Designing an implementation language for a TEX successor

David Kastrup[*]

February 27, 2005

Abstract

Managing the complexity of TEX's codebase is an arduous task, so arduous that few mortals can hope to manage the underlying complexity. Its original author's computational roots date back to a time where the maturity and expressive power of existing programming languages was such that he chose to employ the assembly language of a fictional processor for the examples in his seminal work "The Art of Computer Programming". In a similar vein, TEX is written in a stripped-down subset of a now-extinct Pascal dialect. Current adaptations of the code base include more or less literal translations into Java (NTS and exTeX), C++ (the Omega-2.0 codebase), mechanically generated C (web2c) and a few others. In practically all currently available cases, the data structures and control flow and overall program structure mimick the original program to a degree that again requires the resourcefulness of a highly skilled programmer to manage its complexity. As a result, almost all of those projects have turned out to be basically single-person projects, and few projects have shown significant progress beyond providing an imitation of TEX.

It is the persuasion of the author that progressing significantly beyond the state of the art as represented by TEX will require the expressiveness and ease of use of a tailor-made implementation and extension language. Even a language as thwarted as Emacs Lisp has, due to its conciseness, rapid prototyping nature, extensibility and custom data types and its coevolution with the Emacs editor itself, enabled progress and add-ons reaching far beyond the original state as conceived by its original authors. This talk tries to answer the question what basic features an implementation platform and language for future typesetting needs should possess.

1 Problems of TEX

Managable problems

- Simplest measures such as `\boxstretch`, `\boxfilstretch`, `\boxshrink` etc are not available.
- Boxes can't reliably be deconstructed (`\special`, single characterse etc. can't be removed, boxes can only be taken apart from the end)
- Variables that TEX employs for decisions are partly unavailable (in some cases because of system-dependent rounding)
- Peculiarities like the loss of the first line's baseline (for `\vtop`) by whatsits, `\splittopskip0pt` and other.

Problems of the macro language

- Only global register pools indexed by number are available. There are no lexically local variables, the grouping structure does not match the macro structure.
- macro arguments get `\catcode`too soon, complex patterns are not easily parseable. Maybe `\lazy\def`would help?
- Implementing regular input languages is hard.

[*] dak@gnu.org

Interoperation problems

TEX

- only knows its own font formats, metrics and ligatures.
- does not talk to graphic programs
- can't trigger reformatting of external material.

Algorithmic problems

- TEX is either perfect, or deficient: paragraphs are optimized globally, but the vertical breaks are "local best fit" without feedback to horizontal breaks or future pages.
- TEX has no sane concept for asynchronous user code. `\output` is shielded with the expedient of additional grouping and has no multithreading concept.
- TEX has no possibilities for making use of side-effect free user-defined code. Consequently, user-defined code can't be used in several speculative contexts.

2 Document examples

2.1 Line numbers

Task at hand

If your ultimate goal is to produce a set of files in a different format that can be produced by GhostScript, take a look at the `tightpage` option of the preview package. This will embed the page dimensions into the PostScript code, obliterating the need to use the `-E -i` options to Dvips. You can then produce all image files with a single run of GhostScript from a single PostScript file for all images at once. The `tightpage` option requires setting the `dvips` option as well.

Various options exist that will pass TEX dimensions and other information about the respective shipped out material (including descender size) into the log file, where external applications might make use of it.

The possibility for generating a whole set of graphics with a single run of LATEX, Dvips, and GhostScript increases both speed and robustness of applications. It is to be hoped that applications like LATEX2HTML will be able to make use of this package in future.

Current line number implementations

Implementation with `lineno.sty`:

