatalized	k̟ʲː
k-prevelar-palatalized	k̟ʲ
k-prevelar-preaspirated-long	ʰk̟ː
k-prevelar-unreleased	k̟˺
k-tense	k͈
k-tense-labialized	k͈ʷ
k-tense-long	k͈ː
k-tense-long-labialized	k͈ʷː
k-tense-long-palatalized	k͈ʲː
k-unreleased	k˺
k-unreleased-labialized	k˺ʷ
k-unreleased-tense	k͈˺
k/c-aspirated	kcʰ
k/c-fricative	cç
k/gamma	kɣ
k/gamma-labialized	kɣʷ
k/j-fricative	cʝ
k/p	kp
k/p-unreleased	kp˺
k/x	kx
k/x-aspirated	kxʰ
k/x-ejective	kx’
k/x-labialized	kxʷ
k/x-lateral-ejective	kʟ̥͓’
k/x-prevelar-palatalized	k̟x̟ʲ
l	l

SPA code	Unicode IPA
l-breathy voice	l̤
l-creaky voice	l̰
l-dental	l̪
l-dental-half-voice-velarized	l̪ˠ
l-dental-long	l̪ː
l-dental-palatalized	l̪ʲ
l-dental-syllabic	l̪̩
l-dental-velarized	l̪ˠ
l-flap	ɺ
l-flap-long	ɺː
l-flap-nasalized	ɺ̃
l-flap-palatalized	ɺʲ
l-flap-retroflex	ɺ̠
l-flap-voiceless	ɺ̥
l-flap-voiceless-palatalized	ɺ̥ʲ
l-fricative	ɬ
l-fricative-ejective	ɬʼ
l-fricative-ejective-palatalized	ɬʲʼ
l-fricative-laminal	ɬ̻
l-fricative-long	ɬː
l-fricative-palatalized	ɬʲ
l-fricative-syllabic	ɬ̩
l-fricative-voice	ɮ
l-fricative-voice-palatalized	ɮʲ
l-half-long	lˑ
l-half-voice	l
l-half-voice-long	lː

SPA code	Unicode IPA
l-half-voice-palatalized	lʲ
l-half-voice-velarized	lˠ
l-interdental	l̪
l-labialized	lʷ
l-labiovelarized	lʷˠ
l-labiovelarized-syllabic	l̩ʷˠ
l-laminal	l̻
l-laminal-creaky voice	l̻̰
l-laminal-long	l̻ː
l-laminal-palatalized	l̻ʲ
l-laminal-preglottalized-voiceless	ˀl̻̥
l-laminal-voiceless	l̻̥
l-long	lː
l-long-palatalized	lʲː
l-long-pharyngealized	lˤː
l-nasalized	l̃
l-palatal	ʎ
l-palatal-half-voice	ʎ
l-palatal-voiceless	l̥ʲ
l-palatalized	lʲ
l-palatalized-syllabic	l̩ʲ
l-palatoalveolar	ʎ̟
l-pharyngealized	lˤ
l-pharyngealized-syllabic	l̩ˤ
l-preglottalized	ˀl
l-retroflex	ɭ
l-retroflex-long	ɭː

SPA code	Unicode IPA
l-retroflex-palatalized	ɭʲ
l-retroflex-syllabic	ɭ̩
l-retroflex-voiceless	ɭ̊
l-syllabic	l̩
l-tense-long	l͈ː
l-velarized	lˠ
l-velarized-syllabic	l̩ˠ
l-voiceless	l̥
l-voiceless-half-long	l̥ˑ
l-voiceless-palatalized	l̥ʲ
l-voiceless-velarized	l̥ˠ
low	˩
low-breathy voice-over-short	˩̤̆
low-creaky voice	˩̰
low-falling	˨˩
low-falling-breathy voice	˨˩̤
low-falling-rising	˨˩˥
low-glottalized	˩ˀ
low-rising	˩˨
low-rising-falling	˩˨˩
low-rising-long	˩˧
lower-low	˩
lower-mid	˨
lower-mid-falling	˨˩
lower-mid-falling-breathy voice	˨˩̤
lower-mid-falling-pharyngealized	˨˩ˤ
lower-mid-falling-rising	˧˩˥

SPA code	Unicode IPA
lower-mid-rising	˨˧
lower-mid-rising-falling	˨˧˨
lower-mid-rising-over-short	˨˧̆
m	m
m-breathy voice	m̤
m-creaky voice	m̰
m-glottalized	mˀ
m-half-long	mˑ
m-half-voice	m
m-half-voice-labiovelarized	mʷˠ
m-half-voice-long	mː
m-half-voice-palatalized	mʲ
m-labialized	mʷ
m-labiodental	ɱ
m-labiodental-syllabic	ɱ̩
m-labiovelarized	mʷˠ
m-lax	m̞
m-long	mː
m-long-labialized	mʷː
m-long-labialized-pharyngealized	mʷˤː
m-long-palatalized	mʲː
m-long-pharyngealized	mˤː
m-palatalized	mʲ
m-palatalized-syllabic	m̩ʲ
m-pharyngealized	mˤ
m-postglottalized	mˀ
m-preglottalized	ˀm

SPA code	Unicode IPA
m-preglottalized-voiceless	ˀm̥
m-syllabic	m̩
m-syllabic/v	m̩v
m-tense	m͈
m-tense-long	m͈ː
m-velarized	mˠ
m-voiceless	m̥
m-voiceless-half-long	m̥ˑ
m-voiceless-labialized	m̥ʷ
m/v	mv
mid	˧
mid-falling	˧˩
mid-falling-creaky voice	˧˩̰
mid-falling-creaky voice/glottal stop	˧˩̰ʔ
mid-falling-lower-mid	˧˨
mid-falling-over-short	˧˩̆
mid-over-short	˧̆
mid-rising	˧˥
n	n
n-breathy voice	n̤
n-creaky voice	n̰
n-dental	n̪
n-dental-breathy voice	n̪̤
n-dental-long	n̪ː
n-dental-syllabic	n̪̩
n-glottalized	nˀ
n-half-long	nˑ

SPA code	Unicode IPA
n-half-voice	n
n-half-voice-long	nː
n-half-voice-palatalized	nʲ
n-half-voice-velarized	nˠ
n-interdental	n̪
n-interdental-half-voice	n̪
n-labialized	nʷ
n-labiovelarized	nʷˠ
n-laminal	n̻
n-laminal-long	n̻ː
n-laminal-palatalized	n̻ʲ
n-laminal-syllabic	n̻̩
n-laminal-voiceless	n̻̥
n-laminal-voiceless-palatalized	n̻̥ʲ
n-lax	n̙
n-long	nː
n-long-palatalized	nʲː
n-long-pharyngealized	nˤː
n-palatal	ɲ
n-palatal-half-voice	ɲ
n-palatal-long	ɲː
n-palatal-palatalized	ɲʲ
n-palatal-preglottalized	ˀɲ
n-palatal-syllabic	ɲ̩
n-palatal-voiceless	ɲ̥
n-palatalized	nʲ
n-palatalized-syllabic	n̩ʲ

SPA code	Unicode IPA
n-palatoalveolar	ɲ̟
n-palatoalveolar-voiceless	ɲ̟̥
n-pharyngealized	nˤ
n-postglottalized	nˀ
n-preglottalized	ˀn
n-retroflex	ɳ
n-retroflex-palatalized	ɳʲ
n-retroflex-syllabic	ɳ̩
n-retroflex-voiceless	ɳ̥
n-syllabic	n̩
n-tense	n͈
n-tense-long	n͈ː
n-unreleased	n̚
n-unreleased-palatalized	n̚ʲ
n-uvular	ɴ
n-uvular-long	ɴː
n-velarized	nˠ
n-voiceless	n̥
n-voiceless-half-long	n̥ˑ
n-voiceless-long	n̥ː
n-voiceless-palatalized	n̥ʲ
n-voiceless-tense	n̥͈
n-voiceless-velarized	n̥ˠ
n/m	nm
o	o
o–open-dot-backed	ɞ̠
o-breathy voice	o̤

SPA code	Unicode IPA
o-breathy voice-long	o̤ː
o-creaky voice	o̰
o-dot	ɵ
o-dot/w	ɵu
o-fronted	o̟
o-glide	o̯
o-glide-preglottalized	ˀo̯
o-glide/u	o̯u
o-half-voice-long	oː
o-long	oː
o-long-advanced	o̘ː
o-long-fronted	o̟ː
o-long-nasalized	õː
o-long-nasalized-weak	õː
o-long/w	oːu
o-mid	o̞
o-mid-creaky voice	o̞̰
o-mid-creaky voice-long	o̞̰ː
o-mid-creaky voice-nasalized	õ̞̰
o-mid-dot	ɵ̞
o-mid-dot-backed	ɵ̞̠
o-mid-dot-glide	ɵ̞̯
o-mid-dot-half-voice-long	ɵ̞ː
o-mid-dot-long	ɵ̞ː
o-mid-dot-long-nasalized	ɵ̞̃ː
o-mid-dot-nasalized	ɵ̞̃
o-mid-dot-over-short	ɵ̞̆

SPA code	Unicode IPA
o-mid-fronted	o̞̟
o-mid-glide	o̞̯
o-mid-half-voice-half-long	o̞ˑ
o-mid-long	o̞ː
o-mid-long-nasalized	õ̞ː
o-mid-nasalized	õ̞
o-mid-nasalized-weak	õ̞
o-mid-over-short	ŏ̞
o-mid-retroflexed	o̞˞
o-mid-trema	ø̞
o-mid-trema-long	ø̞ː
o-mid-trema-pharyngealized	ø̞ˤ
o-mid-trema/schwa-glide	ø̞ə̞̯
o-mid-voiceless	o̞̥
o-mid/o-open-glide	o̞ɔ̯
o-mid/schwa-glide	o̞ə̞̯
o-mid/w	o̞u
o-mid/yod	o̞i
o-nasalized	õ
o-nasalized-weak	õ
o-open	ɔ
o-open-breathy voice	ɔ̤
o-open-creaky voice	ɔ̰
o-open-dot	ɞ
o-open-glide	ɔ̯
o-open-half-long	ɔˑ
o-open-half-voice	ɔ

SPA code	Unicode IPA
o-open-half-voice-long	ɔː
o-open-long	ɔː
o-open-long-advanced	ɔ̟ː
o-open-long-nasalized	ɔ̃ː
o-open-long-uvularized	ɔ̴ː
o-open-nasalized	ɔ̃
o-open-nasalized-weak	ɔ̃
o-open-over-short	ɔ̆
o-open-retroflexed	ɔ˞
o-open-trema	œ
o-open-trema-long	œː
o-open-trema-long-nasalized	œ̃ː
o-open-trema-nasalized	œ̃
o-open-uvularized	ɔ̴
o-open-voiceless	ɔ̥
o-open-voiceless-long	ɔ̥ː
o-open/caret-glide	ɔʌ̯
o-open/o-glide	ɔo̯
o-open/o-glide-breathy voice	ɔo̯̤
o-open/o-glide-nasalized	ɔõ̯
o-open/schwa	ɔə
o-over-short	ŏ
o-over-short-nasalized	ŏ̃
o-trema	ø
o-trema-long	øː
o-trema-nasalized	ø̃
o-trema-over-short	ø̆

SPA code	Unicode IPA
o-voiceless	o̥
o/e-trema	oɤ
o/e-trema-retroflexed	oɤ˞
o/o-mid	oo̞
o/o-mid-long	oo̞ː
o/u	ou
o/u-nasalized	oũ
o/u-retracted	ou̙
o/w	ou
o/yod-over-short	oĭ
omega	ʊ
omega-long	ʊː
omega-trema-long	ʏː
p	p
p-aspirated	p$^{\text{h}}$
p-aspirated-labialized	p$^{\text{wh}}$
p-aspirated-labiovelarized	p$^{\text{wɣh}}$
p-aspirated-long	p$^{\text{h}}$ː
p-aspirated-palatalized	p$^{\text{jh}}$
p-aspirated-weak	p$^{\text{h}}$
p-breathy voice	b̤̥
p-ejective	p’
p-ejective-long	p’ː
p-glottalized	p$^{\text{ʔ}}$
p-half-long	pˑ
p-implosive	ɓ̥
p-labialized	p$^{\text{w}}$

SPA code	Unicode IPA
p-labiodental	p̪
p-labiovelarized	pʷˠ
p-lateral-release	pˡ
p-lax	p͉
p-lax-long	p͉ː
p-lax-palatalized	p͉ʲ
p-long	pː
p-long-palatalized	pʲː
p-nasal-release	pⁿ
p-palatalized	pʲ
p-preaspirated	ʰp
p-preaspirated-half-long	ʰpˑ
p-preaspirated-long	ʰpː
p-preglottalized	ˀp
p-prenasalized	mp
p-prenasalized-aspirated	mpʰ
p-tense	p͈
p-tense-long	p͈ː
p-unreleased	p̚
p-unreleased-glottalized	p̚ˀ
p-unreleased-labiovelarized	p̚ʷˠ
p-unreleased-palatalized	p̚ʲ
p-velarized	pˠ
p/f	pf
p/f-aspirated	pfʰ
p/f-ejective	pfʼ
p/phi	pɸ

