# Lexicography with TEX

JÖRGEN L. PIND

Institute of Lexicography
University of Iceland
Reykjavík
Iceland
jorgen@lexis.hi.is

## ABSTRACT

At the Institute of Lexicography at the University of Iceland, TEX is used for the typesetting of dictionaries. Currently we are in the process of bringing out a large etymological dictionary which is typeset in TEX with PostScript fonts. Details of this project are presented. The value of generic or logical coding over typographical coding is emphasized.

## 1. Background

In this paper I will discuss the use of TEX in the work carried out at the Institute of Lexicography of the University of Iceland. The Institute was founded in 1948 and has as its major aim the production of an historical dictionary of Icelandic from 1540 (when the first printed book appeared in Icelandic) up to the present, a dictionary somewhat along the lines of the Oxford English Dictionary.

During the past forty years, a lot of material has been gathered for the dictionary. The main collection of the Institute comprises some 2.5 million dictionary slips; others include, for instance a collection of words from the spoken language. These other collections contain perhaps 300,000 slips in all. Near the end of 1982, it was decided to begin evaluating the collection with the aim of publishing an historical dictionary of the language. At the same time it was decided to embark on computerizing the Institute itself.

The first computational project involved registering the main collection so as to open more paths into the collection itself. A database of all the words contained in the collection was set up. The word class, date of oldest and newest citation, the oldest source, number of citations kept in the collection and the word type (whether the word is a compound, an affixed word or a 'simple' word) were registered for each word. This database contains a total of just over 600,000 words. This is a surprisingly high figure but is explained in part by word-compounding, which is an active process in the Germanic languages, not the least in Icelandic.

In some respects, this database file can be viewed as a first approximation to a dictionary although a very primitive one, since it does not have any grammatical analysis to speak of. Yet, because the material is stored in a database (as opposed to a linear alphabetized order), it does enable us to escape from the "tyranny of the alphabet" and gives us multiple access paths to the collections of the Institute.

The editing of historical dictionaries has usually proceeded in alphabetical order, the work being brought out in installments over a period of decades. This is an approach which is in many respects less than ideal since the editor is forced to deal with words which do not form a coherent set under any reasonable linguistic criterion. We would therefore like to proceed in a different manner, dealing with individual word classes at a time. The availability of the computer makes this relatively easy to accomplish. It has now been decided by the governing board of the Institute that the editing work will concentrate on the verbs with the aim of producing an historical dictionary of verbs as the first volumes of what will hopefully later become a comprehensive historical dictionary of Icelandic.

This work was begun in 1985. The editorial strategy involves some novelties compared with traditional methods (e.g., Kuhn 1982), in that each citation is furnished with a set of editorial descriptors detailing the grammatical and semantic features of the citation itself. This is done on-line with the

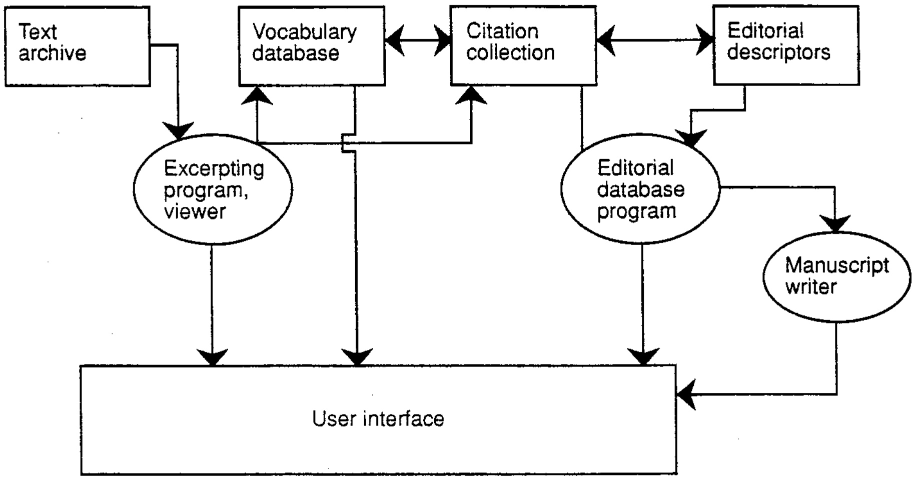


Figure 1: A model for a "lexicographer's workbench"

description being stored in a database system. The database system is then used to make TEX-encoded scripts which TEX then changes into beautifully typeset pages.