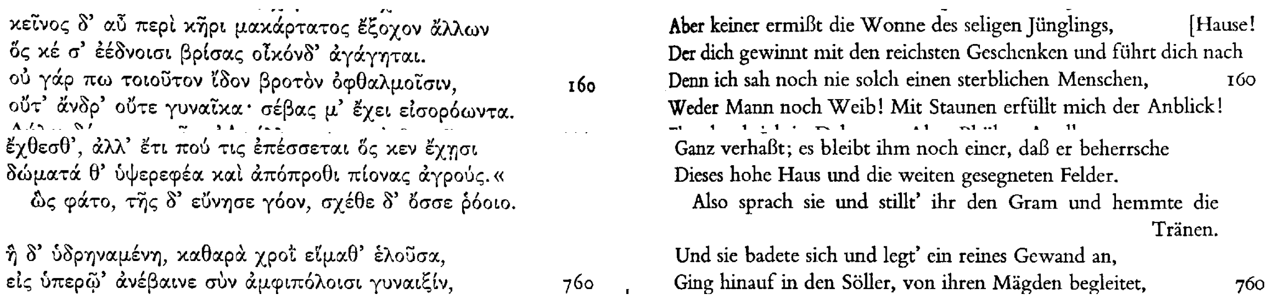
1. Replaces all interline penalties with forced page breaks.
2. This triggers a special output routine placed before the principal output routine.
3. This special routine places the line numbers and reinserts the correct penalties.
4. The normal Output routine is called.
5. A label-like multipass mechanism resets line numbers at the start of the page.

What would be saner for line numbering?

1. For migrating boxes into the main vertical list, a special "context" is defined that assembles a parallel column of 'unfinished' line numbers.
2. The unfinished objects take up constant dimensions and will be translated into glyphs either in the context of the output routine or at shipout time, since then the page start is known.
3. Consequently, a multipass algorithm is not necessary.
4. In the same context `\label`-commands referencing line numbers are expanded.

2.2 More complex Problems

Synchronized texts. . .

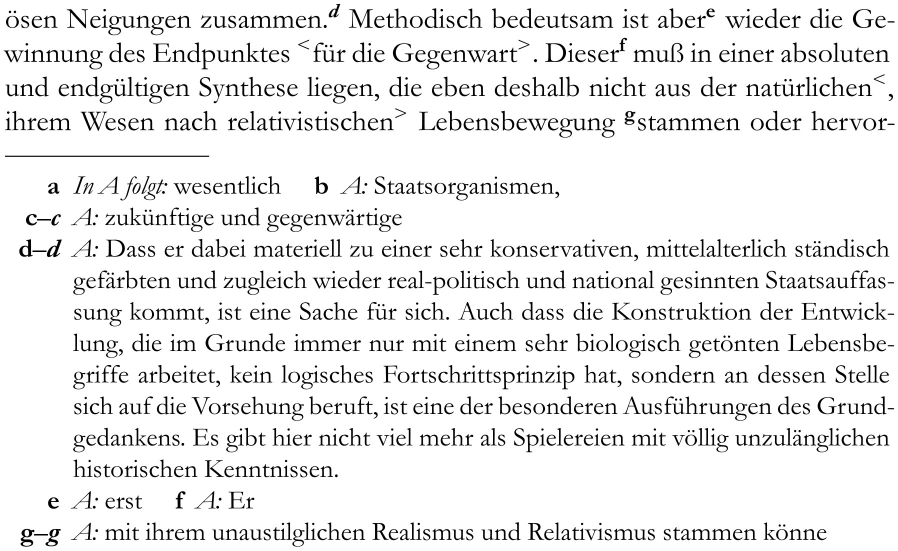


κεῖνος δ' αὖ περὶ κῆρι μακάρτατος ἔξοχον ἄλλων
ὅς κέ σ' ἐέδνοισι βρίσας οἶκόνδ' ἀγάγηται.
οὐ γάρ πω τοιοῦτον ἴδον βροτὸν ὀφθαλμοῖσιν, 160
οὔτ' ἄνδρ' οὔτε γυναῖκα· σέβας μ' ἔχει εἰσορόωντα.
ἔχθεσθ', ἀλλ' ἔτι πού τις ἐπέσσεται ὅς κεν ἔχῃσι
δώματά θ' ὑψερεφέα καὶ ἀπόπροθι πίονας ἀγρούς.«
ὣς φάτο, τῆς δ' εὔνησε γόον, σχέθε δ' ὄσσε ῥόοιο.