SPA code	Unicode IPA
pharyngeal-voice	ʕ
pharyngeal-voice-long	ʕː
pharyngeal-voiceless	ħ
pharyngeal-voiceless-labialized	ħʷ
pharyngeal-voiceless-long	ħː
pharyngeal-voiceless-tense-long	ħ͈ː
phi	ɸ
phi-ejective	ɸ’
phi-labialized	ɸʷ
phi-labiovelarized	ɸʷˠ
phi-labiovelarized-nasalized	ɸ̃ʷˠ
phi-long	ɸː
phi-nasalized	ɸ̃
phi-nasalized-palatalized	ɸ̃ʲ
phi-palatalized	ɸʲ
q	q
q-aspirated	qʰ
q-aspirated-labialized	qʷʰ
q-aspirated-palatalized	qʲʰ
q-aspirated-weak	qʰ
q-creaky voice	q̰
q-ejective	q’
q-ejective-labialized	qʷ’
q-labialized	qʷ
q-long	qː
q-long-pharyngealized	qˤː
q-palatalized	qʲ

SPA code	Unicode IPA
q-pharyngealized	qˤ
q-preaspirated	ʰq
q-tense	q͈
q-tense-labialized	q͈ʷ
q-unreleased	q̚
q-voice	ɢ
q-voice-labialized	ɢʷ
q-voice-long	ɢː
q-voice-palatalized	ɢʲ
q-voice/gamma-uvular	ɢʁ
q/x-uvular	qχ
q/x-uvular-aspirated	qχʰ
q/x-uvular-aspirated-long	qχʰː
q/x-uvular-ejective	qχʼ
q/x-uvular-ejective-labialized	qχʷʼ
q/x-uvular-labialized	qχʷ
r	r
r-approximant	ɹ
r-approximant-retroflex	ɻ
r-approximant-retroflex-syllabic	ɻ̩
r-approximant-retroflex-voiceless	ɻ̥
r-approximant-retroflex-voiceless-syllabic	ɻ̥̩
r-approximant-uvular	ʁ̞
r-approximant-uvular-voiceless	ʁ̞̥
r-approximant-voiceless	ɹ̥
r-flap	ɾ
r-flap-breathy voice	ɾ̤

SPA code	Unicode IPA
r-flap-dental-velarized	ɾ̪ˠ
r-flap-fricative	ɾ̽
r-flap-glottalized	ɾˀ
r-flap-half-voice-long	ɾː
r-flap-half-voice-palatalized	ɾʲ
r-flap-half-voice-velarized	ɾˠ
r-flap-long	ɾː
r-flap-nasalized	ɾ̃
r-flap-nasalized-palatalized	ɾ̃ʲ
r-flap-nasalized-velarized	ɾ̃ˠ
r-flap-palatalized	ɾʲ
r-flap-pharyngealized	ɾˤ
r-flap-retroflex	ɽ
r-flap-retroflex-breathy voice	ɽ̤
r-flap-retroflex-nasalized	ɽ̃
r-flap-retroflex-palatalized	ɽʲ
r-flap-retroflex-voiceless	ɽ̥
r-flap-velarized	ɾˠ
r-flap-voiceless	ɾ̥
r-flap-voiceless-palatalized	ɾ̥ʲ
r-flap-voiceless-velarized	ɾ̥ˠ
r-flap/l	ɾl
r-flap/n	ɾn
r-fricative	z͇
r-fricative-retroflex	z̢͇
r-fricative-retroflex-half-voice	z̢͇
r-fricative-retroflex-voiceless	s̢͇

SPA code	Unicode IPA
r-fricative-voiceless	z̥̳
r-long	rː
r-syllabic	ɹ̩
r-trill	r
r-trill-half-long	rˑ
r-trill-half-voice	r
r-trill-half-voice-long	rː
r-trill-labiovelarized	rʷˠ
r-trill-long	rː
r-trill-long-pharyngealized	rˤː
r-trill-palatalized	rʲ
r-trill-pharyngealized	rˤ
r-trill-preglottalized	ˀr
r-trill-retroflex	r̠
r-trill-retroflex-nasalized	r̠̃
r-trill-syllabic	r̩
r-trill-tense-long	r͈ː
r-trill-uvular	ʀ
r-trill-uvular-voiceless	ʀ̥
r-trill-velarized	rˠ
r-trill-voiceless	r̥
r-trill-voiceless-half-long	r̥ˑ
r-trill-voiceless-palatalized	r̥ʲ
rising	˩˥
s	s
s-approximant-syllabic	s̞̩
s-aspirated	sʰ

SPA code	Unicode IPA
s-dental	s̪
s-dental-lax	s̪̞
s-dental-long	s̪ː
s-ejective	s’
s-ejective-long	s’ː
s-glottalized	sˀ
s-hacek	ʃ
s-hacek-ejective	ʃʼ
s-hacek-half-long	ʃˑ
s-hacek-labialized	ʃʷ
s-hacek-labiovelarized	ʃʷˠ
s-hacek-lax	ʃ̞
s-hacek-long	ʃː
s-hacek-long-pharyngealized	ʃˤː
s-hacek-nasalized	ʃ̃
s-hacek-palatalized	ʃʲ
s-hacek-pharyngealized	ʃˤ
s-hacek-retroflex	ʂ
s-hacek-retroflex/r	ʂɻ
s-hacek-syllabic	ʃ̩
s-hacek-tense	ʃ͈
s-hacek-tense-labialized	ʃ͈ʷ
s-hacek-tense-long	ʃ͈ː
s-hacek-tense-long-palatalized	ʃ͈ʲː
s-hacek-velarized	ʃˠ
s-half-long	sˑ
s-labialized	sʷ

SPA code	Unicode IPA
s-labiovelarized	sʷˠ
s-laminal	s̻
s-laminal-half-long	s̻ˑ
s-laminal-lax	s̻͉
s-laminal-long	s̻ː
s-laminal-palatalized	s̻ʲ
s-laminal/theta	s̻θ
s-lax	s͉
s-long	sː
s-long-palatalized	sʲː
s-long-pharyngealized	sˤː
s-nasalized	s̃
s-palatalized	sʲ
s-pharyngealized	sˤ
s-retroflex	s̢
s-retroflex-long	s̢ː
s-syllabic	s̩
s-tense	s͈
s-tense-long	s͈ː
s-velarized	sˠ
s/l-fricative	sɬ
s/t	st
s/t/s	sts
s/t/s-hacek	st̠ʃ
s/t/s-long	stsː
schwa	ə
schwa-backed	ə̠

SPA code	Unicode IPA
schwa-creaky voice	ə̰
schwa-fronted	ə̟
schwa-glide	ə̯
schwa-glide/i-long	ə̯iː
schwa-glide/iota	ə̯ɪ
schwa-long	əː
schwa-long-advanced	ə̞ː
schwa-long-nasalized	ə̃ː
schwa-long-nasalized-weak	ə̃ː
schwa-long-uvularized	əː
schwa-nasalized	ə̃
schwa-nasalized-retroflexed	ə̃˞
schwa-nasalized-weak	ə̃
schwa-over-short	ə̆
schwa-over-short-fronted	ə̟̆
schwa-retroflexed	ə˞
schwa-uvularized	ə
schwa-voiceless	ə̥
schwa-voiceless-nasalized	ə̥̃
schwa-voiceless-over-short	ə̥̆
schwa/i-bar-retracted	əɨ̞
t	t
t-aspirated	t^{h}
t-aspirated-labialized	t^{wh}
t-aspirated-labiovelarized	$\text{t}^{\text{wɣh}}$
t-aspirated-long	t^{h}ː
t-aspirated-palatalized	t^{jh}

SPA code	Unicode IPA
t-aspirated-weak	tʰ
t-breathy voice	d̤̥
t-dental	t̪
t-dental-aspirated	t̪ʰ
t-dental-aspirated-long	t̪ʰː
t-dental-aspirated-palatalized	t̪ʲʰ
t-dental-aspirated-weak	t̪ʰ
t-dental-breathy voice	d̪̤̥
t-dental-ejective	t̪ʼ
t-dental-lateral-release	t̪ˡ
t-dental-long	t̪ː
t-dental-nasal-release	t̪ⁿ
t-dental-palatalized	t̪ʲ
t-dental-prenasalized	n̪t̪
t-dental-unreleased	t̪̚
t-ejective	tʼ
t-ejective-long	tʼː
t-glottalized	tˀ
t-half-long	tˑ
t-implosive	ɗ̥
t-interdental	t̪
t-interdental-aspirated	t̪ʰ
t-interdental-unreleased	t̪̚
t-labialized	tʷ
t-labiovelarized	tʷˠ
t-laminal	t̻
t-laminal-aspirated	t̻ʰ

SPA code	Unicode IPA
t-laminal-aspirated-long	t̻ʰː
t-laminal-aspirated-weak	t̻ʰ
t-laminal-click	k!
t-laminal-ejective	t̻ʼ
t-laminal-lateral-release-palatalized	t̻ˡʲ
t-laminal-long	t̻ː
t-laminal-nasal-release-palatalized	t̻ⁿʲ
t-laminal-palatalized	t̻ʲ
t-laminal-unreleased	t̻̚
t-lateral-release	tˡ
t-lax	t̞
t-lax-palatalized	t̞ʲ
t-long	tː
t-long-palatalized	tʲː
t-long-pharyngealized	tˤː
t-long/s	tːs
t-long/s-hacek	t̠ːʃ
t-nasal-release	tⁿ
t-palatalized	tʲ
t-pharyngealized	tˤ
t-preaspirated	ʰt
t-preaspirated-half-long	ʰtˑ
t-preaspirated-long	ʰtː
t-preglottalized	ˀt
t-prenasalized	nt
t-prenasalized-aspirated	ntʰ
t-prenasalized-aspirated-palatalized	ntʲʰ

SPA code	Unicode IPA
t-prenasalized/r-trill-retroflex-voiceless	ɳʈr̠̥
t-retroflex	ʈ
t-retroflex-aspirated	ʈʰ
t-retroflex-aspirated-labiovelarized	ʈʷˠʰ
t-retroflex-aspirated-long	ʈʰː
t-retroflex-aspirated-palatalized	ʈʲʰ
t-retroflex-ejective	ʈʼ
t-retroflex-labiovelarized	ʈʷˠ
t-retroflex-lateral-release	ʈˡ
t-retroflex-lax	ʈ̞
t-retroflex-long	ʈː
t-retroflex-nasal-release	ʈⁿ
t-retroflex-palatalized	ʈʲ
t-retroflex-unreleased	ʈ̚
t-retroflex-unreleased-glottalized	ʈ̚ˀ
t-retroflex/r-flap-retroflex	ʈɽ
t-tense	t͈
t-tense-long	t͈ː
t-unreleased	t̚
t-unreleased-glottalized	t̚ˀ
t-unreleased-palatalized	t̚ʲ
t-unreleased-pharyngealized	t̚ˤ
t-unreleased-tense	t͈̚
t-velarized	tˠ
t/c-fricative	cç
t/c-fricative-aspirated	cçʰ
t/c-fricative-aspirated-labialized	cçʷʰ

SPA code	Unicode IPA
t/c-fricative-aspirated-weak	cçʰ
t/c-fricative-long	cçː
t/l-fricative	tɬ
t/l-fricative-aspirated	tɬʰ
t/l-fricative-click	k‖
t/l-fricative-ejective	tɬ’
t/l-fricative-ejective-palatalized	tɬʲ’
t/l-fricative-ejective-syllabic	tɬ̩’
t/l-fricative-voice	tɮ
t/p	tp
t/r-fricative-retroflex-voiceless	ʈs̳
t/r-trill-retroflex-voiceless	ʈr̥̠
t/s	ts
t/s-aspirated	tsʰ
t/s-aspirated-labialized	tsʷʰ
t/s-aspirated-labiovelarized	tsʷˠʰ
t/s-aspirated-long	tsʰː
t/s-aspirated-palatalized	tsʲʰ
t/s-aspirated-weak	tsʰ
t/s-breathy voice	tz̤̥
t/s-click	k|
t/s-dental	t̪s̪
t/s-ejective	ts’
t/s-ejective-labialized	tsʷ’
t/s-ejective-long	ts’ː
t/s-fricative-ejective	ts’
t/s-hacek	t̠ʃ

SPA code	Unicode IPA
t/s-hacek-aspirated	t̠ʃʰ
t/s-hacek-aspirated-labialized	t̠ʃʷʰ
t/s-hacek-aspirated-labiovelarized	t̠ʃʷˠʰ
t/s-hacek-aspirated-long	t̠ʃʰː
t/s-hacek-aspirated-palatalized	t̠ʃʲʰ
t/s-hacek-aspirated-weak	t̠ʃʰ
t/s-hacek-ejective	t̠ʃʼ
t/s-hacek-ejective-labialized	t̠ʃʷʼ
t/s-hacek-ejective-labialized-syllabic	t̠ʃ̩ʷʼ
t/s-hacek-ejective-long	t̠ʃʼː
t/s-hacek-ejective-palatalized	t̠ʃʲʼ
t/s-hacek-ejective-syllabic	t̠ʃ̩ʼ
t/s-hacek-glottalized	t̠ʃˀ
t/s-hacek-half-long	t̠ʃˑ
t/s-hacek-labialized	t̠ʃʷ
t/s-hacek-labiovelarized	t̠ʃʷˠ
t/s-hacek-lax	t̠ʃ̘
t/s-hacek-long	t̠ʃː
t/s-hacek-palatalized	t̠ʃʲ
t/s-hacek-preaspirated	ʰt̠ʃ
t/s-hacek-preglottalized	ˀt̠ʃ
t/s-hacek-prenasalized	n̠t̠ʃ
t/s-hacek-prenasalized-aspirated	n̠t̠ʃʰ
t/s-hacek-retroflex	ʈʂ
t/s-hacek-retroflex-aspirated	ʈʂʰ
t/s-hacek-retroflex-ejective	ʈʂ’
t/s-hacek-retroflex-prenasalized	ɳʈʂ