In 1986, in a talk presented at the NordData Conference in Stockholm, I outlined the approach we were using and illustrated it with a figure of a "lexicographer's workbench" (Figure 1), commenting that a number of features had not been implemented. "This holds especially for the 'manuscript writer'. Our work has not yet reached the stage where this is in great demand, but we envisage, for example, the possibility of using the database to turn out manuscripts for a typesetting program like TEX" (Pind 1986:87).

Well, this was written before we even had a version of TEX running at the Institute! As a matter of fact, though we expected that typesetting would be something that we would deal with much later, a lot of work over the past couple of years has been devoted to the typesetting side of lexicography. There are two reasons for this. The first is that the lexicographer very much wants to be able to print proofs from the lexicographic database that show some resemblance to a traditional dictionary. The second is the fact that we have been engaged in producing an Icelandic etymological dictionary working from the author's manuscript. This project will be described in detail below.

## 2. Icelandic TEX

In October 1986 I first acquired TEX. Unfortunately it was not possible for me at that time to work with Icelandic in TEX since a number of characters were missing from the Computer Modern fonts which are needed in Icelandic, such as *thorn* and *eth* (`\char'034` and `\char'037` in Figure 2). Additionally, Icelandic has a number of accented characters and, as is well known, TEX will not hyphenate words which contain floating accents. In January 1987, however, I acquired Doug Henderson's METAFONT for MS-DOS and this enabled me to get started on making Icelandic versions of the Computer Modern fonts. The first version was limited to 128 characters, due to limitations in the drivers then available. The special Icelandic characters are accessed as ligatures, as recommended by Knuth (1984:46). An Icelandic hyphenation table was made using Frank Liang's PATGEN-program. This table has turned out to perform excellently. A number of changes have also been made to the plain and LaTEX macros to accommodate Icelandic. The development of Icelandic TEX was originally carried out on an IBM PC/AT. In early 1988 we switched over to AIX on an IBM PC/RT and got Rick Simpson's excellent port of TEX and METAFONT to that machine. This has since been the platform on which we have operated.

|  | ´0 | ´1 | ´2 | ´3 | ´4 | ´5 | ´6 | ´7 |  |
|---|---|---|---|---|---|---|---|---|---|
| ´00x | Γ | Δ | Θ | Λ | Ξ | Π | Σ | Υ | "0x |
| ´01x | Φ | Ψ | Ω | á | é | í | ó | ú |  |
| ´02x | ı | ȷ | ` | ´ | ˇ | ˘ | ¯ | ˚ | "1x |
| ´03x | ¸ | ý | æ | ö | þ | Æ | Þ | ð |  |
| ´04x | Đ | ! | ” | # | $ | % | & | ’ | "2x |
| ´05x | ( | ) | * | + | , | - | . | / |  |
| ´06x | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | "3x |
| ´07x | 8 | 9 | : | ; | ø | = | Ø | ? |  |
| ´10x | @ | A | B | C | D | E | F | G | "4x |
| ´11x | H | I | J | K | L | M | N | O |  |
| ´12x | P | Q | R | S | T | U | V | W | "5x |
| ´13x | X | Y | Z | [ | “ | ] | ˆ | ˙ |  |
| ´14x | ‘ | a | b | c | d | e | f | g | "6x |
| ´15x | h | i | j | k | l | m | n | o |  |
| ´16x | p | q | r | s | t | u | v | w | "7x |
| ´17x | x | y | z | – | — | ˝ | ˜ | ¨ |  |
|  | "8 | "9 | "A | "B | "C | "D | "E | "F |  |

Figure 2: The Icelandic Font imr10

## 3. The Making of an Etymological Dictionary

I turn now to a discussion of the making of one particular dictionary, a 1250-page etymological dictionary of Icelandic which will appear later this year. The etymological dictionary is the work of one man, the late Ásgeir Blöndal Magnússon, who worked at the Institute for over forty years. When keyboarding of the text began in 1985, it was expected that the dictionary would eventually be typeset by a professional printer. Indeed, at that time we did not even have TEX at the Institute as mentioned above. However, I acquired the *TEXbook* early in 1985 and TEX did influence the keyboarding of the manuscript.