ἣ δ' ὑδρηναμένη, καθαρὰ χροῒ εἵμαθ' ἑλοῦσα,
εἰς ὑπερῷ' ἀνέβαινε σὺν ἀμφιπόλοισι γυναιξίν, 760

Aber keiner ermißt die Wonne des seligen Jünglings, [Hause!
Der dich gewinnt mit den reichsten Geschenken und führt dich nach
Denn ich sah noch nie solch einen sterblichen Menschen, 160
Weder Mann noch Weib! Mit Staunen erfüllt mich der Anblick!
Ganz verhaßt; es bleibt ihm noch einer, daß er beherrsche
Dieses hohe Haus und die weiten gesegneten Felder.
Also sprach sie und stillt' ihr den Gram und hemmte die Tränen.
Und sie badete sich und legt' ein reines Gewand an,
Ging hinauf in den Söller, von ihren Mägden begleitet, 760

Footnotes in running paragraphs



ösen Neigungen zusammen.[d] Methodisch bedeutsam ist aber[e] wieder die Gewinnung des Endpunktes <für die Gegenwart>. Dieser[f] muß in einer absoluten und endgültigen Synthese liegen, die eben deshalb nicht aus der natürlichen<, ihrem Wesen nach relativistischen> Lebensbewegung [g]stammen oder hervor-