SPA code	Unicode IPA
t/s-hacek-tense	t̠͈ʃ
t/s-hacek-tense-labialized	t̠͈ʃʷ
t/s-hacek-tense-long	t̠͈ʃː
t/s-labialized	tsʷ
t/s-labiovelarized	tsʷˠ
t/s-laminal	t̻s̻
t/s-laminal-aspirated	t̻s̻ʰ
t/s-laminal-aspirated-weak	t̻s̻ʰ
t/s-laminal-ejective	t̻s̻’
t/s-laminal-ejective-syllabic	t̻s̻̩’
t/s-lax	t͉s
t/s-lax-long	d͉zː
t/s-long	tsː
t/s-palatalized	tsʲ
t/s-preaspirated	ʰts
t/s-preaspirated-long	ʰtsː
t/s-prenasalized	nts
t/s-prenasalized-aspirated	ntsʰ
t/s-retroflex	ʈʂ
t/s-retroflex-aspirated	ʈʂʰ
t/s-retroflex-aspirated-weak	ʈʂʰ
t/s-retroflex-ejective	ʈʂ’
t/s-tense	t͈s
t/s-tense-labialized	t͈sʷ
t/s/c-fricative	tsç
t/s/x	tsx
t/s/x-labialized	tsxʷ

SPA code	Unicode IPA
t/theta	t̪θ
t/theta-aspirated	t̪θʰ
t/theta-ejective	t̪θ’
t/theta-glottalized	t̠θˀ
t/theta-lax	t̪θ̙
t/x	tx
t/x-labialized	txʷ
t/x-uvular	tχ
theta	θ
theta-half-long	θˑ
theta-lax	θ̙
theta-long	θː
theta-prenasalized	n̪θ
u	u
u-breathy voice-long	ṳː
u-creaky voice	ṵ
u-creaky voice-long	ṵː
u-dot	ʉ
u-dot-long	ʉː
u-fronted	u̟
u-half-long	uˑ
u-half-voice	u
u-half-voice-long	uː
u-long	uː
u-long-fronted	u̟ː
u-long-nasalized	ũː
u-long-nasalized-weak	ũː

SPA code	Unicode IPA
u-nasalized	ũ
u-nasalized-weak	ũ
u-over-short	ŭ
u-over-short-nasalized	ŭ̃
u-retroflexed	u˞
u-trema	y
u-trema-long	yː
u-trema-nasalized	ỹ
u-trema-over-short	y̆
u-trema-voiceless	y̥
u-trema/schwa-glide	yə̯
u-voiceless	u̥
u-voiceless-long	u̥ː
u-voiceless-over-short	ŭ̥
u/e-dot	uɘ
u/schwa-glide	uə̯
u/w	uu̯
u/yod	ui
upsilon	ʊ
upsilon-breathy voice	ʊ̤
upsilon-creaky voice	ʊ̰
upsilon-creaky voice-long	ʊ̰ː
upsilon-dot	ʊ̈
upsilon-dot-long	ʊ̈ː
upsilon-dot-long-nasalized	ʊ̈̃ː
upsilon-dot-nasalized	ʊ̈̃
upsilon-fronted	ʊ̟

SPA code	Unicode IPA
upsilon-glide	ʊ̯
upsilon-long	ʊː
upsilon-long-nasalized	ʊ̃ː
upsilon-long-retracted	ʊ̠ː
upsilon-nasalized	ʊ̃
upsilon-nasalized-weak	ʊ̃
upsilon-over-short	ʊ̆
upsilon-retracted	ʊ̠
upsilon-retroflexed	ʊ˞
upsilon-trema	ʏ
upsilon-trema-voiceless-over-short	ʏ̥̆
upsilon-voiceless	ʊ̥
upsilon-voiceless-over-short	ʊ̥̆
upsilon/schwa-glide	ʊə̯
upsilon/u	ʊu
upsilon/w	ʊu
v	v
v-approximant	ʋ
v-approximant-long	ʋː
v-approximant-nasalized	ʋ̃
v-approximant-palatalized	ʋʲ
v-flap	ⱱ
v-half-long	vˑ
v-half-voice	v
v-labialized	vʷ
v-labiovelarized	vʷˠ
v-long	vː

SPA code	Unicode IPA
v-nasalized	ṽ
v-palatalized	vʲ
v-syllabic	v̩
v-tense	v͈
v-velarized	vˠ
w	w
w-creaky voice	w̰
w-creaky voice-nasalized	w̰̃
w-front	ɥ
w-front-nasalized	ɥ̃
w-front-voiceless	ɥ̥
w-glottalized	wˀ
w-half-voice	w
w-half-voice-long	wː
w-long	wː
w-long-nasalized	w̃ː
w-nasalized	w̃
w-over-short	w̆
w-preglottalized	ˀw
w-preglottalized-voiceless	ˀw̥
w-retroflexed	w˞
w-voiceless	ʍ
w/a	ua
w/epsilon	uɛ
w/iota	uɪ
w/o	uo
w/o-mid	uo̞

SPA code	Unicode IPA
w/o-open	uɔ
x	x
x-ejective	x’
x-half-long	xˑ
x-labialized	xʷ
x-long	xː
x-long-labialized	xʷː
x-long/r-trill-uvular-voiceless	xːʀ̥
x-palatalized	xʲ
x-prevelar	x̟
x-prevelar-palatalized	x̟ʲ
x-tense	x͈
x-tense-labialized	x͈ʷ
x-tense-long	x͈ː
x-tense-long-labialized	x͈ʷː
x-uvular	χ
x-uvular-labialized	χʷ
x-uvular-lax	χ̙
x-uvular-long	χː
x-uvular-long-pharyngealized	χˤː
x-uvular-palatalized	χʲ
x-uvular-pharyngealized	χˤ
x-uvular-tense	χ͈
x-uvular-tense-labialized	χ͈ʷ
x-velarized	xˠ
x/h	xh
x/r-trill-uvular-voiceless	xʀ̥

SPA code	Unicode IPA
yod	j
yod-creaky voice	j̰
yod-creaky voice-nasalized	j̰̃
yod-dot	ɨ̯
yod-glottalized	jˀ
yod-half-voice	j
yod-half-voice-long	jː
yod-lax	j͉
yod-lax-half-voice	j͉
yod-long	jː
yod-long-nasalized	j̃ː
yod-nasalized	j̃
yod-over-short	j̆
yod-preglottalized	ˀj
yod-preglottalized-voiceless	ˀj̥
yod-trema	ɰ
yod-trema-half-voice	ɰ
yod-trema-nasalized	ɰ̃
yod-trema-voiceless	ɰ̥
yod-trema/i-bar	ɰɨ
yod-trema/schwa	ɰə
yod-voiceless	j̥
yod/a	ia
yod/ash	iæ
yod/e	ie
yod/e-dot	iɘ
yod/e-mid	ie̞

SPA code	Unicode IPA
yod/e-nasalized	iẽ
yod/epsilon	iɛ
yod/o-long	ioː
yod/o-mid	io̞
z	z
z-approximant-labialized-syllabic	z̞̩ʷ
z-approximant-nasalized-velarized-syllabic	z̞̩̃ˠ
z-approximant-syllabic	z̞̩
z-approximant-velarized-syllabic	z̞̩ˠ
z-dental	z̪
z-dental-long	z̪ː
z-hacek	ʒ
z-hacek-half-voice	ʒ
z-hacek-half-voice-long	ʒː
z-hacek-long	ʒː
z-hacek-long-pharyngealized	ʒˤː
z-hacek-nasalized	ʒ̃
z-hacek-palatalized	ʒʲ
z-hacek-pharyngealized	ʒˤ
z-hacek-prenasalized	n̠ʒ
z-hacek-retroflex	ʐ
z-hacek-retroflex-glottalized	ʐˀ
z-hacek-syllabic	ʒ̩
z-hacek-tense	ʒ̤
z-hacek-tense-long	ʒ̤ː
z-hacek-velarized	ʒˠ
z-half-long-nasalized	z̃ˑ

SPA code	Unicode IPA
z-half-voice	z
z-half-voice-long	zː
z-labiovelarized	zʷˠ
z-laminal	z̻
z-long	zː
z-long-pharyngealized	zˤː
z-nasalized	z̃
z-palatalized	zʲ
z-pharyngealized	zˤ
z-prenasalized	nz
z-retroflex	z̢
z-syllabic	z̩
z-tense	z͈
z-tense-long	z͈ː
z-velarized	zˠ

Appendix F
UPSID AND IPA SEGMENT CORRESPONDENCES

For the mapping of UPSID ASCII segment codes and segment descriptions to Unicode IPA segments, the following points should be taken into consideration:

- affricated clicks represented with a “frictionalized” diacritic COMBINING X BELOW (U+0353), e.g. <k|͓>
- non-strident coronal fricatives are represented as their strident counterparts with a COMBINING EQUALS SIGN BELOW (U+0347), e.g. <ʃ͇>
- palatal lateral clicks are represented as palatal with a lateral release diacritic, the MODIFIER LETTER SMALL L (U+02E1) <ˡ>, e.g. <gǂ͓ˡ>
- “palatal sibilant” is mapped to the LATIN SMALL LETTER C WITH CURL <ɕ> (U+0255)
- UPSID’s tap/flap distinction is preserved; flaps are marked with the LATIN SMALL LETTER R WITH FISHHOOK (U+027E) <ɾ> ; taps are marked with the LATIN LETTER SMALL CAPITAL D (U+1D05) <ᴅ>
- the dental/alveolar underspecification in UPSID$_{451}$ is kept and is signified with a vertical bar, e.g. UPSID$_{451}$ ``voiceless dental/alveolar plosive" "t is represented as t̪|t
- both ``glottalized" and ``laryngealized" are mapped to MODIFIER LETTER GLOTTAL STOP (U+02C0) <ˀ> if the base segment is voiceless; if the base segment is voiced, each is mapped to COMBINING TILDE BELOW (U+0330) <o̰>

The full list of UPSID$_{451}$ segment descriptions and IPA interpretations is given below. The columns include “CCID”, “Description” and “CharCode” from the original UPSID$_{451}$ database tables. My IPA interpretation is given in the “IPA” column.

CCID	Description	CharCode	IPA
1	labialized voiceless labio-velar plosive	kpW	kp^{w}
2	labialized voiced labio-velar plosive	gbW	gb^{w}
3	prenasalized voiced labial-velar plosive	Nmgb	ŋmgb
4	voiceless aspirated labial-velar plosive	kph	kp^{h}
5	voiceless labial-velar plosive	kp	kp
6	voiced labial-velar plosive	gb	gb
7	labialized velarized voiceless aspirated bilabial plosive	pW-h	p^{wɣh}
8	labialized velarized voiced bilabial plosive	bW-	b^{wɣ}
9	prenasalized labialized voiced bilabial plosive	mbW	mb^{w}
10	labialized voiceless bilabial plosive	pW	p^{w}
11	labialized voiced bilabial plosive	bW	b^{w}
12	prenasalized palatalized voiced bilabial plosive	mbJ	mb^{j}
13	palatalized voiceless aspirated bilabial plosive	pJh	p^{jh}
14	palatalized voiceless bilabial plosive	pJ	p^{j}
15	palatalized breathy voiced bilabial plosive	bJh	b̤^{j}
16	palatalized voiced bilabial plosive	bJ	b^{j}
17	prenasalized voiceless aspirated bilabial plosive	mph	mp^{h}
18	prenasalized voiceless bilabial plosive	mp	mp
19	prenasalized voiced bilabial plosive	mb	mb
20	nasally-released voiced bilabial plosive	bm	b^{n}
21	voiceless aspirated bilabial plosive	ph	p^{h}
22	laryngealized voiceless bilabial plosive	p*	p^{ʔ}
23	long voiceless bilabial plosive	p:	pː
24	voiceless bilabial plosive with breathy release	phh	pɦ
25	voiceless preaspirated bilabial plosive	hp	^{h}p

CCID	Description	CharCode	IPA
26	voiceless bilabial plosive	p	p
27	laryngealized voiced bilabial plosive	b*	b̰
28	long voiced bilabial plosive	b:	bː
29	breathy voiced bilabial plosive	bh	b̤
30	voiced bilabial plosive	b	b
31	voiced labiodental plosive	bD	b̪
32	palatalized voiceless dental plosive	tDJ	t̪ʲ
33	palatalized voiced dental plosive	dDJ	d̪ʲ
34	pharyngealized voiceless dental plosive	tD9	t̪ˤ
35	pharyngealized voiced dental plosive	dD9	d̪ˤ
36	prenasalized voiceless aspirated dental plosive	ntDh	n̪t̪ʰ
37	prenasalized voiceless dental plosive	ntD	n̪t̪
38	prenasalized voiced dental plosive	ndD	n̪d̪
39	nasally released voiced dental plosive	dDn	d̪ⁿ
40	voiceless aspirated dental plosive	tDh	t̪ʰ
41	laryngealized voiceless dental plosive	tD*	t̪ˀ
42	voiceless dental plosive with breathy release	tDhh	t̪ɦ
43	voiceless dental plosive	tD	t̪
44	laryngealized voiced dental plosive	dD*	d̪̰
45	breathy voiced dental plosive	dDh	d̪̤
46	voiced dental plosive	dD	d̪
47	prenasalized labialized voiced dental/alveolar plosive	"ndW	n̪d̪ʷ|ndʷ
48	labialized voiceless aspirated dental/alveolar plosive	"tWh	t̪ʷʰ|tʷʰ
49	labialized voiceless dental/alveolar plosive	"tW	t̪ʷ|tʷ
50	labialized voiced dental/alveolar plosive	"dW	d̪ʷ|dʷ

