

T_EX and Linguistics

Keywords

babel, bigfoot, colortab, devnag, edmac, ednotes, fontenc, latexsym, ledmac, makor, omega, pscyr, psgreek, pstricks, syntax, xetex, grammar, tipa, philology, phonology, lemmata, linguistics, semantics, cyrillic, teubner, unicode, arabic, armenian, chinese, devanagari, greek, hebrew, japanese, korean, russian, sanskrit

Abstract

T_EX has long been associated with mathematics and “hard” sciences such as physics. But even during the early days of T_EX, linguists were attracted to the system, and today a growing number of them are turning to T_EX (LaT_EX, ConT_EXt). Aside from the general advantages of T_EX 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 T_EX 2004 talk, I show some notational issues and their solutions in T_EX.

Why T_EX?

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

For me, the main reason to use T_EX 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 T_EX to format certain constructs aided their adoption by the field.

In this paper, I discuss various aspects of using T_EX and LaT_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 T_EX. Until its development is further along than it currently is, 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 has been painful. Your mileage may of course vary. *Digital Typography Using LaT_EX*

([1]) has a good overview of language support for Omega.

I should note that the union of T_EX and linguistics goes back quite far. For example, see the articles by Christina Thiele in *TUGboat*, [2], [3]. Donald E. Knuth² notes that linguists were among the first outside of mathematics to embrace T_EX.

The field of linguistics

Linguistics is a large field that stands at the crossroads of several other fields, but that 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 T_EXnicians is that the notational issues presented here fall essentially into either special symbols or special layouts. In order to partition the field into units we can deal with here, let us adopt a fairly traditional division that linguists often use, rather than the coarser special symbols vs. special layouts.

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

Fortunately for us, the “hyphenated” subdivisions largely make due 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 T_EX. Most of the discussion deals with various LaT_EX packages, which reflects the market share of LaT_EX. Some of what follows can be done with more or less difficulty in Plain T_EX or ConT_EXt.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} , \mathring{s} , 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 Kuryłowicz, *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 ar , al , am , an , with the same a that appears outside laryngeal environments, and this is what we often get, e.g. $\epsilon\beta\alpha\lambda\omicron\nu$ ‘they threw’ < $*erh^mE-$, $\epsilon\kappa\alpha\mu\omicron\nu$ ‘they toiled’ < $*ekm^m A-$. But where the following laryngeal was O , we generally get or , ol : $\epsilon\mu\omicron\lambda\omicron\nu$ ‘they went’, $\epsilon\theta\omicron\rho\omicron\nu$ ‘they leapt’, $\epsilon\tau\omicron\rho\epsilon$ ‘pierced’, $\epsilon\pi\omicron\rho\omicron\nu$ ‘they granted’ to the presents $\beta\lambda\omicron\sigma\kappa\omega$, $\theta\rho\acute{\omega}(i)\sigma\kappa\omega$, $\tau\iota\tau\rho\acute{\omega}\sigma\kappa\omega$ and the perfect $\pi\acute{\epsilon}\pi\tau\omega\tau\alpha\iota$. An ablaut ow : or or ow : ol seems not to occur in Greek. F. B. J. Kuiper has suggested, *India Antiqua* 199, that the development to or and ol here was phonetically regular, the laryngeal influencing the vocalization of the resonant. I doubt that this is correct, because I believe that Dor. $\pi\rho\acute{\alpha}\tau\omicron\varsigma$ represents a direct outcome of $*prO-$ (cf. §2.3.2). If pre-consonantal 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. The requirements of line numbers, cross-references, lemmata, layer upon layer of notes, marginalia, et cetera can bring a typesetter to the brink of madness, but for the edmac (Plain \TeX), ledmac and ednotes (La \TeX) packages. Unfortunately, none of these packages offers 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 ([5]) has created the bigfoot package, but I have yet to try it, so I cannot offer an opinion.

```

1  A dhuine gan chéill do mhaisligh an chléir
   b is tharcainsigh naomhscriupt na bhfáige,
   c na haitheanta réab 's an t-aifreann thréig
   d re taitheamh do chlaonchreideamh Mhártain,
   e cá rachair 'od dhíon ar Iosa Nasardha
   f nuair chaithfidimid cruinn bheith ar mhaoileann
   g Josephá?
   h Ní caraid Mac Crae chuim t'anama ' phlé
   ná Calvin bhiais taobh ris an lá sin.

2  Nách damanta an scéal don chreachaire chlaon
   b ghlac baiste na cléire 'na pháiste
   c 's do glanadh mar ghréin ón bpeaca ró-dhaor
   d trí ainibhfios Eva rinn Ádam,
   e tuitim arís fé chuing na haicme sin
   f tug atharrach brí don scríbhinn bhannaithe,
   g d'aistrigh béasa agus reachta na cléire
   h 's nách tugann aon ghéilleadh don Phápa?

3  Gach scoilaire baoth, ní mholaim a cheird
   b 'tá ag obair le géilleadh dá tháille
   c don doirbhchoin chlaon dá ngorthar Mac Crae,
   d deisceabal straeigh as an gcolláiste.
   e Tá adaithe thíos in íochtar ifrinn,
   f gan solas gan soilse i dtíorthaibh dorcha,
   g tuigsint an léinn, gach cuirpeacht déin
   h is Lucifer aosta 'na mháistir.