a *In A folgt:* wesentlich **b** *A:* Staatsorganismen,

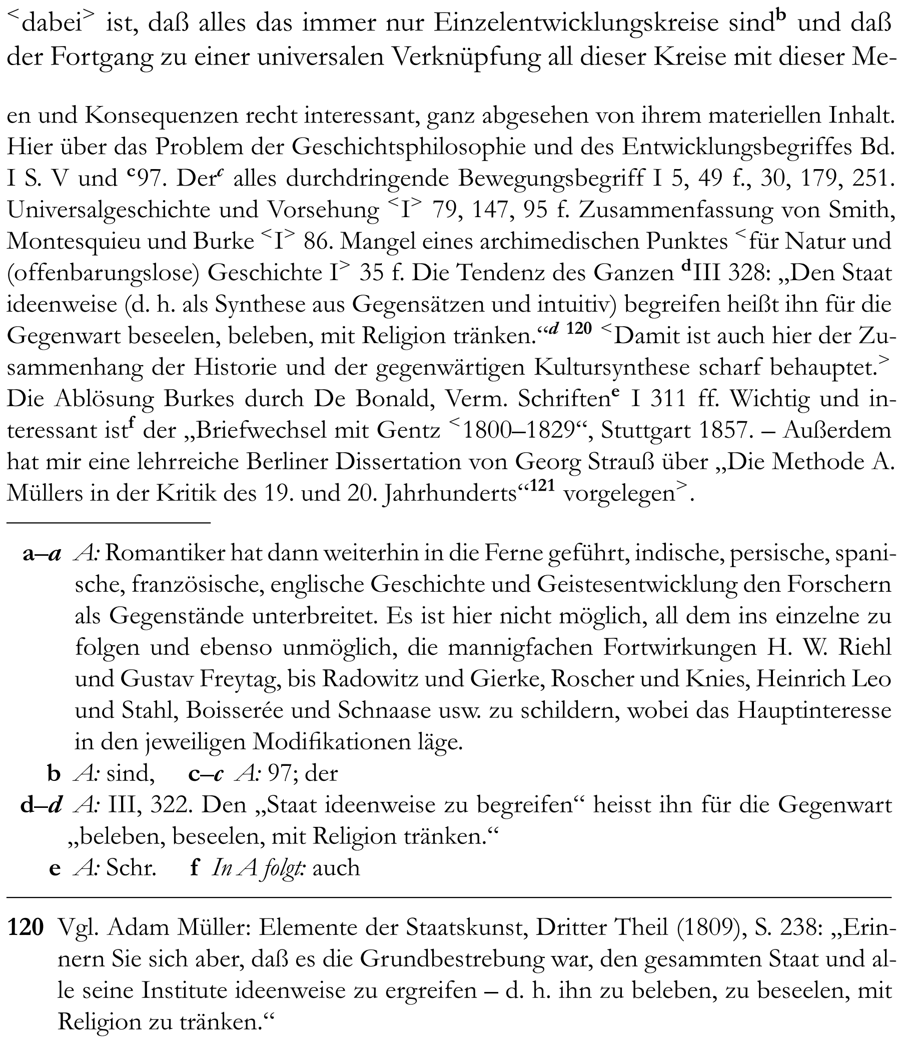
c–*c* *A:* zukünftige und gegenwärtige

d–*d* *A:* Dass er dabei materiell zu einer sehr konservativen, mittelalterlich ständisch gefärbten und zugleich wieder real-politisch und national gesinnten Staatsauffassung kommt, ist eine Sache für sich. Auch dass die Konstruktion der Entwicklung, die im Grunde immer nur mit einem sehr biologisch getönten Lebensbegriffe arbeitet, kein logisches Fortschrittsprinzip hat, sondern an dessen Stelle sich auf die Vorsehung beruft, ist eine der besonderen Ausführungen des Grundgedankens. Es gibt hier nicht viel mehr als Spielereien mit völlig unzulänglichen historischen Kenntnissen.

e *A:* erst **f** *A:* Er

g–*g* *A:* mit ihrem unaustilglichen Realismus und Relativismus stammen könne

Nested footnotes



<dabei> ist, daß alles das immer nur Einzelentwicklungskreise sind[b] und daß der Fortgang zu einer universalen Verknüpfung all dieser Kreise mit dieser Me-

en und Konsequenzen recht interessant, ganz abgesehen von ihrem materiellen Inhalt. Hier über das Problem der Geschichtsphilosophie und des Entwicklungsbegriffes Bd. I S. V und [c]97. Der[c] alles durchdringende Bewegungsbegriff I 5, 49 f., 30, 179, 251. Universalgeschichte und Vorsehung <I> 79, 147, 95 f. Zusammenfassung von Smith, Montesquieu und Burke <I> 86. Mangel eines archimedischen Punktes <für Natur und (offenbarungslose) Geschichte I> 35 f. Die Tendenz des Ganzen [d]III 328: „Den Staat ideenweise (d. h. als Synthese aus Gegensätzen und intuitiv) begreifen heißt ihn für die Gegenwart beseelen, beleben, mit Religion tränken.“[d] [120] <Damit ist auch hier der Zusammenhang der Historie und der gegenwärtigen Kultursynthese scharf behauptet.> Die Ablösung Burkes durch De Bonald, Verm. Schriften[e] I 311 ff. Wichtig und interessant ist[f] der „Briefwechsel mit Gentz <1800–1829“, Stuttgart 1857. – Außerdem hat mir eine lehrreiche Berliner Dissertation von Georg Strauß über „Die Methode A. Müllers in der Kritik des 19. und 20. Jahrhunderts“[121] vorgelegen>.

a–*a* *A:* Romantiker hat dann weiterhin in die Ferne geführt, indische, persische, spanische, französische, englische Geschichte und Geistesentwicklung den Forschern als Gegenstände unterbreitet. Es ist hier nicht möglich, all dem ins einzelne zu folgen und ebenso unmöglich, die mannigfachen Fortwirkungen H. W. Riehl und Gustav Freytag, bis Radowitz und Gierke, Roscher und Knies, Heinrich Leo und Stahl, Boisserée und Schnaase usw. zu schildern, wobei das Hauptinteresse in den jeweiligen Modifikationen läge.