CCID	Description	CharCode	IPA
51	prenasalized palatalized voiced dental/alveolar plosive	"ndJ	n̪d̪ʲ|ndʲ
52	palatalized voiceless aspirated dental/alveolar plosive	"tJh	t̪ʲʰ|tʲʰ
53	palatalized voiceless dental/alveolar plosive	"tJ	t̪ʲ|tʲ
54	palatalized voiced dental/alveolar plosive	"dJ	d̪ʲ|dʲ
55	velarized voiceless aspirated dental/alveolar plosive	"t-h	t̪ˠʰ|tˠʰ
56	pharyngealized voiceless dental/alveolar plosive	"t9	t̪ˤ|tˤ
57	pharyngealized voiced dental/alveolar plosive	"d9	d̪ˤ|dˤ
58	prenasalized voiceless dental/alveolar plosive	"nt	n̪t̪|nt
59	prenasalized voiced dental/alveolar plosive	"nd	n̪d̪|nd
60	voiceless aspirated dental/alveolar plosive	"th	t̪ʰ|tʰ
61	laryngealized voiceless dental/alveolar plosive	"t*	t̪ˀ|tˀ
62	long voiceless dental/alveolar plosive	"t:	t̪ː|tː
63	voiceless dental/alveolar plosive with breathy release	"thh	t̪ɦ|tɦ
64	voiceless dental/alveolar plosive	"t	t̪|t
65	laryngealized voiced dental/alveolar plosive	"d*	d̪̰|d̰
66	breathy voiced dental/alveolar plosive	"dh	d̪̤|d̤
67	voiced dental/alveolar plosive	"d	d̪|d
68	prenasalized palatalized voiced alveolar plosive	ndJ	ndʲ
69	palatalized voiceless alveolar plosive	tJ	tʲ
70	palatalized voiced alveolar plosive	dJ	dʲ
71	velarized voiceless alveolar plosive	t-	tˠ
72	velarized voiced alveolar plosive	d-	dˠ

CCID	Description	CharCode	IPA
73	prenasalized voiceless alveolar plosive	nt	nt
74	prenasalized voiced alveolar plosive	nd	nd
75	nasally-released voiced alveolar plosive	dn	dⁿ
76	voiceless aspirated alveolar plosive	th	tʰ
77	long voiceless alveolar plosive	t:	tː
78	voiceless alveolar plosive with breathy release	thh	tɦ
79	voiceless preaspirated alveolar plosive	ht	ʰt
80	voiceless alveolar plosive	t	t
81	laryngealized voiced alveolar plosive	d*	d̰
82	voiced alveolar plosive	d	d
83	prenasalized voiceless palato-alveolar plosive	nt_	n̠t̠
84	prenasalized voiced palato-alveolar plosive	nd_	n̠d̠
85	nasally-released voiced palato-alveolar plosive	d_n	d̠ⁿ
86	voiceless aspirated palato-alveolar plosive	t_h	t̠ʰ
87	voiceless palato-alveolar plosive	t_	t̠
88	laryngealized voiced palato-alveolar plosive	d_*	d̠̰
89	voiced palato-alveolar plosive	d_	d̠
90	prenasalized voiceless retroflex plosive	nt.	ɳʈ
91	prenasalized voiced retroflex plosive	nd.	ɳɖ
92	nasally-released voiced retroflex plosive	d.n	ɖⁿ
93	voiceless aspirated retroflex plosive	t.h	ʈʰ
94	laryngealized voiceless retroflex plosive	t.*	ʈˀ
95	voiceless retroflex plosive	t.	ʈ
96	laryngealized voiced retroflex plosive	d.*	ɖ̰
97	breathy voiced retroflex plosive	d.h	ɖ̤
98	voiced retroflex plosive	d.	ɖ
99	prenasalized voiceless palatal plosive	nc	ɲc

CCID	Description	CharCode	IPA
100	prenasalized voiced palatal plosive	ndj	ɲɟ
101	voiceless aspirated palatal plosive	ch	cʰ
102	long voiceless palatal plosive	c:	cː
103	voiceless palatal plosive	c	c
104	laryngealized voiced palatal plosive	dj*	ɟ̰
105	long voiced palatal plosive	dj:	ɟː
106	voiced palatal plosive	dj	ɟ
107	prenasalized labialized voiceless velar plosive	NkW	ŋkʷ
108	prenasalized labialized voiced velar plosive	NgW	ŋɡʷ
109	labialized voiceless aspirated velar plosive	kWh	kʷʰ
110	laryngealized labialized voiceless velar plosive	kW*	kʷˀ
111	long labialized voiceless velar plosive	kW:	kʷː
112	labialized voiceless velar plosive	kW	kʷ
113	labialized breathy voiced velar plosive	gWh	ɡ̤ʷ
114	labialized voiced velar plosive	gW	ɡʷ
115	prenasalized palatalized voiceless velar plosive	NkJ	ŋkʲ
116	palatalized voiceless aspirated velar plosive	kJh	kʲʰ
117	palatalized voiceless velar plosive	kJ	kʲ
118	palatalized voiced velar plosive	gJ	ɡʲ
119	pharyngealized voiceless velar plosive	k9	kˤ
120	prenasalized voiceless aspirated velar plosive	Nkh	ŋkʰ
121	prenasalized voiceless velar plosive	Nk	ŋk
122	prenasalized voiced velar plosive	Ng	ŋɡ
123	nasally-released voiced velar plosive	gn	ɡⁿ
124	laterally-released voiced velar plosive	gL	ɡˡ
125	voiceless aspirated velar plosive	kh	kʰ
126	laryngealized voiceless velar plosive	k*	kˀ

CCID	Description	CharCode	IPA
127	long voiceless velar plosive	k:	kː
128	voiceless velar plosive with breathy release	khh	kɦ
129	voiceless preaspirated velar plosive	hk	ʰk
130	voiceless velar plosive	k	k
131	breathy voiced velar plosive	gh	ɡ̤
132	voiced velar plosive	g	ɡ
133	labialized pharyngealized voiceless aspirated uvular plosive	qW9h	qʷˤʰ
134	labialized pharyngealized voiced uvular plosive	GW9	ɢʷˤ
135	labialized voiceless aspirated uvular plosive	qWh	qʷʰ
136	laryngealized labialized voiceless uvular plosive	qW*	qʷˀ
137	long labialized voiceless uvular plosive	qW:	qʷː
138	labialized voiceless uvular plosive	qW	qʷ
139	labialized voiced uvular plosive	GW	ɢʷ
140	pharyngealized voiceless aspirated uvular plosive	q9h	qˤʰ
141	pharyngealized voiced uvular plosive	G9	ɢˤ
142	prenasalized voiceless aspirated uvular plosive	Nqh	ɴqʰ
143	prenasalized voiceless uvular plosive	Nq	ɴq
144	voiceless aspirated uvular plosive	qh	qʰ
145	laryngealized voiceless uvular plosive	q*	qˀ
146	long voiceless uvular plosive	q:	qː
147	voiceless uvular plosive	q	q
148	voiced uvular plosive	G	ɢ
149	voiced pharyngeal plosive	99	ʕ̝
150	labialized glottal plosive	?W	ʔʷ

CCID	Description	CharCode	IPA
151	pharyngealized glottal plosive	?9	ʔˤ
152	glottal plosive	?	ʔ
153	voiced glottal plosive	??	ʔ̬
154	palatalized voiced bilabial implosive	bJ<	ɓʲ
155	voiceless bilabial implosive	p<	ɓ̥
156	voiced bilabial implosive	b<	ɓ
157	voiced dental implosive	dD<	ɗ̪
158	voiced dental/alveolar implosive	"d<	ɗ̪|ɗ
159	voiceless alveolar implosive	t<	ɗ̥
160	voiced alveolar implosive	d<	ɗ
161	voiced palato-alveolar implosive	d_<	ɗ̠
162	voiced retroflex implosive	d.<	ᶑ
163	voiced palatal implosive	dj<	ʄ
164	voiced velar implosive	g<	ɠ
165	voiceless uvular implosive	q<	ʛ̥
166	voiced uvular implosive	G<	ʛ
167	voiceless bilabial ejective stop	p'	pʼ
168	voiced bilabial ejective stop	b'	bʼ
169	voiceless dental ejective stop	tD'	t̪ʼ
170	voiceless dental/alveolar ejective stop	"t'	t̪ʼ|tʼ
171	voiceless alveolar ejective stop	t'	tʼ
172	voiced alveolar ejective stop	d'	dʼ
173	voiceless palatal ejective stop	c'	cʼ
174	labialized voiceless velar ejective stop	kW'	kʷʼ
175	palatalized voiceless velar ejective stop	kJ'	kʲʼ
176	voiceless velar ejective stop	k'	kʼ
177	voiced velar ejective stop	g'	ɡʼ

CCID	Description	CharCode	IPA
178	labialized pharyngealized voiceless uvular ejective stop	qW9'	qʷˤʼ
179	labialized voiceless uvular ejective stop	qW'	qʷʼ
180	pharyngealized voiceless uvular ejective stop	q9'	qˤʼ
181	voiceless uvular ejective stop	q'	qʼ
182	glottalized nasalized velarized voiceless alveolar click	hn/x?	ŋ̊!ˠˀ
183	velar-fricated voiceless aspirated alveolar click	/xh	k!xʰ
184	velar-fricated voiceless alveolar click	/x	k!x
185	glottalized velar-fricated voiced alveolar click	g/x?	g̰!x
186	velar-fricated voiced alveolar click	g/x	g!x
187	nasalized voiceless aspirated alveolar click	hn/h	ŋ̥!
188	glottalized nasalized voiceless alveolar click	hn/?	ŋ̥!ˀ
189	nasalized breathy voiced alveolar click	n/h	ŋ̤!
190	nasalized voiced alveolar click	n/	ŋ!
191	voiceless aspirated alveolar click	/h	k!ʰ
192	voiceless alveolar click	/	k!
193	breathy voiced alveolar click	g/h	g̤!
194	voiced alveolar click	g/	g!
195	velar-fricated voiceless aspirated palato-alveolar click	!xh	k!̠xʰ
196	nasalized voiceless aspirated palato-alveolar click	hn!h	ŋ̥!̠ʰ
197	glottalized nasalized voiceless palato-alveolar click	hn!?	ŋ̥!̠ˀ
198	nasalized voiced palato-alveolar click	n!	ŋ!̠
199	voiceless aspirated palatal-alveolar click	!h	k!̠ʰ
200	glottalized voiceless palatal-alveolar click	!?	k!̠ˀ

CCID	Description	CharCode	IPA
201	voiceless palato-alveolar click	!	k!̠
202	voiced palatal-alveolar click	g!	g!̠
203	glottalized nasalized velar-fricated voiceless palatal click	hn/=x?	ŋ̊ǂxˀ
204	velar-fricated voiceless palatal click	/=x	kǂx
205	glottalized velar-fricated voiced palatal click	g/=x?	g̰ǂx
206	velar-fricated voiced palatal click	g/=x	gǂx
207	nasalized voiceless aspirated palatal click	hn/=h	ŋ̊ǂʰ
208	glottalized nasalized voiceless palatal click	hn/=?	ŋ̊ǂˀ
209	nasalized breathy voiced palatal click	n/=h	ŋ̤ǂ
210	nasalized voiced palatal click	n/=	ŋǂ
211	voiceless aspirated palatal click	/=h	kǂʰ
212	voiceless palatal click	/=	kǂ
213	breathy voiced palatal click	g/=h	g̤ǂ
214	voiced palatal click	g/=	gǂ
215	voiced alveolar fricative flap	r[F	ɾ͓
216	voiced dental/alveolar fricative trill	"rF	r̪͓|r͓
217	fricative high front unrounded vowel	iF	i͓
218	fricative high back rounded vowel	uF	u͓
219	fricative high back unrounded lip-compressed vowel	uuF	ɯ͓
220	palatalized voiceless dental lateral fricative	hlDFJ	ɬ̪ʲ
221	palatalized voiced dental lateral fricative	lDFJ	ɮ̪ʲ
222	long voiceless dental lateral fricative	hlDF:	ɬ̪ː
223	voiceless dental lateral fricative	hlDF	ɬ̪
224	voiced dental lateral fricative	lDF	ɮ̪
225	long labialized voiceless dental/alveolar lateral fricative	"hlFW:	ɬ̪ʷː|ɬʷː