```

22 Teideal: Dhuinnhuing T, Seóghan Mac Domhnaill cct B

1.a dhuinne T 1.a mhaislaidh T, mhaislaig B 1.c raob T 1.d le B 1.e dod B 1.f chaithfidimid T 1.f maoilinn B 1.g phleidh T 1.h bhíos B 1.h leis B 2.a claon B 2.c glannuig T 2.d ainibhfios T, ainibhfios B 2.d Eabha B 2.g is B 2.h tuiginn T 3.a sgollaire T 3.a mholuim T 3.b 'tág coobar T 3.b re B 3.c dorbhchon daor B 3.d straothaig T 3.e fhadoghthe tsíos T 3.e fadaighthe B 3.f sollus T 3.g cuirpeacht T 3.h Lucifer T, Lúcifer 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. For 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) 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 grkfst 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 lifesaver for switching Greek fonts.

The Greek in Figure 1 was produced with PSGreek. For example, to get ‘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 aspects to typesetting languages in alphabets other than Latin. First, 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 T_EX, 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 endeavours in L^AT_EX is babel ([9]) by Johannes Braams. It is included in (I believe) all T_EX distros, the manual is comprehensive and well written, and you should spend some time familiarising yourself with it.

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.

Languages that use Indic scripts, such as Devān, agarī, 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 T_EX and L^AT_EX. Using devnag makes it possible to typeset a bilingual critical edition with essentially the same input for both the Devānagarī and the transliterated text. Figure 3 shows the vowels of Marāṭhī, typeset with the devnag package.

For languages written in the Arabic alphabet (such as Arabic, Persian, Pashto), Klaus Lagally’s ArabT_EX 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]).

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

Figure 3. The vowels of Marāṭhī

While it is possible to typeset Hebrew using ArabT_EX, Alan Hoenig’s Makor ([14]) is worth every penny.⁸

Typesetting Chinese using T_EX is possible with the CJK ([15]) package (which provides for much more than just Chinese, Japanese, and Korean support). However, I prefer ConT_EXt, 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 T_EX. So the function of the set of things similar to houses is denoted by $\lambda x \text{Similar_to}(x, \text{houses})$. The T_EX to get this is `\lambda x \mathit{Similar_to}(x, \mathit{houses})`. We had to wrap the ‘English’ inside the function with `\mathit` to prevent T_EX from interpreting the words as a series of variables. In some cases `\mbox` will work, and note that sometimes spaces inside the `\mboxes` are important. So a possible interpretation for the sentence *I have told one friend of mine all those stories*⁹ is given as $\exists x[\forall y[(x \in \text{friends of mine} \wedge y \in \text{those stories}) \rightarrow \text{I have told } y \text{ to } x]]$, or in T_EX 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 stmaryd package ([16]). So, typing `\llbracket(MN)\rrbracket^{\mathcal{M}}` yields $\llbracket(MN)\rrbracket^{\mathcal{M}}$. You may also need to load the latexsym package for a few symbols.

Phonetics and Phonology

Phonetics is a branch of acoustics that deals with speech sounds and their production and perception.

- [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://www.tug.org/applications/PSTricks>
- [21] <http://contextgarden.net/MetaFun>
- [22] CTAN/macros/generic/colortab

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