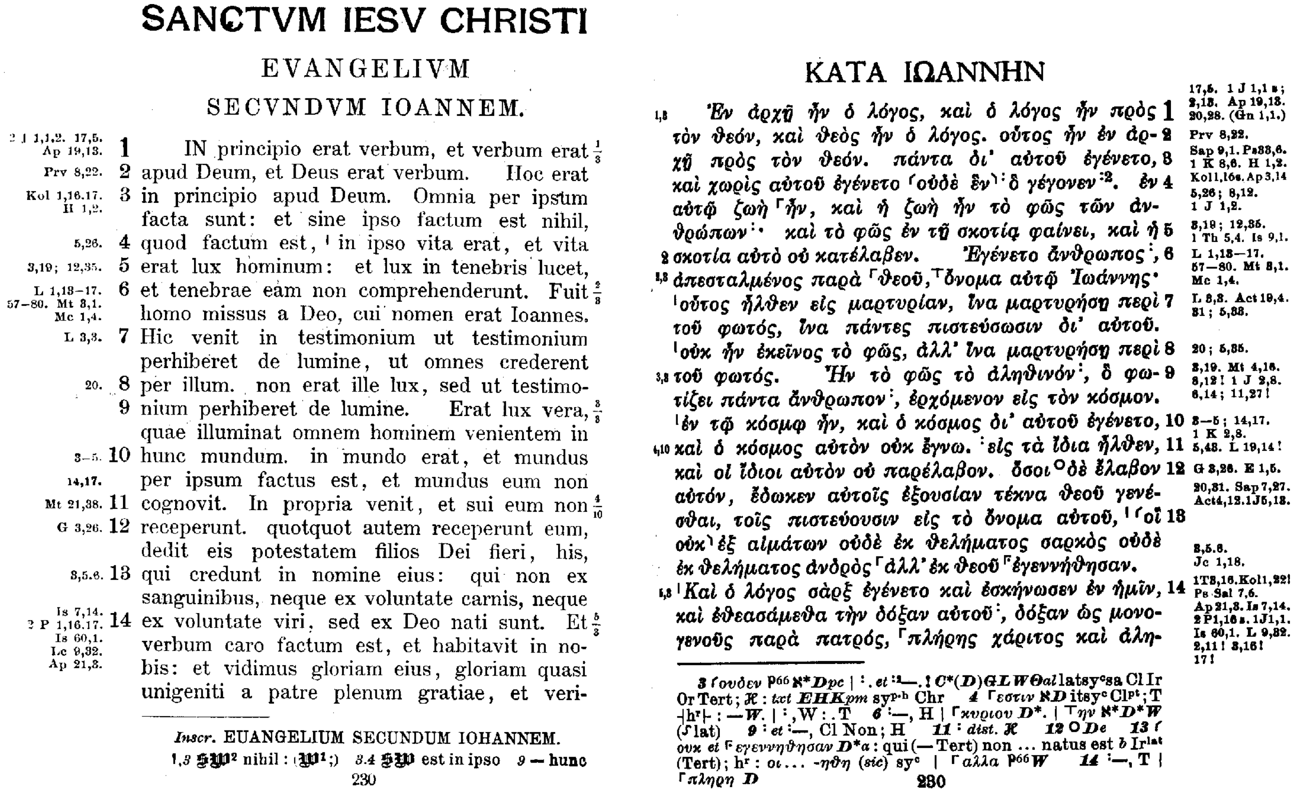
b *A:* sind, **c–*c*** *A:* 97; der

d–*d* *A:* III, 322. Den „Staat ideenweise zu begreifen“ heisst ihn für die Gegenwart „beleben, beseelen, mit Religion tränken.“

e *A:* Schr. **f** *In A folgt:* auch

[120] Vgl. Adam Müller: Elemente der Staatskunst, Dritter Theil (1809), S. 238: „Erinnern Sie sich aber, daß es die Grundbestrebung war, den gesammten Staat und alle seine Institute ideenweise zu ergreifen – d. h. ihn zu beleben, zu beseelen, mit Religion zu tränken.“

Tough stuff...



SANCTVM IESV CHRISTI
EVANGELIVM
SECVNDVM IOANNEM.

1 J 1,1.2. 17,5. Ap 19,13. 1 IN principio erat verbum, et verbum erat $\frac{1}{3}$
Prv 8,22. 2 apud Deum, et Deus erat verbum. Hoc erat
Kol 1,16.17. H 1,2. 3 in principio apud Deum. Omnia per ipsum
facta sunt: et sine ipso factum est nihil,
5,26. 4 quod factum est, ˡ in ipso vita erat, et vita
3,19; 12,35. 5 erat lux hominum: et lux in tenebris lucet,
L 1,13–17. 57–80. Mt 3,1. Mc 1,4. 6 et tenebrae eam non comprehenderunt. Fuit $\frac{2}{8}$
homo missus a Deo, cui nomen erat Ioannes.
L 3,3. 7 Hic venit in testimonium ut testimonium
perhiberet de lumine, ut omnes crederent
20. 8 per illum. non erat ille lux, sed ut testimo-
9 nium perhiberet de lumine. Erat lux vera, $\frac{3}{8}$
quae illuminat omnem hominem venientem in
3–5. 10 hunc mundum. in mundo erat, et mundus
14,17. per ipsum factus est, et mundus eum non
Mt 21,38. 11 cognovit. In propria venit, et sui eum non $\frac{4}{10}$
G 3,26. 12 receperunt. quotquot autem receperunt eum,
dedit eis potestatem filios Dei fieri, his,
3,5.6. 13 qui credunt in nomine eius: qui non ex
sanguinibus, neque ex voluntate carnis, neque
Is 7,14. 2 P 1,16.17. 14 ex voluntate viri, sed ex Deo nati sunt. Et $\frac{5}{8}$
Is 60,1. Lc 9,32. Ap 21,3. verbum caro factum est, et habitavit in no-
bis: et vidimus gloriam eius, gloriam quasi
unigeniti a patre plenum gratiae, et veri-