CCID	Description	CharCode	IPA
226	labialized voiceless dental/alveolar lateral fricative	"hlFW	ɬ̪ʷ|ɬʷ
227	palatalized voiceless dental/alveolar lateral fricative	"hlFJ	ɬ̪ʲ|ɬʲ
228	long voiceless dental/alveolar lateral fricative	"hlF:	ɬ̪ː|ɬː
229	voiceless dental/alveolar lateral fricative	"hlF	ɬ̪|ɬ
230	voiced dental/alveolar lateral fricative	"lF	ɮ̪|ɮ
231	voiceless velar-alveolar lateral fricative	hxlF	ɬʟ̥͓
232	palatalized voiceless alveolar lateral fricative	hlFJ	ɬʲ
233	palatalized voiced alveolar lateral fricative	lFJ	ɮʲ
234	voiceless alveolar lateral fricative	hlF	ɬ
235	voiced alveolar lateral fricative	lF	ɮ
236	voiced retroflex lateral fricative	l.F	ɭ͓
237	voiceless velar lateral fricative	hLF	ʟ̥͓
238	palatalized voiceless dental sibilant fricative	sDJ	s̪ʲ
239	palatalized voiced sibilant dental fricative	zDJ	z̪ʲ
240	pharyngealized voiceless dental sibilant fricative	sD9	s̪ˤ
241	pharyngealized voiced dental sibilant fricative	zD9	z̪ˤ
242	prenasalized voiced dental sibilant fricative	nzD	n̪z̪
243	laryngealized voiceless dental sibilant fricative	sD*	s̪ˀ
244	voiceless dental sibilant fricative	sD	s̪
245	voiced dental sibilant fricative	zD	z̪
246	long labialized voiceless dental/alveolar sibilant fricative	"sW:	s̪ʷː|sʷː
247	labialized voiceless dental/alveolar sibilant fricative	"sW	s̪ʷ|sʷ

CCID	Description	CharCode	IPA
248	labialized voiced dental/alveolar sibilant fricative	"zW	z̪ʷ|zʷ
249	palatalized voiceless dental/alveolar sibilant fricative	"sJ	s̪ʲ|sʲ
250	palatalized voiced dental/alveolar sibilant fricative	"zJ	z̪ʲ|zʲ
251	pharyngealized voiceless dental/alveolar sibilant fricative	"s9	s̪ˤ|sˤ
252	pharyngealized voiced dental/alveolar sibilant fricative	"z9	z̪ˤ|zˤ
253	prenasalized voiceless dental/alveolar sibilant fricative	"ns	n̪s̪|ns
254	prenasalized voiced dental/alveolar sibilant fricative	"nz	n̪z̪|nz
255	voiceless aspirated dental/alveolar sibilant fricative	"sh	s̪ʰ|sʰ
256	long voiceless dental/alveolar sibilant fricative	"s:	s̪ː|sː
257	voiceless preaspirated dental/alveolar sibilant fricative	"hs	ʰs̪|ʰs
258	voiceless dental/alveolar sibilant fricative	"s	s̪|s
259	voiced dental/alveolar sibilant fricative	"z	z̪|z
260	palatalized voiceless alveolar sibilant fricative	sJ	sʲ
261	pharyngealized voiceless alveolar sibilant fricative	s9	sˤ
262	prenasalized voiced alveolar sibilant fricative	nz	nz
263	laryngealized voiceless alveolar sibilant fricative	s*	sˀ
264	long voiceless alveolar sibilant fricative	s:	sː

CCID	Description	CharCode	IPA
265	voiceless alveolar sibilant fricative	s	s
266	voiced alveolar sibilant fricative	z	z
267	long labialized voiceless palato-alveolar sibilant fricative	SW:	ʃʷː
268	labialized voiceless palato-alveolar sibilant fricative	SW	ʃʷ
269	labialized voiced palatal-alveolar sibilant fricative	ZW	ʒʷ
270	palatalized voiceless palato-alveolar sibilant fricative	SJ	ʃʲ
271	palatalized voiced palato-alveolar sibilant fricative	ZJ	ʒʲ
272	velarized voiceless palato-alveolar sibilant fricative	S-	ʃˠ
273	velarized voiced palato-alveolar sibilant fricative	Z-	ʒˠ
274	prenasalized voiced palato-alveolar sibilant fricative	nZ	n̠ʒ
275	long voiceless palato-alveolar sibilant fricative	S:	ʃː
276	voiceless preaspirated palatal-alveolar sibilant fricative	hS	ʰʃ
277	voiceless palato-alveolar sibilant fricative	S	ʃ
278	breathy voiced palato-alveolar sibilant fricative	Zh	ʒ̤
279	voiced palato-alveolar sibilant fricative	Z	ʒ
280	voiceless retroflex sibilant fricative	s.	ʂ
281	laryngealized voiced retroflex sibilant fricative	z.*	ʐ̰
282	voiced retroflex sibilant fricative	z.	ʐ

CCID	Description	CharCode	IPA
283	voiceless palatal sibilant fricative	C,	ɕ
284	voiced palatal sibilant fricative	z,	ʑ
285	labialized velarized voiceless bilabial fricative	PW-	ɸʷˠ
286	labialized voiceless bilabial fricative	PW	ɸʷ
287	palatalized voiceless bilabial fricative	PJ	ɸʲ
288	palatalized voiced bilabial fricative	BJ	βʲ
289	voiceless bilabial fricative	P	ɸ
290	voiced bilabial fricative	B	β
291	labialized voiceless labiodental fricative	fW	fʷ
292	labialized voiced labiodental fricative	vW	vʷ
293	palatalized voiceless labio-dental fricative	fJ	fʲ
294	palatalized voiced labio-dental fricative	vJ	vʲ
295	prenasalized voiced labiodental fricative	mv	ɱv
296	long voiceless labio-dental fricative	f:	fː
297	voiceless labio-dental fricative	f	f
298	breathy voiced labiodental fricative	vh	v̤
299	voiced labio-dental fricative	v	v
300	palatalized voiced dental fricative	6DJ	ðʲ
301	voiceless dental fricative	0D	θ
302	voiced dental fricative	6D	ð
303	voiced dental/alveolar fricative	"6	z̳̪|z̳
304	voiced alveolar fricative	6	z̳
305	voiceless palato-alveolar fricative	0_	ʃ̳
306	voiceless palato-alveolar fricative	6_	ʒ̳
307	voiceless retroflex fricative	0.	ʂ̳
308	voiced retroflex fricative	6.	ʐ̳
309	labialized voiceless palatal fricative	CW	çʷ

CCID	Description	CharCode	IPA
310	voiceless palatal fricative	C	ç
311	voiced palatal fricative	jF	ʝ
312	long labialized voiceless velar fricative	xW:	xʷː
313	labialized voiceless velar fricative	xW	xʷ
314	labialized voiced velar fricative	gFW	ɣʷ
315	palatalized voiceless velar fricative	xJ	xʲ
316	palatalized voiced velar fricative	gFJ	ɣʲ
317	long voiceless velar fricative	x:	xː
318	voiceless velar fricative	x	x
319	laryngealized voiced velar fricative	gF*	ɣ̰
320	voiced velar fricative	gF	ɣ
321	long labialized pharyngealized voiceless uvular fricative	XW9:	χʷˤː
322	labialized pharyngealized voiceless uvular fricative	XW9	χʷˤ
323	labialized pharyngealized voiced uvular fricative	RFW9	ʁʷˤ
324	long labialized voiceless uvular fricative	XW:	χʷː
325	labialized voiceless uvular fricative	XW	χʷ
326	voiced uvular fricative	RFW	ʁʷ
327	long pharyngealized voiceless uvular fricative	X9:	χˤː
328	pharyngealized voiceless uvular fricative	X9	χˤ
329	pharyngealized voiced uvular fricative	RF9	ʁˤ
330	long voiceless uvular fricative	X:	χː
331	voiceless uvular fricative	X	χ
332	voiced uvular fricative	RF	ʁ
333	voiceless pharyngeal fricative	H	ħ
334	voiced pharyngeal fricative	9	ʕ

CCID	Description	CharCode	IPA
335	palatalized voiceless dental lateral ejective fricative	hlDFJ'	ɬ̪ʲ’
336	voiceless dental/alveolar lateral ejective fricative	"hlF'	ɬ̪’|ɬ’
337	voiceless dental sibilant ejective fricative	sD'	s̪’
338	voiceless dental/alveolar sibilant ejective fricative	"s'	s̪’|s’
339	voiceless alveolar sibilant ejective fricative	s'	s’
340	voiceless palato-alveolar sibilant ejective fricative	S'	ʃ’
341	voiceless retroflex sibilant ejective fricative	s.'	ʂ’
342	voiceless palatal sibilant ejective fricative	C,'	ɕ’
343	voiceless bilabial ejective fricative	P'	ɸ’
344	voiceless labio-dental ejective fricative	f'	f’
345	labialized voiceless velar ejective fricative	xW'	xʷ’
346	voiceless velar ejective fricative	x'	x’
347	labialized voiceless uvular ejective fricative	XW'	χʷ’
348	voiceless uvular ejective fricative	X'	χ’
349	voiceless retroflex affricated trill	t.r	ʈɹ̠̥
350	voiced retroflex affricated trill	d.r	ɖr̠͓
351	labialized voiceless aspirated dental/alveolar lateral affricate	"tlFWh	t̪ɬ̪ʷʰ|tɬʷʰ
352	voiceless aspirated dental/alveolar lateral affricate	"tlFh	t̪ɬ̪ʰ|tɬʰ
353	laryngealized voiceless dental/alveolar lateral affricate	"tlF*	t̪ɬ̪ˀ|tɬˀ
354	long voiceless dental/alveolar lateral affricate	"tlF:	t̪ɬ̪ː|tɬː
355	voiceless dental/alveolar lateral affricate	"tlF	t̪ɬ̪|tɬ

CCID	Description	CharCode	IPA
356	voiced dental/alveolar lateral affricate	"dlF	d̪ɮ̪|dɮ
357	voiceless velar plosive with alveolar lateral fricative release	klF	kɬ
358	voiceless aspirated alveolar lateral affricate	tlFh	tɬ^{h}
359	voiceless alveolar lateral affricate	tlF	tɬ
360	voiced alveolar lateral affricate	dlF	dɮ
361	voiceless palatalized dental sibilant affricate	tDsJ	t̪s̪^{j}
362	prenasalized voiceless aspirated dental sibilant affricate	ntDsh	n̪t̪s̪^{h}
363	prenasalized voiceless dental sibilant affricate	ntDs	n̪t̪s̪
364	prenasalized voiced dental sibilant affricate	ndDz	n̪d̪z̪
365	voiceless aspirated dental sibilant affricate	tDsh	t̪s̪^{h}
366	voiceless dental sibilant affricate	tDs	t̪s̪
367	voiced dental sibilant affricate	dDz	d̪z̪
368	labialized voiceless aspirated dental/alveolar sibilant affricate	"tsWh	t̪s̪^{wh}|ts^{wh}
369	long labialized voiceless dental/alveolar sibilant affricate	"tsW:	t̪s̪^{w}ː|ts^{w}ː
370	palatalized voiceless dental/alveolar sibilant affricate	"tsJ	t̪s̪^{j}|ts^{j}
371	palatalized voiced dental/alveolar sibilant affricate	"dzJ	d̪z̪^{j}|dz^{j}
372	prenasalized voiceless dental/alveolar sibilant affricate	"nts	n̪t̪s̪|nts
373	voiceless aspirated dental/alveolar sibilant affricate	"tsh	t̪s̪^{h}|ts^{h}
374	laryngealized voiceless dental/alveolar sibilant affricate	"ts*	t̪s̪^{ʔ}|ts^{ʔ}

CCID	Description	CharCode	IPA
375	long voiceless dental/alveolar sibilant affricate	"ts:	t̪s̪ː|tsː
376	voiceless dental/alveolar sibilant affricate with breathy release	"tshh	t̪s̪ɦ|tsɦ
377	voiceless dental/alveolar sibilant affricate	"ts	t̪s̪|ts
378	breathy voiced dental/alveolar sibilant affricate	"dzh	d̪z̪̤|dz̤
379	voiced dental/alveolar sibilant affricate	"dz	d̪z̪|dz
380	velarized voiceless alveolar sibilant affricate	ts-	tsˠ
381	velarized voiced alveolar sibilant affricate	dz-	dzˠ
382	prenasalized voiced alveolar sibilant affricate	ndz	ndz
383	voiceless aspirated alveolar sibilant affricate	tsh	tsʰ
384	laryngealized voiceless alveolar sibilant affricate	ts*	tsˀ
385	voiceless aspirated alveolar sibilant affricate with breathy release	tshh	tsɦ
386	voiceless alveolar sibilant affricate	ts	ts
387	breathy voiced alveolar sibilant affricate	dzh	dz̤
388	voiced alveolar sibilant affricate	dz	dz
389	labialized voiceless aspirated palato-alveolar sibilant affricate	tSWh	t̠ʃʷʰ
390	long labialized voiceless palato-alveolar sibilant affricate	tSW:	t̠ʃʷː
391	labialized voiceless palato-alveolar sibilant affricate	tSW	t̠ʃʷ
392	labialized voiced palato-alveolar sibilant affricate	dZW	d̠ʒʷ
393	palatalized voiceless aspirated palato-alveolar sibilant affricate	tSJh	t̠ʃʲʰ

