

TeX and Linguistics

Steve Peter
Beech Stave Press
310 Hana Road
Edison, NJ 08817, USA
speter@beechstave.com

Abstract

TeX has long been associated with mathematics and “hard” sciences such as physics. But even during the early days of TeX, linguists were attracted to the system, and today a growing number of them are turning to TeX (L^AT_EX, ConTeXt). Aside from the general advantages of TeX for producing academic papers, it offers linguists largely intuitive means for dealing with often complex notational issues. In this paper, an abbreviated version of my Practical TeX 2004 talk, I show some notational issues and their solutions in TeX.

Why TeX?

As a linguist and an avid user of TeX, I’m frequently asked why linguists would want to use TeX, as opposed to a word processor, to write their papers. Of course, there are the general reasons why any academic would benefit from TeX, such as easy handling of numbered examples, footnotes that make sense, bibliographic management via BIBTeX, etc.

For me, the main reason to use TeX to typeset linguistic papers and books is due to the complex, but often mathematically-inspired, notational systems used in the various subfields of linguistics. In fact, there are some cases¹ where the ability of TeX to format certain constructs aided their adoption by the field.

In this paper, I discuss various aspects of using TeX and L^AT_EX to typeset linguistics. One thing you will find largely absent here is a discussion of Omega, which should offer great hope for linguists using TeX. Until its development is further along, discussion of it only presents users with a utopian taste of what might be. I would dearly love for Omega to advance, but the wait (for me) has been painful. Your mileage may of course vary. *Digital Typography Using L^AT_EX* [1] has a good overview of language support in Omega.

I should note that the union of TeX and linguistics goes back quite far, even here in the pages of *TUGboat* [2, 3]. Donald E. Knuth² notes that linguists were among the first outside of mathematics to embrace TeX.

¹ Such as Attribute Value Matrices in Head-driven Phrase Structure Grammar.

² Personal communication at the Practical TeX banquet.

The field of linguistics

Linguistics is a large field that stands at the crossroads of several other disciplines, but is united in dealing with the scientific study of language. As you might expect from a large field, linguistics is commonly subdivided into various disciplines, each of which has various notational traditions and goals.

What is important for TeXnicians is that the notational issues presented here fall essentially into the general categories of “special symbols” or “special layouts”. In order to partition the field into units we can deal with here, though, let us instead adopt a fairly traditional division that linguists often use.

1. philology
2. phonology
3. phonetics
4. syntax
5. semantics
6. “hyphenated”
 - (a) psycholinguistics
 - (b) sociolinguistics
 - (c) biolinguistics
 - (d) discourse analysis
 - (e) ...

Fortunately for us, the “hyphenated” subdivisions largely make do with notation from other subfields, so they need not concern us further here.

In the remainder of the paper, I will take up each subfield in turn and discuss notational issues and how they may be solved with TeX. Most of the discussion deals with various L^AT_EX packages, which reflects the “market share” of L^AT_EX. Some of what

follows can be done with more or less difficulty in Plain TeX or ConTeXt.³

Due to the complexity of syntactic notation and the generality of application of tree structures, I will postpone discussion of Syntax until installment two of this paper.

Philology

Philology was once the general term for linguistic science, but is now more commonly used to refer to textual (rhetoric, poetics, textual criticism) and historical/diachronic linguistics. Most of the notational issues here deal either with different writing systems or with modifications of Latin.

The latter is usually quite straightforward in TeX, such as \bar{a} , \bar{e} , \bar{i} , \bar{o} , \bar{u} , \check{s} , \check{s} , etc., obtained by $\backslash=a$, $\backslash=e$, $\backslash=i$, $\backslash=o$, $\backslash=u$, $\backslash\{s\}$, $\backslash d\{s\}$. For example, Figure 1 is an example of something you might encounter in a paper on Indo-European.⁴

to me (although the suggestion of Kurylowicz, *Apophonie* 170, that the ablaut *CeHi* : *CHi* is paralleled by a type *CeRi* : *CRi* seems worth considering).