We were immediately confronted with the diverse floating accents which any etymological dictionary contains and I decided that we would adopt the TEX coding scheme as our model for the keyboarding. I now have some doubts about the suitability of this scheme as I will elaborate on later. However, it must be emphasized that I never expected that in fact the dictionary would end up being typeset with TEX.

The dictionary was keyboarded directly from the author's handwritten manuscript (having been collected on slips of the traditional kind so loved by lexicographers before the advent of computing). The PC-Write editor was used to input the manuscript, since it uses near ASCII-files and is easily configured. It was limited to 60K files but this did not cause any trouble. When the whole manuscript had been input, it amounted to 151 files containing just over 7Mb of text.

Proofreading and checking the manuscript turned out to be a major task, even more so since the author died in 1987. In December 1988, it became clear that it would be possible to publish the dictionary this year and arrangements were made with the largest printing house in Iceland to take care of the typesetting and printing. The typesetting was to be done on a Linotronic 300. At that time it turned out, however, that the typesetting process would be difficult since a lot of accents were missing from the fonts which were available. Some of these could be ordered from Linotype, others had to be made specifically at what we felt was an exorbitant price. I think I can fairly say that the printers were none too happy with the prospect of typesetting this massive book.

At this point, I decided that I would have a go at typesetting the dictionary myself with TEX. This fitted also very well into our overall strategy since we had decided that we would use TEX in the future as our typesetting engine and had, as a matter of fact, already made some experiments with the dictionary of verbs. Using the etymological dictionary as the first major test-case was of course in some respects ideal since if we could accomplish *that*, we felt we could cope with any dictionary. I should point out that this is by no means the first Icelandic book typeset with TEX. The first one was actually a book about the Macintosh personal computer written by the present author (Pind 1987) — my apologies for not having chosen a weightier subject for the occasion! A number of other books have appeared in Icelandic typeset with TEX, with more on their way. The dictionary is, however, by far the most ambitious and also the first (Icelandic) book of its kind typeset with TEX.

## 4. Some Details

Figure 3 shows a page from the dictionary printed on the Linotronic 300.

### 4.1 The Dictionary Entry

A typical article from this page is the one for *áðan*, meaning 'just now'. This is coded as follows:

```
\hword{áðan} ao. 'fyrir skömmu'; \shword{áður},
\dag\shword{áðr} ao. 'fyrr'. Sbr. fær. \wform{áðan(i)},
\wform{áður}, nno. \wform{\aa{}dan}, \wform{\aa{}der}, fd.
\wform{adens}, fsæ. \wform{aþans}, nsæ. \wform{ij\aa{}ns\/};
sk. fe. \wform{\={æ}dre}, fsax. \wform{\=adro\/} 'undir
eins', fhþ. \wform{\=atar\/} 'fljótur, skilningsskarpur'; líkl.
einnig í ætt við lettn. \wform{\~atrs} 'bráður, fljótur til'
og lith. \wform{otr\'us\/} 'ákafur'.
```