CCID	Description	CharCode	IPA
394	palatalized voiceless palato-alveolar sibilant affricate	tSJ	t̠ʃʲ
395	palatalized voiced palato-alveolar sibilant affricate	dZJ	d̠ʒʲ
396	velarized voiceless palato-alveolar sibilant affricate	tS-	t̠ʃˠ
397	velarized voiced palato-alveolar sibilant affricate	dZ-	d̠ʒˠ
398	prenasalized voiceless aspirated palato-alveolar sibilant affricate	ntSh	n̠t̠ʃʰ
399	prenasalized voiceless palato-alveolar sibilant affricate	ntS	n̠t̠ʃ
400	prenasalized voiced palato-alveolar sibilant affricate	ndZ	n̠d̠ʒ
401	voiceless aspirated palato-alveolar sibilant affricate	tSh	t̠ʃʰ
402	laryngealized voiceless palato-alveolar sibilant affricate	tS*	t̠ʃˀ
403	long voiceless palato-alveolar sibilant affricate	tS:	t̠ʃː
404	voiceless preaspirated palato-alveolar sibilant affricate	htS	ʰt̠ʃ
405	voiceless palato-alveolar sibilant affricate	tS	t̠ʃ
406	breathy voiced palato-alveolar sibilant affricate	dZh	d̠ʒ̤
407	voiced palato-alveolar sibilant affricate	dZ	d̠ʒ
408	prenasalized voiceless aspirated retroflex sibilant affricate	nt.sh	ɳʈʂʰ

CCID	Description	CharCode	IPA
409	prenasalized voiceless retroflex sibilant affricate	nt.s	ɳʈʂ
410	prenasalized voiced retroflex sibilant affricate	nd.z	ɳɖʐ
411	voiceless aspirated retroflex sibilant affricate	t.sh	ʈʂʰ
412	voiceless retroflex sibilant affricate	t.s	ʈʂ
413	voiced retroflex sibilant affricate	d.z	ɖʐ
414	prenasalized voiceless palatal sibilant affricate	ncC,	ɲcɕ
415	prenasalized voiced palatal sibilant affricate	ndjz,	ɲɟʑ
416	voiceless aspirated palatal sibilant affricate	cC,h	cɕʰ
417	voiceless palatal sibilant affricate	cC,	cɕ
418	voiced sibilant palatal affricate	djz,	ɟʑ
419	voiceless aspirated labio-dental affricate	pfh	pfʰ
420	voiceless labio-dental affricate	pf	pf
421	voiced labio-dental affricate	bv	bv
422	voiceless aspirated dental affricate	tD0h	t̪θʰ
423	voiceless dental affricate	tD0	t̪θ
424	voiced dental affricate	dD6	d̪ð
425	voiceless aspirated alveolar affricate	t0h	ts̳ʰ
426	voiceless alveolar affricate	t0	ts̳
427	voiceless retroflex affricate	t.0	ʈʂ̳
428	labialized voiceless palatal affricate	cCW	cçʷ
429	labialized voiced palatal affricate	djjFW	ɟʝʷ
430	prenasalized voiced palatal affricate	ndjjF	ɲɟʝ
431	voiceless aspirated palatal affricate	cCh	cçʰ
432	voiceless palatal affricate	cC	cç
433	voiced palatal affricate	djjF	ɟʝ
434	labialized voiceless aspirated velar affricate	kxWh	kxʷʰ
435	voiceless aspirated velar affricate	kxh	kxʰ

CCID	Description	CharCode	IPA
436	long voiceless velar affricate	kx:	kxː
437	voiceless velar affricate	kx	kx
438	labialized pharyngealized voiceless uvular affricate	qXW9	qχʷˤ
439	labialized voiceless uvular affricate	qXW	qχʷ
440	pharyngealized voiceless uvular affricate	qX9	qχˤ
441	long voiceless uvular affricate	qX:	qχː
442	voiceless uvular affricate	qX	qχ
443	labialized voiceless dental/alveolar lateral ejective affricate	"tlFW'	t̪ɬ̪ʷ’|tɬʷ’
444	long voiceless dental/alveolar lateral ejective affricate	"tlF':	t̪ɬ̪’ː|tɬ’ː
445	voiceless dental/alveolar lateral ejective affricate	"tlF'	t̪ɬ̪’|tɬ’
446	voiceless alveolar lateral ejective affricate	tlF'	tɬ’
447	voiceless velar lateral ejective affricate	klF'	kʟ̥’
448	voiceless dental sibilant ejective affricate	tDs'	t̪s̪’
449	labialized voiceless dental/alveolar sibilant ejective affricate	"tsW'	t̪s̪ʷ’|tsʷ’
450	long voiceless dental/alveolar sibilant ejective affricate	"ts':	t̪s̪’ː|ts’ː
451	voiceless dental/alveolar sibilant ejective affricate	"ts'	t̪s̪’|ts’
452	voiceless alveolar sibilant ejective affricate	ts'	ts’
453	voiced alveolar sibilant ejective affricate	dz'	dz’
454	labialized voiceless palato-alveolar sibilant ejective affricate	tSW'	t̠ʃʷ’

CCID	Description	CharCode	IPA
455	long voiceless palatal-alveolar sibilant ejective affricate	tS':	t̠ʃʼː
456	voiceless palato-alveolar sibilant ejective affricate	tS'	t̠ʃʼ
457	voiced palato-alveolar sibilant ejective affricate	dZ'	d̠ʒʼ
458	voiceless retroflex sibilant ejective affricate	t.s'	ʈʂʼ
459	voiceless palatal sibilant ejective affricate	cC,'	cçʼ
460	voiceless labiodental ejective affricate	pf'	pfʼ
461	voiceless dental ejective affricate	tD0'	t̪θʼ
462	long voiceless velar ejective affricate	kx':	kxʼː
463	labialized pharyngealized voiceless uvular ejective affricate	qXW9'	qχʷˤʼ
464	labialized voiceless uvular ejective affricate	qXW'	qχʷʼ
465	long pharyngealized voiceless uvular ejective affricate	qX9':	qχˤː
466	pharyngealized voiceless uvular ejective affricate	qX9'	qχˤʼ
467	long voiceless uvular ejective affricate	qX':	qχʼː
468	voiceless uvular ejective affricate	qX'	qχʼ
469	velar-fricated voiceless aspirated alveolar lateral affricated click	#xh	k‖͓xʰ
470	nasalized voiceless aspirated alveolar lateral affricated click	hn#h	ŋ̥‖͓ʰ
471	glottalized nasalized voiceless alveolar lateral affricated click	hn#?	ŋ̥‖͓ˀ
472	nasalized voiced alveolar lateral affricated click	n#	ŋ‖͓

CCID	Description	CharCode	IPA
473	voiceless aspirated alveolar lateral affricated click	#h	kǁ͓ʰ
474	glottalized voiceless alveolar lateral affricated click	#?	kǁ͓ˀ
475	voiceless alveolar lateral affricated click	#	kǁ͓
476	voiced alveolar lateral affricated click	g#	ɡǁ͓
477	glottalized nasalized velar-fricated voiceless palatal lateral affricated click	hn#jx?	ŋ̊ǂ͓ˡxˀ
478	velar-fricated voiceless palatal lateral affricated click	#jx	kǂ͓ˡx
479	glottalized velar-fricated voiced palatal lateral affricated click	g#jx?	ɡ̰ǂ͓ˡx
480	velar-fricated voiced palatal lateral affricated click	g#jx	ɡǂ͓ˡx
481	nasalized voiceless aspirated palatal lateral affricated click	hn#jh	ŋ̊ǂ͓ˡʰ
482	glottalized nasalized voiceless palatal lateral affricated click	hn#j?	ŋ̊ǂ͓ˡˀ
483	nasalized breathy voiced palatal lateral affricated click	n#jh	ŋ̤ǂ͓ˡ
484	nasalized voiced palatal lateral affricated click	n#j	ŋǂ͓ˡ
485	voiceless aspirated palatal lateral affricated click	#jh	kǂ͓ˡʰ
486	voiceless palatal lateral affricated click	#j	kǂ͓ˡ
487	breathy voiced palatal lateral affricated click	g#jh	ɡ̤ǂ͓ˡ
488	voiced palatal lateral affricated click	g#j	ɡǂ͓ˡ
489	glottalized nasalized velar-fricated voiceless dental affricated click	hn|x?	ŋ̊ǀ͓xˀ

CCID	Description	CharCode	IPA
490	velarized voiceless aspirated dental affricated click	|xh	k|͓ˠʰ
491	velar-fricated voiceless dental affricated click	|x	k|͓x
492	glottalized velar-fricated voiced dental affricated click	g|x?	g̰|͓x
493	velar-fricated voiced dental affricated click	g|x	g|͓x
494	nasalized voiceless aspirated dental affricated click	hn|h	ŋ̊|͓ʰ
495	glottalized nasalized voiceless dental affricated click	hn|?	ŋ̊|͓ˀ
496	nasalized breathy voiced dental affricated click	n|h	ŋ̤|͓
497	nasalized voiced dental affricated click	n|	ŋ|͓
498	voiceless aspirated dental affricated click	|h	k|͓ʰ
499	glottalized voiceless affricated dental click	|?	k|͓ˀ
500	voiceless dental affricated click	|	k|͓
501	breathy voiced dental affricated click	g|h	g̤|͓
502	voiced dental affricated click	g|	g|͓
503	voiceless alveolar affricated click	/s	k!͓
504	voiced dental r-sound	rrD	*R̪
505	voiceless dental/alveolar r-sound	"hrr	*R̪̥|*R̥
506	laryngealized voiced dental/alveolar r-sound	"rr*	*R̪̰|*R̰
507	voiced dental/alveolar r-sound	"rr	*R̪|*R
508	palatalized voiced alveolar r-sound	rrJ	*Rʲ
509	voiced alveolar r-sound	rr	*R
510	laryngealized voiced dental tap	rDT*	ᴅ̪̰
511	voiced dental tap	rDT	ᴅ̪
512	voiced alveolar tap	rT	ᴅ

CCID	Description	CharCode	IPA
513	voiced dental/alveolar lateral flap	"l[	ɺ̪|ɺ
514	palatalized voiced alveolar lateral flap	l[J	ɺʲ
515	voiced alveolar lateral flap	l[	ɺ
516	laryngealized voiced retroflex lateral flap	l.[*	ɺ̠̰
517	voiced retroflex lateral flap	l.[	ɺ̠
518	voiced labio-dental flap	v[	ⱱ
519	voiced dental flap	rD[	ɾ̪
520	palatalized voiced dental/alveolar flap	"r[J	ɾ̪ʲ|ɾʲ
521	voiced dental/alveolar flap	"r[	ɾ̪|ɾ
522	palatalized voiceless alveolar flap	hr[J	ɾ̥ʲ
523	palatalized voiced alveolar flap	r[J	ɾʲ
524	velarized voiceless alveolar flap	hr[-	ɾ̥ˠ
525	velarized voiced alveolar flap	r[-	ɾˠ
526	glottalized voiced alveolar flap	r[*	ɾ̰
527	voiced alveolar flap	r[	ɾ
528	voiced palato-alveolar flap	r_[	ɾ̠
529	nasalized voiced retroflex flap	r.[{˜}	ɽ̃
530	voiced retroflex flap	r.[	ɽ
531	palatalized voiced dental trill	rDJ	r̪ʲ
532	voiced dental trill	rD	r̪
533	palatalized voiced dental/alveolar trill	"rJ	r̪ʲ|rʲ
534	velarized voiced dental/alveolar trill	"r-	r̪ˠ|rˠ
535	pharyngealized voiced dental/alveolar trill	"r9	r̪ˤ|rˤ
536	voiceless dental/alveolar trill	"hr	r̪̥|r̥
537	laryngealized voiced dental/alveolar trill	"r*	r̪̰|r̰
538	voiced dental/alveolar trill	"r	r̪|r
539	palatalized voiced alveolar trill	rJ	rʲ

CCID	Description	CharCode	IPA
540	prenasalized voiced alveolar trill	nr	nr
541	voiceless alveolar trill	hr	r̥
542	voiced alveolar trill	r	r
543	voiced retroflex trill	r.	r̠
544	voiced palatal trill	rj	r̠̝
545	voiced uvular trill	R	ʀ
546	palatalized voiced dental lateral approximant	lDJ	l̪ʲ
547	velarized voiced dental lateral approximant	lD-	l̪ˠ
548	voiced prestopped dental lateral approximant	dlD	d̪l̪
549	voiceless dental lateral approximant	hlD	l̪̊
550	long voiced dental lateral approximant	lD:	l̪ː
551	voiced dental lateral approximant	lD	l̪
552	palatalized voiced dental/alveolar lateral approximant	"lJ	l̪ʲ|lʲ
553	velarized voiced dental/alveolar lateral approximant	"l-	l̪ˠ|lˠ
554	pharyngealized voiced dental/alveolar lateral approximant	"l9	l̪ˤ|lˤ
555	voiceless dental/alveolar lateral approximant	"hl	l̪̊|l̥
556	laryngealized voiced dental/alveolar lateral approximant	"l*	l̪̰|l̰
557	breathy voiced dental/alveolar lateral approximant	"lh	l̪̤|l̤
558	voiced dental/alveolar lateral approximant	"l	l̪|l
559	palatalized voiceless alveolar lateral approximant	hlJ	l̥ʲ
560	palatalized voiced alveolar lateral approximant	lJ	lʲ