2.2.4. When the laryngeal followed **r, l, m, or n*, we expect the resonants to become α , ω , μ , ν , with the same *a* that appears outside laryngeal environments, and this is what we often get, e.g. ἔβαλον ‘they threw’ < **egʷʰ!E-*, ἔκαμον ‘they torched’ < **ekmA-*. But where the following laryngeal was *O*, we generally get α , ω : ἔμολον ‘they went’, ἔθορον ‘they leapt’, ἔπορε ‘pierced’, ἔπορον ‘they granted’ to the presents θώσκω, θώ(ι)σκω, τίτσώσκω and the perfect πέπωσται. An ablaut *ρω* : α or ω : α seems not to occur in Greek. F. B. J. Kuiper has suggested, *India Antiqua* 199, that the development to α and ω here was phonetically regular, the laryngeal influencing the vocalization of the resonant. I doubt that this is correct, because I believe that Dor. πάτρος represents a direct outcome of **pr̥O-tos* (cf. §2.3.2). If preconsonantal *O* failed to color the

Figure 1: What an Indo-Europeanist reads at the breakfast table

Paleography, for example, often uses a dot below a letter to indicate an obscured reading, which is quite easy to do in TeX (via $\backslash d\{ \}$), but in Word requires a special font, and a utility to access the needed characters.⁵ A few other symbols require the use of the TIPA package, which we will discuss below in the section on Phonology.

A major undertaking in philology is the production of critical editions (see Figure 2). The requirements of line numbers, cross-references lemmata, layer upon layer of notes, marginalia, etc. can bring a typesetter to the brink of madness, but for the edmac (Plain TeX), ledmac and ednotes (LATEX) packages. Unfortunately, none of these packages of-

³ Hans Hagen is aware of many of these issues, and we are working on adding more support for linguistics to ConTeXt.

⁴ From *The Collected Writings of Warren Cowgill*, Beech Stave Press, 2005.

⁵ I once tried to explain to my advisor how to get at some of these characters on a Mac. He shook his head and told me, “If I learn how to do that, I’ll forget Hittite.”

fers a complete solution, so you will need to select one based on your specific goals and circumstances. Uwe Lück offers a critical overview of these critical edition packages in [4]. Recently, David Kastrup has created the bigfoot package [5], but I have yet to try it, so I cannot offer an opinion.

1	A	A	dhuiñe gan chéill do mhaisligh an chléir
b	is	is	tharcaisnigh naomhscript na bhfáige,
c	na	na	haitheanta réab’s an t-aifreamh thréig
d	re	re	taithneamh do chlaonchreideamh Mhártain,
e	cá	cá	racairh do dhion ar fósa Nasardha
f	nuair	nuair	chaithimid cruimh bheith ar mhaoileann
			Joseph?
g	Ní	Ní	caraid Mac Crae chuim t’ana ‘phlé
h			ná Calvin bhias taobh ris an lá sin.
2	Nách	Nách	damanta an scéal don chreachaire chlaon
b	ghlac	ghlac	baiste na cléire ‘na pháiste
c	’s	’s	do glanadh mar ghréin ón bpeaca ró-dhaor
d	trí	trí	ainmhíofas Éva rinn Ádam,
e	tuitim	tuitim	arís f’chuing na haicme sin
f	tug	tug	atharrach brí don scribhinn bheannaithe,
g	d’alstrigh	d’alstrigh	béasa agus reachta na cléire
h	’s	’s	nách tugann aon ghéilleadh don Phápa?
3	Gach	Gach	scolaire baoth, ní mholaím a cheird
b	’tá	’tá	ag obair le gileáidh dá tháille
c	don	don	doirbhchoin chlaon dá ngorthar Mac Crae,
d	deisceabal	deisceabal	straeigh as an gcolláiste.
e	Tá	Tá	adaithe thíos in iochtar ifrim,
f	gan	gan	solas gan solse i dtíorthaibh dorcha,
g	tuigsint	tuigsint	an léinn, gach cuirpeacht déin
h	is Lucifer	is Lucifer	aosta ‘na mháistír.