Inscr. EUANGELIUM SECUNDUM IOHANNEM.
1,3 𝔖𝔚𝔚^2 nihil : (𝔚𝔚^1;) *3.4* 𝔖𝔚𝔚 est in ipso *9* — **hunc**

230

ΚΑΤΑ ΙΩΑΝΝΗΝ

1,8 Ἐν ἀρχῇ ἦν ὁ λόγος, καὶ ὁ λόγος ἦν πρὸς 1 17,5. 1 J 1,1 s; 2,13. Ap 19,13. 20,28. (Gn 1,1.)
τὸν θεόν, καὶ θεὸς ἦν ὁ λόγος. οὗτος ἦν ἐν ἀρ- 2 Prv 8,22.
χῇ πρὸς τὸν θεόν. πάντα δι' αὐτοῦ ἐγένετο, 3 Sap 9,1. Ps33,6. 1 K 8,6. H 1,2.
καὶ χωρὶς αὐτοῦ ἐγένετο ⸂οὐδὲ ἕν⸃⁚ ὃ γέγονεν⁚². ἐν 4 Kol1,16s. Ap3,14
αὐτῷ ζωὴ ⸀ἦν, καὶ ἡ ζωὴ ἦν τὸ φῶς τῶν ἀν- 5,26; 8,12. 1 J 1,2.
θρώπων⁚· καὶ τὸ φῶς ἐν τῇ σκοτίᾳ φαίνει, καὶ ἡ 5 3,19; 12,35. 1 Th 5,4. Is 9,1.
2 σκοτία αὐτὸ οὐ κατέλαβεν. Ἐγένετο ἄνθρωπος⁚, 6 L 1,13–17. 57–80. Mt 3,1. Mc 1,4.
5,8 ἀπεσταλμένος παρὰ ⸀θεοῦ, ⸆ὄνομα αὐτῷ Ἰωάννης·
ˡοὗτος ἦλθεν εἰς μαρτυρίαν, ἵνα μαρτυρήσῃ περὶ 7 L 3,3. Act 19,4. 31; 5,33.
τοῦ φωτός, ἵνα πάντες πιστεύσωσιν δι' αὐτοῦ.
ˡοὐκ ἦν ἐκεῖνος τὸ φῶς, ἀλλ' ἵνα μαρτυρήσῃ περὶ 8 20; 5,35.
3,8 τοῦ φωτός. Ἦν τὸ φῶς τὸ ἀληθινόν⁚, ὃ φω- 9 3,19. Mt 4,16. 8,12! 1 J 2,8. 6,14; 11,27!
τίζει πάντα ἄνθρωπον⁚, ἐρχόμενον εἰς τὸν κόσμον.
ˡἐν τῷ κόσμῳ ἦν, καὶ ὁ κόσμος δι' αὐτοῦ ἐγένετο, 10 3–5; 14,17. 1 K 2,8.
4,10 καὶ ὁ κόσμος αὐτὸν οὐκ ἔγνω. ⁚εἰς τὰ ἴδια ἦλθεν, 11 5,43. L 19,14!
καὶ οἱ ἴδιοι αὐτὸν οὐ παρέλαβον. ὅσοι °δὲ ἔλαβον 12 G 3,26. E 1,5.
αὐτόν, ἔδωκεν αὐτοῖς ἐξουσίαν τέκνα θεοῦ γενέ- 20,31. Sap 7,27. Act4,12. 1J5,13.
σθαι, τοῖς πιστεύουσιν εἰς τὸ ὄνομα αὐτοῦ, ˡ⸂οἳ 13
οὐκ⸃ ἐξ αἱμάτων οὐδὲ ἐκ θελήματος σαρκὸς οὐδὲ 3,5.6.
ἐκ θελήματος ἀνδρὸς ⸀ἀλλ' ἐκ θεοῦ ⸀ἐγεννήθησαν. Jc 1,18.
1,8 ˡΚαὶ ὁ λόγος σὰρξ ἐγένετο καὶ ἐσκήνωσεν ἐν ἡμῖν, 14 1T3,16. Kol1,22! Ps Sal 7,6.
καὶ ἐθεασάμεθα τὴν δόξαν αὐτοῦ⁚, δόξαν ὡς μονο- Ap21,3. Is 7,14. 2P1,16s. 1J1,1.
γενοῦς παρὰ πατρός, ⸀πλήρης χάριτος καὶ ἀλη- Is 60,1. L 9,32. 2,11! 3,16! 17!