CCID	Description	CharCode	IPA
561	velarized voiceless alveolar lateral approximant	hl-	l̥ˠ
562	velarized voiced alveolar lateral approximant	l-	lˠ
563	pharyngealized voiced alveolar lateral approximant	l9	lˤ
564	nasalized voiced alveolar lateral approximant	l{˜}	l̃
565	voiceless alveolar lateral approximant	hl	l̥
566	laryngealized voiced alveolar lateral approximant	l*	l̰
567	voiced alveolar lateral approximant	l	l
568	breathy voiced palato-alveolar lateral approximant	l_h	l̠̤
569	voiced palato-alveolar lateral approximant	l_	l̠
570	voiceless retroflex lateral approximant	hl.	ɭ̥
571	voiced retroflex lateral approximant	l.	ɭ
572	voiced palatal lateral approximant	lj	ʎ
573	voiced velar lateral approximant	L	ʟ
574	voiced labial-palatal approximant	wj	ɥ
575	nasalized voiced labial-velar approximant	w{˜}	w̃
576	voiceless labial-velar approximant	hw	ʍ
577	laryngealized voiced labial-velar approximant	w*	w̰
578	voiced labial-velar approximant	w	w
579	palatalized voiced bilabial approximant	BAJ	β̞ʲ
580	velarized voiced bilabial approximant	BA-	β̞ˠ
581	voiceless bilabial approximant	PA	ɸ̞
582	long voiced bilabial approximant	BA:	β̞ː
583	voiced bilabial approximant	BA	β̞
584	voiced labio-dental approximant	vA	ʋ

CCID	Description	CharCode	IPA
585	voiceless dental/alveolar approximant	"hrA	ɹ̪̊|ɹ̥
586	voiced dental/alveolar approximant	"rA	ɹ̪|ɹ
587	voiced alveolar approximant	rA	ɹ
588	voiced palatal-alveolar approximant	j_	j̟
589	voiced retroflex approximant	r.A	ɻ
590	nasalized voiced palatal approximant	j{˜}	j̃
591	voiceless palatal approximant	hj	j̊
592	laryngealized voiced palatal approximant	j*	j̰
593	voiced palatal approximant	j	j
594	laryngealized voiced velar approximant	gA*	ɰ̰
595	voiced velar approximant	gA	ɰ
596	laryngealized labialized voiced uvular approximant	RAW*	ʁ̞̰^{w}
597	labialized voiced uvular approximant	RAW	ʁ̞^{w}
598	voiced uvular approximant	RA	ʁ̞
599	voiceless labial-velar nasal	hNm	ŋ̊m̥
600	voiced labial-velar nasal	Nm	ŋm
601	labialized velarized voiced bilabial nasal	mW-	m^{wɣ}
602	labialized voiced bilabial nasal	mW	m^{w}
603	palatalized voiced bilabial nasal	mJ	m^{j}
604	voiceless bilabial nasal	hm	m̥
605	laryngealized voiced bilabial nasal	m*	m̰
606	long voiced bilabial nasal	m:	mː
607	breathy voiced bilabial nasal	mh	m̤
608	voiced bilabial nasal	m	m
609	voiced labio-dental nasal	mD	ɱ
610	palatalized voiced dental nasal	nDJ	n̪^{j}
611	voiceless dental nasal	hnD	n̪̊

CCID	Description	CharCode	IPA
612	laryngealized voiced dental nasal	nD*	n̪̰
613	breathy voiced dental nasal	nDh	n̪̤
614	voiced dental nasal	nD	n̪
615	labialized voiced dental/alveolar nasal	"nW	n̪ʷ|nʷ
616	palatalized voiced dental/alveolar nasal	"nJ	n̪ʲ|nʲ
617	voiceless dental/alveolar nasal	"hn	n̪̊|n̥
618	laryngealized voiced dental/alveolar nasal	"n*	n̪̰|n̰
619	long voiced dental/alveolar nasal	"n:	n̪ː|nː
620	breathy voiced dental/alveolar nasal	"nh	n̪̤|n̤
621	voiced dental/alveolar nasal	"n	n̪|n
622	palatalized voiceless alveolar nasal	hnJ	n̥ʲ
623	palatalized voiced alveolar nasal	nJ	nʲ
624	velarized voiceless alveolar nasal	hn-	n̥ˠ
625	velarized voiced alveolar nasal	n-	nˠ
626	voiceless alveolar nasal	hn	n̥
627	laryngealized voiced alveolar nasal	n*	n̰
628	long voiced alveolar nasal	n:	nː
629	voiced alveolar nasal	n	n
630	voiceless palatal-alveolar nasal	hn_	n̠̊
631	breathy voiced palato-alveolar nasal	n_h	n̠̤
632	voiced palato-alveolar nasal	n_	n̠
633	voiceless retroflex nasal	hn.	ɳ̊
634	voiced retroflex nasal	n.	ɳ
635	labialized voiced palatal nasal	njW	ɲʷ
636	voiceless palatal nasal	hnj	ɲ̊
637	laryngealized voiced palatal nasal	nj*	ɲ̰
638	long voiced palatal nasal	nj:	ɲː

CCID	Description	CharCode	IPA
639	voiced palatal nasal	nj	ɲ
640	labialized voiceless velar nasal	hNW	ŋ̊ʷ
641	labialized voiced velar nasal	NW	ŋʷ
642	palatalized voiceless velar nasal	hNJ	ŋ̊ʲ
643	palatalized voiced velar nasal	NJ	ŋʲ
644	pharyngealized voiced velar nasal	N9	ŋˤ
645	voiceless velar nasal	hN	ŋ̊
646	laryngealized voiced velar nasal	N*	ŋ̰
647	breathy voiced velar nasal	Nh	ŋ̤
648	voiced velar nasal	N	ŋ
649	voiced uvular nasal	nU	ɴ
650	nasalized high front unrounded vowel with velar stricture	i{˜}-	ɨ̃
651	high front unrounded vowel with velar stricture	i-	ɨ
652	nasalized pharyngealized mid back rounded vowel	"o9{˜}	ọ̃ˤ
653	long nasalized pharyngealized lower mid back rounded vowel	O9{˜}:	ɔ̃ˤː
654	long nasalized pharyngealized low central unrounded vowel	a9{˜}:	ãˤː
655	nasalized pharyngealized low central unrounded vowel	a9{˜}	ãˤ
656	pharyngealized lowered high front unrounded vowel	I9	ɪˤ
657	long pharyngealized high front unrounded vowel	i9:	iˤː
658	pharyngealized high front unrounded vowel	i9	iˤ

CCID	Description	CharCode	IPA
659	pharyngealized lowered high back rounded vowel	U9	ʊˤ
660	long pharyngealized high back rounded vowel	u9:	uˤː
661	pharyngealized high back rounded vowel	u9	uˤ
662	pharyngealized mid front rounded vowel	"o/9	ø̞ˤ
663	pharyngealized mid front unrounded vowel	"e9	e̞ˤ
664	long pharyngealized mid back rounded vowel	"o9:	o̞ˤː
665	pharyngealized mid back rounded vowel	"o9	o̞ˤ
666	pharyngealized lower mid back rounded vowel	O9	ɔˤ
667	pharyngealized raised low front unrounded vowel	aa9	æˤ
668	pharyngealized raised low central unrounded vowel	49	ɐˤ
669	long pharyngealized low central unrounded vowel	a9:	aˤː
670	pharyngealized low central unrounded vowel	a9	aˤ
671	nasalized lowered high front unrounded vowel	I{˜}	ɪ̃
672	nasalized high front rounded vowel	y{˜}	ỹ
673	laryngealized nasalized high front unrounded vowel	i{˜}*	ḭ̃
674	long nasalized high front unrounded vowel	i{˜}:	ĩː
675	nasalized high front unrounded vowel	i{˜}	ĩ
676	nasalized lowered high central unrounded vowel	I_{˜}	ɨ̞̃
677	long nasalized high central unrounded vowel	i_{˜}:	ɨ̃ː
678	nasalized high central unrounded vowel	i_{˜}	ɨ̃
679	nasalized lowered high back rounded vowel	U{˜}	ʊ̃
680	long nasalized high back rounded vowel	u{˜}:	ũː

CCID	Description	CharCode	IPA
681	nasalized high back rounded vowel	u{˜}	ũ
682	nasalized high back unrounded vowel	uu{˜}	ɯ̃
683	long nasalized higher mid front unrounded vowel	e{˜}:	ẽː
684	nasalized higher mid front unrounded vowel	e{˜}	ẽ
685	nasalized higher mid central rounded vowel	@){˜}	ɵ̃
686	long nasalized higher mid central unrounded vowel	@{˜}:	ɘ̃ː
687	nasalized higher mid central unrounded vowel	@{˜}	ɘ̃
688	long nasalized higher mid back rounded vowel	o{˜}:	õː
689	nasalized higher mid back rounded vowel	o{˜}	õ
690	laryngealized nasalized mid front unrounded vowel	"e{˜}*	ẽ̞̰
691	nasalized mid front unrounded vowel	"e{˜}	ẽ̞
692	nasalized mid central unrounded vowel	"@{˜}	ə̃
693	nasalized fronted mid back unrounded vowel	"o(+{˜}	ɤ̞̟̃
694	laryngealized nasalized mid back rounded vowel	"o{˜}*	õ̞̰
695	nasalized mid back rounded vowel	"o{˜}	õ̞
696	nasalized mid back unrounded vowel	"o({˜}	ɤ̞̃
697	nasalized lower mid front rounded vowel	E){˜}	œ̃
698	long nasalized lower mid front unrounded vowel	E{˜}:	ɛ̃ː
699	nasalized lower mid front unrounded vowel	E{˜}	ɛ̃
700	nasalized lower mid central unrounded vowel	3{˜}	ɜ̃
701	long nasalized lower mid back rounded vowel	O{˜}:	ɔ̃ː
702	nasalized lower mid back rounded vowel	O{˜}	ɔ̃
703	nasalized lower mid back unrounded vowel	{ˆ}{˜}	ʌ̃

CCID	Description	CharCode	IPA
704	nasalized raised low front unrounded vowel	aa{˜}	æ̃
705	nasalized low front unrounded vowel	a+{˜}	ã̟
706	laryngealized nasalized low central unrounded vowel	a{˜}*	ã̰
707	long nasalized low central unrounded vowel	a{˜}:	ãː
708	nasalized low central unrounded vowel	a{˜}	ã
709	nasalized low back rounded vowel	a_){˜}	ɒ̃
710	long nasalized low back unrounded vowel	a_{˜}:	ɑ̃ː
711	nasalized low back unrounded vowel	a_{˜}	ɑ̃
712	lowered high front rounded vowel	Y	ʏ
713	overshort lowered high front unrounded vowel	IS	ɪ̆
714	lowered high front unrounded vowel	I	ɪ
715	long high front rounded vowel	y:	yː
716	high front rounded vowel	y	y
717	voiceless high front unrounded vowel	hi	i̥
718	laryngealized high front unrounded vowel	i*	ḭ
719	long high front unrounded vowel	i:	iː
720	breathy voiced high front unrounded vowel	ih	i̤
721	overshort high front unrounded vowel	iS	ĭ
722	high front unrounded vowel	i	i
723	lowered high central rounded vowel	U+	ʉ̞
724	lowered high central unrounded vowel	I_	ɨ̞
725	long high central rounded vowel	u+:	ʉː
726	overshort high central rounded vowel	u+S	ʉ̆
727	high central rounded vowel	u+	ʉ
728	retroflexed high central unrounded vowel	i_.	ɨ˞
729	long high central unrounded vowel	i_:	ɨː
730	overshort high central unrounded vowel	i_S	ɨ̆

CCID	Description	CharCode	IPA
731	high central unrounded vowel	i_	ɨ
732	long lowered high back rounded vowel	U:	ʊː
733	overshort lowered high back rounded vowel	US	ʊ̆
734	lowered high back rounded vowel	U	ʊ
735	overshort lowered high back unrounded vowel	UUS	ɯ̞̆
736	lowered high back unrounded vowel	UU	ɯ̞
737	voiceless high back rounded vowel	hu	u̥
738	laryngealized high back rounded vowel	u*	ṵ
739	long high back rounded vowel	u:	uː
740	breathy voiced high back rounded vowel	uh	ṳ
741	overshort high back rounded vowel	uS	ŭ
742	high back rounded vowel	u	u
743	long high back unrounded vowel	uu:	ɯː
744	breathy voiced high back unrounded vowel	uuh	ɯ̤
745	high back unrounded vowel	uu	ɯ
746	higher mid retracted front rounded vowel	o/_	ø̠
747	higher mid retracted front unrounded vowel	e_	e̠
748	long higher mid front rounded vowel	o/:	øː
749	overshort higher mid front rounded vowel	o/S	ø̆
750	higher mid front rounded vowel	o/	ø
751	laryngealized higher mid front unrounded vowel	e*	ḛ
752	long higher mid front unrounded vowel	e:	eː
753	breathy voiced higher mid front unrounded vowel	eh	e̤
754	higher mid front unrounded vowel	e	e
755	higher mid central rounded vowel	@)	ɵ
756	long higher mid central unrounded vowel	@:	ɘː