²² *Teideal*: Dhuimhluing T, Seághan Mac Domhnaill cct B
1.a. dhuiñe T 1.a. mhaslaidh T, mhaslaig B 1.c. raob T 1.d. le B 1.e
dod B 1.f. chaithfamid T 1.f. maolinn B 1.g. phleidh T 1.h. bhíos B
1.h. leis B 2.a. claoñ B 2.c. glaumig T 2.d. ainmhíofas T, ainmhíofas B
2.d. Éabha B 2.g. is B 2.h. tuiginn T 3.a. sgollaíre T 3.a. mholluim T
3.b. ’tág ecobar T 3.b. re B 3.c. dorblchón daor B 3.d. straodhaig T
3.e. fhadoghthe thíos T 3.e. fadaigthe B 3.f. sollus T 3.g. cuirripeacht T
3.h. Luicifer T, Lucifer B 3.h. mhaighistir T

Figure 2: A critical edition

Typesetting Greek critical editions presents the same problems as above, plus the need for good Greek fonts. Claudio Beccari [6] has extended babel to produce a remarkable facsimile of the famous Teubner editions. It still lacks some refinement for producing the critical apparatus, but the package is under active development, and the results thus far are quite pleasing.

But Greek fonts aren’t an issue just when doing Greek critical editions. Due to whatever historical accident, Greek examples in philology are usually typeset in Greek, even while other languages that don’t use the Latin alphabet (such as Sanskrit, Russian, Armenian, Tocharian, etc.) are transliterated. Fortunately there are several options for getting and entering Greek examples. The Beccari Greek fonts are excellent, and there is also the PSGreek package [7], which bundles Greek PostScript fonts and a style file to make accessing them easier, by hiding some

of the horrors of encoding vectors. The quality of the PS fonts bundled is somewhat uneven, and installing new fonts for use in the same manner is not easy. To do so requires the `grkinst` fontinst plugin [8] and some time configuring. I wish it were a bit easier, since the `PSGreek` interface is one I find quite comfortable to use, and it has proven to be a life-saver for switching Greek fonts.

The Greek in Figure 1 was produced with `PSGreek`. For example, to get ‘ἔμολον’ ‘they went’, you enter `\textgreek{>}{emolon}`. I am now quite used to entering Greek in this manner, and therefore I can do it quite rapidly. However, you may be more comfortable entering Greek in Unicode, given an appropriate text editor. For that, put the following in your preamble:

```
\usepackage{ucs}
\usepackage[utf8]{inputenc}
\usepackage[polutonikogreek,english]{babel}
```

There are two general contexts for typesetting languages in alphabets other than Latin. First, of course there are times when you need to typeset a single language solely for speakers of that language, such as setting a Russian text in Cyrillic for a Russian reader. On the other hand, at times it is necessary to mix two or more languages, such as in dictionaries or instructional material.

Both scenarios are supported in `TeX`, although dealing with encoding vectors can cause a headache or two.⁶ Since I can't detail all possible language packages, let me limit myself here to a couple of packages I've found to be useful.

Underpinning nearly all multilingual endeavors in `LATeX` is `babel` by Johannes Braams [9]. It is included in (I believe) all `TeX` distributions, the manual is comprehensive and well written, and you should spend some time familiarizing yourself with it if you plan to do multilingual typesetting.

For Russian and the other Cyrillic-alphabet languages, there is the default Computer Modern Cyrillic font, which matches the standard Soviet look nicely.⁷ At some point, though, you'll no doubt want a change of pace. The `pscyr` package [10] contains a number of serif, sans serif, and a couple of display faces.

⁶ For Mac OS X users, the `XeTeX` system frees you from many of these problems. http://scripts.sil.org/cms/scripts/page.php?site_id=nrsi&item_id=xetex. However, you cannot exchange source files with colleagues who use other operating systems.