3 ⸂ουδεν 𝔓^{66} ***ℵ*D pc*** | ⁚ . *et* ⁚² —. ! ***C*(D)GLWΘ al*** lat sy^{c} sa Cl Ir Or Tert; 𝔎 : *txt* ***EHK pm*** sy$^{p.h}$ Chr **4** ⸀εστιν ***ℵD*** it sy^{c} Cl^{pt}; T ⫞h^{r}⊦ : —***W***. | ⁚ ,W : . T **6** ⁚ —, H | ⸀κυριου ***D****. | ⸆ην ***ℵ*D*W*** (⸉lat) **9** ⁚ *et* ⁚ —, Cl Non; H **11** ⁚ *dist.* 𝔎 **12** ° ***De*** **13** ⸂ ουκ *et* ⸀ εγεννηθησαν ***D*a*** : qui (— Tert) non ... natus est *b* Ir^{lat} (Tert); h^{r} : οι ... -ηθη (*sic*) sy^{c} | ⸀αλλα 𝔓^{66}***W*** **14** ⁚ —, T | ⸀πληρη ***D***

230

3 Concepts

Contexts

- A context is a programmatic entity with its own control flow and local variables.
- Example: an output context continuously requests material from the main vertical list and insertions. Collections of page matter are then scored (currently this happens using `\brokenpenalty`, `\widowpenalty`, `\clubpenalty`, `\badness` and others).
- The output context thus is coupled with the migration of page material from the vertical list to the current page.
- Other contexts may be coupled with other migrations.
- For example, a color context would have the current color as a local variable for material migrating to the page and into insertions.

Migrations

- Actions get triggered when objects of a class migrate from one list to another.
- Migrations can be penalized.
- When different migrations are possible, the combination with the smallest total penalties survives.
- Line breaking is a special example of penalized breakpoints during the migration of a horizontal into a vertical list.

Objects

- are elements of the various horizontal and vertical lists.
- can belong to different classes.
- classes can be added as well as extended.
- objects can have their own contexts for particular migrations.

Optimization

Global optimization leads to combinatorical explosion of run time. Countermeasures:

1. reduction of interdependencies by separated contexts
2. serialization by tying the optimization to migrations
3. limited backfeed, preferring multiple passes.
4. make do with less than full optimization.

Disadvantages

- higher memory impact since decisions need to remain revertible to some degree.
- higher computational resources because of backtracking
- quite a bit of potential for infinite or almost infinite loops and calculations.
- Programming a full TEX clone on such a platform appears possible, but pointless.
- Decomposition or analysis of several variants can be expensive.

Implementation language

- should offer natural expressivity for lists, TEX-typical strings and token lists.
- should make the required mechanism natively available.
- automatic garbage collection.
- need not be a single layer: instead of TEX's Pascal/TEX-macro layering a more tiered concept like C/Scheme/TEX-core/TEX-Macros would be possible.
- Problematic: Coroutines. Smalltalk? Ada?
- Problematic: I/O (memory for tentative I/O)?
- Combination with low-level languages like C desirable.
- Low-level implementation of fast algorithms on custom data structures should be possible
- Avoidance of unnecessary language features.