CCID	Description	CharCode	IPA
757	higher mid central unrounded vowel	@	ɘ
758	fronted higher mid back rounded vowel	o+	o̟
759	fronted higher mid back unrounded vowel	o(+	ɤ̟
760	laryngealized higher mid back rounded vowel	o*	o̰
761	long higher mid back rounded vowel	o:	oː
762	breathy voiced higher mid back rounded vowel	oh	o̤
763	higher mid back rounded vowel	oS	ŏ
764	higher mid back rounded vowel	o	o
765	breathy voiced higher mid back unrounded vowel	o(h	ɤ̤
766	higher mid back unrounded vowel	o(	ɤ
767	retracted mid front unrounded vowel	"e_	e̞̠
768	long mid front rounded vowel	"o/:	ø̞ː
769	mid front rounded vowel	"o/	ø̞
770	voiceless mid front unrounded vowel	"he	e̞̥
771	laryngealized mid front unrounded vowel	"e*	e̞̰
772	long mid front unrounded vowel	"e:	e̞ː
773	overshort mid front unrounded vowel	"eS	ĕ̞
774	mid front unrounded vowel	"e	e̞
775	overshort mid central rounded vowel	"@)S	ɵ̞̆
776	mid central rounded vowel	"@)	ɵ̞
777	retroflexed mid central unrounded vowel	"@.	ə˞
778	long mid central unrounded vowel	"@:	əː
779	overshort mid central unrounded vowel	"@S	ə̆
780	mid central unrounded vowel	"@	ə
781	fronted mid back rounded vowel	"o+	o̞̟
782	voiceless mid back rounded vowel	"ho	o̞̥
783	laryngealized mid back rounded vowel	"o*	o̞̰

CCID	Description	CharCode	IPA
784	long mid back rounded vowel	"o:	o̝ː
785	overshort mid back rounded vowel	"oS	ŏ̝
786	mid back rounded vowel	"o	o̝
787	long mid back unrounded vowel	"o(:	ɤ̝ː
788	mid back unrounded vowel	"o(	ɤ̝
789	lower mid front rounded vowel	E)	œ
790	laryngealized lower mid front unrounded vowel	E*	ɛ̰
791	long lower mid front unrounded vowel	E:	ɛː
792	breathy voiced lower-mid front unrounded vowel	Eh	ɛ̤
793	overshort lower mid front unrounded vowel	ES	ɛ̆
794	lower mid front unrounded vowel	E	ɛ
795	lower mid central rounded vowel	3)	ɞ
796	long lower mid central unrounded vowel	3:	ɜː
797	lower mid central unrounded vowel	3	ɜ
798	laryngealized lower mid back rounded vowel	O*	ɔ̰
799	long lower mid back rounded vowel	O:	ɔː
800	breathy voiced lower-mid back rounded vowel	Oh	ɔ̤
801	overshort lower mid back rounded vowel	OS	ɔ̆
802	lower mid back rounded vowel	O	ɔ
803	breathy voiced lower mid back unrounded vowel	{ˆ}h	ʌ̤
804	lower mid back unrounded vowel	{ˆ}	ʌ
805	long raised low front unrounded vowel	aa:	æː
806	raised low front unrounded vowel	aa	æ
807	long low front unrounded vowel	a+:	a̟ː
808	low front unrounded vowel	a+	a̟

CCID	Description	CharCode	IPA
809	overshort raised low central rounded vowel	4)S	ɐ̹̆
810	overshort raised low central unrounded vowel	4S	ɐ̆
811	raised low central unrounded vowel	4	ɐ
812	retroflexed low central unrounded vowel	a.	a˞
813	voiceless low central unrounded vowel	ha	ḁ
814	laryngealized low central unrounded vowel	a*	a̰
815	long low central unrounded vowel	a:	aː
816	breathy voiced low central unrounded vowel	ah	a̤
817	overshort low central unrounded vowel	aS	ă
818	low central unrounded vowel	a	a
819	raised low back rounded vowel	4)_	ɒ̝
820	raised low back unrounded vowel	4_	ɑ̝
821	long low back rounded vowel	a_):	ɒː
822	breathy voiced low back rounded vowel	a_)h	ɒ̤
823	overshort low back rounded vowel	a_)S	ɒ̆
824	low back rounded vowel	a_)	ɒ
825	long low back unrounded vowel	a_:	ɑː
826	low back unrounded vowel	a_	ɑ
827	nasalized pharyngealized mid back rounded to high front unrounded diphthong	oi9{˜}	õ̞ĩˤ
828	nasalized pharyngealized low central unrounded to mid front unrounded diphthong	ae9{˜}	ãẽ̞ˤ
829	nasalized pharyngealized low central unrounded to mid back rounded diphthong	ao9{˜}	ãõ̞ˤ
830	nasalized pharyngealized mid back rounded to low central unrounded diphthong	oa9{˜}	õ̞ãˤ
831	pharyngealized mid back rounded to high front unrounded diphthong	oi9	o̞iˤ

CCID	Description	CharCode	IPA
832	pharyngealized low central unrounded to mid front unrounded diphthong	ae9	ae̞ˤ
833	pharyngealized low central unrounded to mid back rounded diphthong	ao9	ao̞ˤ
834	pharyngealized mid back rounded to low central unrounded diphthong	oa9	o̞aˤ
835	nasalized mid front unrounded to high back rounded diphthong	eu{˜}	ẽ̞ũ
836	nasalized mid back rounded to high front unrounded diphthong	oi{˜}	õ̞ĩ
837	nasalized high front unrounded to mid front unrounded diphthong	ie{˜}	ĩẽ̞
838	nasalized mid front unrounded to high front unrounded diphthong	ei{˜}	ẽ̞ĩ
839	nasalized mid back rounded to high back rounded diphthong	ou{˜}	õ̞ũ
840	nasalized lower mid back rounded to high front unrounded diphthong	Oi{˜}	ɔ̃ĩ
841	nasalized low central unrounded to high front unrounded diphthong	ai{˜}	ãĩ
842	nasalized low front unrounded to high front unrounded diphthong	a+i{˜}	ã̟ĩ
843	nasalized high central unrounded to high front unrounded diphthong	i_i{˜}	ɨ̃ĩ
844	nasalized high back rounded to high front unrounded diphthong	ui{˜}	ũĩ
845	nasalized mid back rounded to low central unrounded diphthong	oa{˜}	õ̞ã

CCID	Description	CharCode	IPA
846	nasalized low back unrounded to lower mid back rounded diphthong	a_O{˜}	ɑ̃ɔ̃
847	nasalized lower mid front unrounded to lower mid back rounded diphthong	EO{˜}	ɛ̃ɔ̃
848	breathy voiced higher mid back rounded to high front unrounded diphthong	oih	o̤i̤
849	breathy voiced higher mid back unrounded to high front unrounded diphthong	o(ih	ɤ̤i̤
850	higher mid back unrounded to high front unrounded diphthong	o(i	ɤi
851	higher mid front rounded to high front rounded diphthong	o/y	øy
852	breathy voiced high front unrounded to mid central unrounded diphthong	i@h	i̤ə̤
853	high front unrounded to mid central unrounded diphthong	i@	iə
854	mid central unrounded to high front unrounded diphthong	@i	əi
855	high front unrounded to mid back rounded diphthong	io	io̞
856	mid front unrounded to high back rounded diphthong	eu	e̞u
857	mid back rounded to high front unrounded diphthong	oi	o̞i
858	lowered high front unrounded to mid front unrounded diphthong	Ie	ɪe̞
859	high front rounded to mid front rounded diphthong	yo/	yø̞

CCID	Description	CharCode	IPA
860	high front unrounded to mid front unrounded diphthong	ie	iẹ
861	mid front unrounded to high front unrounded diphthong	ei	ẹi
862	mid central unrounded to high back rounded diphthong	@u	əu
863	breathy voiced high back rounded to mid central unrounded diphthong	u@h	ṳə̤
864	high back rounded to mid central unrounded diphthong	u@	uə
865	mid central unrounded to high back unrounded diphthong	@uu	əɯ
866	breathy voiced high back unrounded to mid central unrounded diphthong	uu@h	ɯ̤ə̤
867	high back unrounded to mid central unrounded diphthong	uu@	ɯə
868	high central unrounded to mid central unrounded diphthong	i_@	ɨə
869	high back rounded to mid back rounded diphthong	uo	uo
870	mid back rounded to high back rounded diphthong	ou	ọu
871	lower mid central unrounded to high front unrounded diphthong	3i	ɜi
872	breathy voiced lower mid back rounded to high front unrounded diphthong	Oih	ɔ̤i̤
873	lower mid back rounded to high front unrounded diphthong	Oi	ɔi

CCID	Description	CharCode	IPA
874	lower mid back rounded to high front rounded diphthong	Oy	ɔy
875	high front unrounded to lower mid back unrounded diphthong	i{^}	iʌ
876	lower mid front unrounded to high back unrounded diphthong	Euu	ɛɯ
877	breathy voiced high front unrounded to lower mid front unrounded diphthong	iEh	i̤ɛ̤
878	high front unrounded to lower mid front unrounded diphthong	iE	iɛ
879	lower mid front unrounded to high front unrounded diphthong	Ei	ɛi
880	high back rounded to lower mid back unrounded diphthong	u{^}	uʌ
881	high back unrounded to lower mid back unrounded diphthong	uu{^}	ɯʌ
882	raised low front unrounded to high central rounded diphthong	aau+	æʉ
883	breathy voiced high front unrounded to low central unrounded diphthong	iah	i̤a̤
884	high front unrounded to low central unrounded diphthong	ia	ia
885	breathy voiced low central unrounded to high front unrounded diphthong	aih	a̤i̤
886	low central unrounded to high front unrounded diphthong	ai	ai
887	raised low front unrounded to high front unrounded diphthong	aai	æi

CCID	Description	CharCode	IPA
888	high front unrounded to low front unrounded diphthong	ia+	ia̟
889	low front unrounded to high front unrounded diphthong	a+i	a̟i
890	breathy voiced low central unrounded to high back rounded diphthong	auh	a̤ṳ
891	low central unrounded to high back rounded diphthong	au	au
892	breathy voiced high back rounded to low central unrounded diphthong	uah	ṳa̤
893	high back rounded to low central unrounded diphthong	ua	ua
894	breathy voiced low central unrounded to high back unrounded diphthong	auuh	a̤ɯ̤
895	low central unrounded to high back unrounded diphthong	auu	aɯ
896	high back unrounded to low central unrounded diphthong	uua	ɯa
897	high central unrounded to low central unrounded diphthong	i_a	ɨa
898	low central unrounded to high central unrounded diphthong	ai_	aɨ
899	high central unrounded to high front unrounded diphthong	i_i	ɨi
900	lowered high back rounded to high front unrounded diphthong	Ui	ʊi
901	breathy voiced high front unrounded to high back rounded diphthong	iuh	i̤ṳ

CCID	Description	CharCode	IPA
902	high front unrounded to high back rounded diphthong	iu	iu
903	breathy voiced high back rounded to high front unrounded diphthong	uih	ṳi̤
904	high back rounded to high front unrounded diphthong	ui	ui
905	breathy voiced high back unrounded to high front unrounded diphthong	uuih	ɯ̤i̤
906	high back unrounded to high front unrounded diphthong	uui	ɯi
907	lower mid front unrounded to mid back rounded diphthong	Eo	ɛo̝
908	mid front unrounded to low central unrounded diphthong	ea	e̝a
909	low central unrounded to mid front unrounded diphthong	ae	ae̝
910	low central unrounded to mid back rounded diphthong	ao	ao̝
911	mid back rounded to low central unrounded diphthong	oa	o̝a
912	mid front unrounded to mid central unrounded diphthong	e@	e̝ə
913	mid front unrounded to mid back rounded diphthong	eo	e̝o̝
914	mid back rounded to mid front unrounded diphthong	oe	o̝e̝
915	low central unrounded to lower mid front unrounded diphthong	aE	aɛ

CCID	Description	CharCode	IPA
916	labialized voiceless ”h”	hW	h^{w}
917	palatalized voiceless ”h”	hJ	h^{j}
918	laryngealized voiceless ”h”	h*	h^{ʔ}
919	voiceless ”h”	h	h
920	voiced ”h”	hh	ɦ
921	”h”	h2	*

VITA

Steven Moran is an alumnus of Eastern Michigan University, where he earned a Bachelor of Arts in Linguistics, German and Teaching English as a Second Language (TESOL). He was awarded the annual Distinguished Undergraduate Award in Linguistics. The Linguist List hired Moran as an undergraduate and subsequently awarded him a graduate student fellowship towards his studies at EMU. In 2006, Moran received his Master of Arts in Linguistics and Language Technology. His MA thesis, *A Grammatical Sketch of Isaalo (Western Sisaala)*, was written after four months of fieldwork in Northwestern Ghana and later published as a book. While working as a researcher for The Linguist List, Moran was the project lead and architect of the E-MELD School of Best Practices in Digital Language Documentation. The NSF-funded E-MELD project (2001–2005) generated and promoted digital standards for endangered languages documentation. After his work at The Linguist List, Moran undertook PhD studies in computational linguistics at the University of Washington. While at UW, Moran has worked in computer assisted language learning, user interface design, digital archiving, semantic web frameworks and in computational phonetics. He has been a consultant for digital archiving projects with native peoples of the Pacific Northwest. In 2010 he received the annual Distinguished Research Award in Linguistics from the UW Linguistics Department.