⁷ The Soviets heavily standardized book typefaces at a time when “modern” fonts were popular. There were some fantastic Russian typefaces developed during the 1920s that were neglected for decades.

Languages that use Indic scripts, such as Devānagari, have a complication that not all graphemes occur in the same order as they are pronounced, plus there are many, many di- and trigraphs. The `devnag` package [11] provides a preprocessor to take care of these complexities, plus good fonts and macros for both Plain `TeX` and `LATeX`. Using `devnag` makes it possible to typeset a bilingual critical edition with essentially the same input for both the Devānagari and the transliterated text. Figure 3 shows the vowels of Marāṭhī, typeset with the `devnag` package.

अ	आ	इ	ई	उ	ऊ
about	car	sit	seat	put	root
कृ	ल	ए	ऐ	ओ	औ
under	bottle	say	by	road	loud
अं	अः				

Figure 3: The vowels of Marāṭhī

For languages written in the Arabic alphabet (such as Arabic, Persian, Pashto, etc.), Klaus Lagally's `ArabTeX` is a must. The system is by now quite stable, and the output is very good. Several people are working on various extensions, especially for typesetting Arabic mathematics. See for example, Lazrek et al. [12, 13]. While it is possible to typeset Hebrew using `ArabTeX`, Alan Hoenig's `Makor` [14] is worth every penny.⁸

Typesetting Chinese using `TeX` is possible with the `CJK` [15] package (which provides for much more than just Chinese, Japanese, and Korean support). However, I prefer `ConTeXt`, due to its support of visual debugging via `\tracechinesetrue`. Numbering can be toggled between Chinese and western styles via `[conversion=chinese]` or `[conversion=numbers]`. More traditional vertical typesetting is possible essentially by flowing the text into narrow columns.

Semantics

Semantics is the study of meaning, and the notation used is tied closely to formal logic. Thus it is very straightforward to typeset with `TeX`. So the function of the set of things similar to houses is denoted by $\lambda x \text{Similar_to}(x, \text{houses})$. The `TeX` to get this is $\$\\lambda x \\textit{Similar_to}(x, \\textit{houses})\$$. We had to wrap the ‘English’ inside the function with `\textit` to prevent `TeX`

⁸ Yes, it is free software, and yes, I am making an exception to not discussing Omega software.

from interpreting the words as a series of variables. In some cases `\mbox` will work, and note that sometimes spaces inside the `\mbox`s are important. So a possible interpretation for the sentence *I have told one friend of mine all those stories*⁹ is given as $\exists x \forall y [(x \in \text{friends of mine} \wedge y \in \text{those stories}) \rightarrow I \text{ have told } y \text{ to } x]$, or in TeX terms

```
$\exists x \ [\forall y [
  (x \in \mbox{\{friends of mine\}}
   \wedge y \in \mbox{\{those stories\}})
  \rightarrow \mbox{\{I have told\}}
  y \mbox{\{ to \}} x]]$
```

Double brackets (representing semantic evaluation) are provided by the `stmaryrd` package [16]. With that loaded, typing $\llbracket(MN)\rrbracket^M$ yields $\llbracket(MN)\rrbracket^M$. You may also need to load the `latexsym` package for an occasional symbol.

Phonetics and Phonology

Phonetics is a branch of acoustics that deals with speech sounds and their production and perception. The notation used is a combination of transcription symbols (as covered below) and diagrams representing articulatory spaces. For example, Figure 4 shows a typical representation of a vowel system.¹⁰

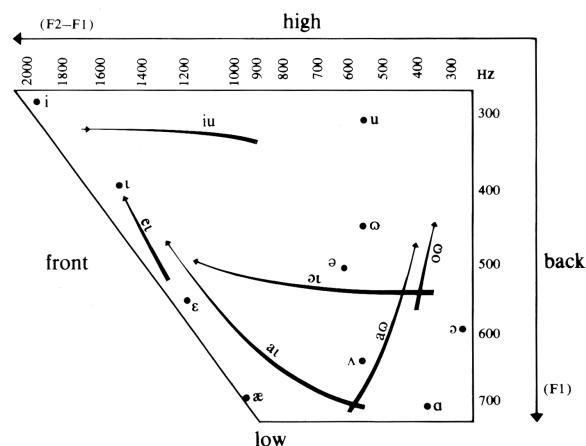


Figure 4: Some American English vowels

Also frequent are diagrams of the human vocal tract. Unfortunately there is no easy way to handle automatic generation of notation of this type. A typical way to handle them is to create the illustra-

⁹ From Ray Jackendoff, *Semantic Interpretation in Generative Grammar*, Cambridge: MIT Press, 1972, p. 308.

¹⁰ From Peter Ladefoged, *A Course in Phonetics*, San Diego: Harcourt Brace Jovanovich, 1982, p. 198.

tion in a vector program (such as Inkscape or Illustrator) and then to import it into TeX. Given the complexity of creating them, publishers (and therefore also authors) have been reluctant to use them aside from in very specialized books.¹¹

The International Phonetic Association came into existence in 1886 with a goal of promoting phonetics in education and the creation of an international phonetic alphabet (now known as the IPA) for the universal transcription of languages.¹² A separate tradition of transcription developed among anthropological linguists in America. Both systems of transcription [18, 19] are supported via the `TIPA` package [17].

While there are numerous fonts that provide IPA symbols that more or less match existing typefaces, there are to my knowledge still only a small number of type families that have complete sets of corresponding IPA symbols: Computer Modern, Lucida, Times, Le Monde, Gentium, Garamond, and Stone—and two of them are provided for by `TIPA`. To wit, the Computer Modern IPA symbols work best with Computer Modern,¹³ but they will fit in reasonably with other vertical-stress typefaces. The Times IPA symbols, again, work best with Times, but in a pinch, they fit in with other oblique-stress typefaces. In a sans-serif environment, `TIPA` provides a Helvetica-like symbol set.

In addition to the IPA fonts and the interface (more on which below), `TIPA` provides a style file to produce simple vowel diagrams (simpler than the one shown in Figure 4). I could conceive—given enough labor—of creating the more complex charts like Figure 4, with arrows and swooshes, programmatically via the tools provided by `PSTricks` [20] or `MetaFun` [21].

`TIPA` provides for a couple of different ways to enter phonetic notation. There are long forms that have generally mnemonic names, so I can write `[əh'a]` as `[\text{schwa} \text{ h}\text{\textit{textprimstress}} \text{ a}]` if my paper uses a limited set of symbols, and I don't want to learn the more involved transcription.

On the other hand, if you need to input larger amounts of transcription, it is useful to enter the IPA environment via `\textipa{}`, `\tipaencoding`, or `\begin{IPA}` and `\end{IPA}`. So, if we enter the IPA environment and type

```
D@ "n0;T "wInd @nd D@ "s2n w@ dIs"pju;tIN
wItS w@z D@ "str6Ng5, wEn @ "tr\ae v15
keIm @"16N "r\ae pt In @ "w0:m "kloUk. DeI
```

¹¹ They can hardly be avoided in an introduction to phonetics, for example.

¹² Remember that there were no tape recorders in 1886!

¹³ And variants such as Latin Modern.

@"gri:d D@t D@ "w2n hu; f3;st s@k"si;dId
 In "meIkiN D@ "tr\ae v15 teIk hIz "kloUk
 6f SUd bI k@n"sId@d "str6Ng@ D@n DI "2D@.
 this will be rendered into IPA as follows:¹⁴
 ðə 'nɔθ 'wm̩ ənd ðə 'sæn wə dis'pjurtnj wɪtʃ wəz ðə
 'strɔŋgə, wən ə 'trævle kem ə'lɔŋ 'ræpt m ə 'wɔ:m
 'kloʊk. ðei ə'grɪ:d ðət ðə 'wʌn hu' f3;st sək'si:did m
 'merkiŋ ðə 'trævle teik hrz 'kloʊk pf ſud br kən'sɪdə
 'strɔŋgə ðən ði 'ʌðə.

TIPA allows you to enter tonal specifications and has many other nice features to explore. I heartily recommend reading the excellent manual included in the package.

As I mentioned earlier in the paper, the TIPA package also allows you to enter Indo-European reconstructed forms. For example, the work for ‘100’ is reconstructed as **kmtóm*, which can be entered as \textroundcap{k}\textsubring{m}t\om or as \textipa{*|\c{k}\r*mt\om}.

A new notational twist entered both phonology and syntax within the past decade as Optimality Theory grew in popularity. The central part of its notation are the so-called optimality tableaux. There are a number of ways to enter them, but I’ve settled on using PStricks together with colortab [22]. Here is source and output using some totally non-sensical data.

```
\begin{tabular}[t]{r|c|c|c|c}
\cline{2-4}
&/ba/ & \textipa{!*} & b\textturna\\
\hline
\w & [ba] & & *\hline
& [*ba] & & !\hline
\end{tabular}
```

/ba/	bə	be	
[ba]		*	
[*ba]	*!		

One other subdivision of phonology, computational phonology, uses a mixture of standard phonological notation plus tree structures. As such it will be covered in the second installment of this paper.

Next time

The subfield of linguistics with perhaps the widest variety of notations is syntax. I will postpone un-

¹⁴ *The north wind and the sun were disputing which was the stronger, when a traveler came along wrapped in a warm cloak. They agreed that the one who first succeeded in making the traveler take his cloak off should be considered stronger than the other.* The text comes from the International Phonetic Association.

til installment two of this paper a discussion of the trees, matrices, and derivations, as I wish to cover them in greater detail than space or time allows at present. In particular, the macros for drawing trees have far wider application than just linguistics.¹⁵

References

- [1] Apostolos Syropoulos, Antonis Tsolomitis, and Nick Sofroniou, *Digital Typography Using LATEX*, New York: Springer, 2003.
- [2] Christina Thiele, “TEX and Linguistics”, *TUGboat* **16**(1), 42–44.
- [3] Christina Thiele, “TEX and the Humanities”, *TUGboat* **17**(4), 388–393.
- [4] Uwe Lück, “ednotes—critical edition typesetting with LATEX”, *TUGboat* **24**(2), 224–236.
- [5] David Kastrup, “The bigfoot bundle for critical editions”, Preprints for the 2004 Annual TUG Meeting, 105–110.
- [6] CTAN/macros/latex/contrib/teubner
- [7] CTAN/fonts/greek/psgreek
- [8] CTAN/fonts/utilities/fontinst-contrib/grkfinst
- [9] <http://www.braams.cistron.nl/babel/>
- [10] <http://www.opennet.ru/prog/info/1117.shtml>
- [11] <http://devnag.sarovar.org/>
- [12] Mustapha Eddahibi, Azzeddine Lazrek, and Khalid Sami, “Arabic mathematical e-documents”, Preprints for the 2004 Annual TUG Meeting, 42–47.
- [13] Azzeddine Lazrek, “CurExt, typesetting variable-sized curved symbols”, *TUGboat* **24**(3), 323–327.
- [14] CTAN/language/hebrew/makor
- [15] CTAN/language/chinese/CJK
- [16] CTAN/fonts/stmaryrd
- [17] CTAN/fonts/tipa
- [18] Geoff Pullum and William Ladusaw, *Phonetic Symbol Guide*, Chicago: University of Chicago Press, 1996.
- [19] *Handbook of the International Phonetic Association*, Cambridge: Cambridge University Press, 1999.
- [20] <http://tug.org/applications/PSTricks>
- [21] <http://contextgarden.net/MetaFun>
- [22] CTAN/macros/generic/colortab

¹⁵ As Nelson Beebe remarked at the PracTEX conference. Installment two will, I hope, serve as his requested paper.