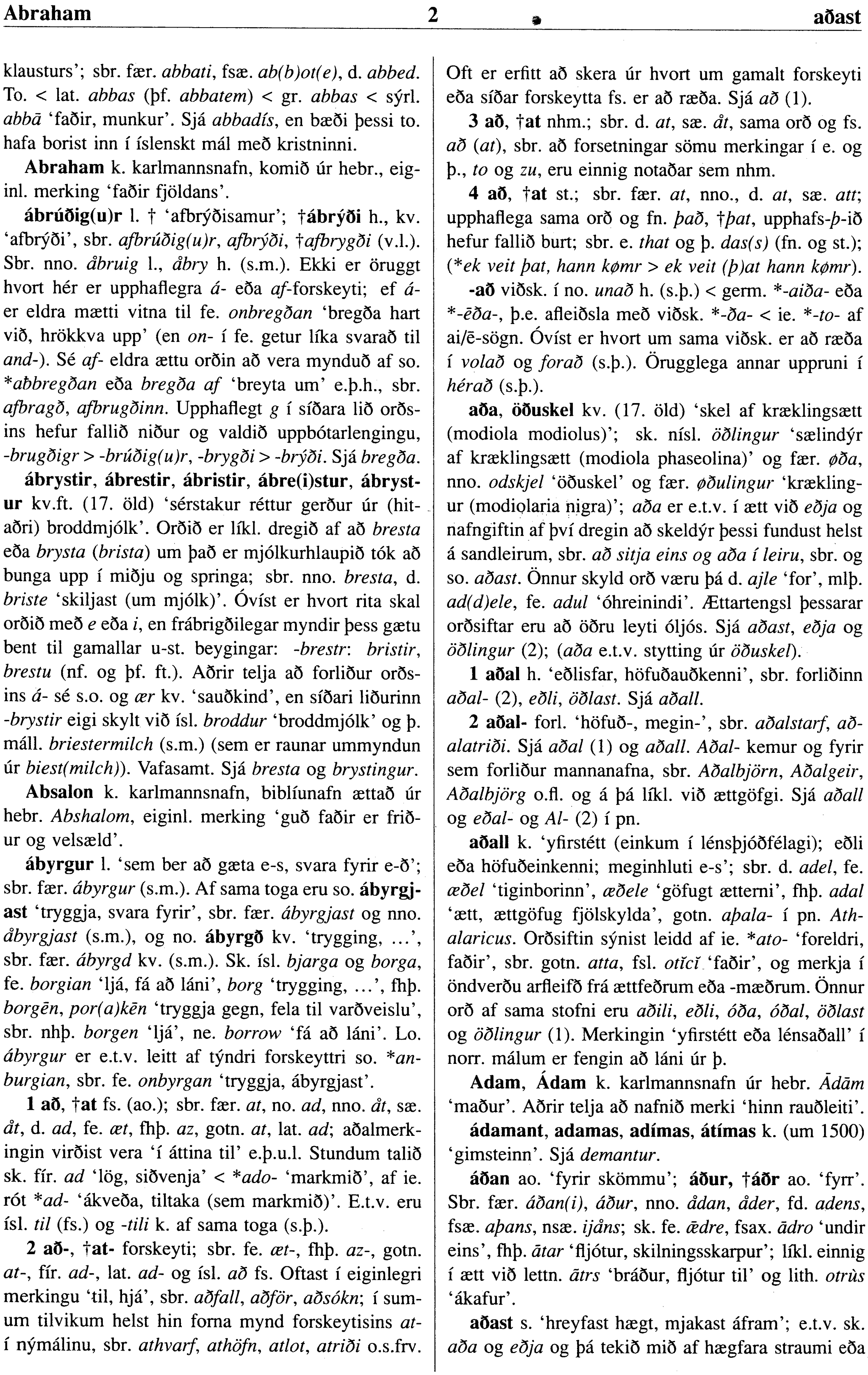
This, admittedly, doesn't look particularly nice, but the output from the Linotronic sure does and that is what counts. Each article starts with a headword which is given by the `\hword` macro. Other categories shown in the extract are `\shword` which identifies a 'subsidiary headword' and `\wform` which identifies a word, either one from a different language or one cross-referenced in the dictionary.

As can be seen in Figure 3, it is often the case that there are multiple meanings for one word, each one entered as a separate headword. These are distinguished by a decimal number in front of the word itself. These numbers are given as an optional parameter (enclosed in square brackets) for the `\hword` macro which is defined as follows:

```
\def\hword{\futurelet\PossBracket\hwordbranch}
\def\hwordbranch{\ifx\PossBracket [%
    \let\next=\hwordwithno
    \else
    \let\next=\hwordwithoutno
    \fi
    \next
}
\def\hwordwithno[#1]#2{%
    {\leavevmode\hbox to 10pt{}\bf#1 #2\mark{#2}}}
\def\hwordwithoutno#1{%
    {\leavevmode\hbox to 10pt{}\bf#1\mark{#1}}}
```

We use `\futurelet` to check for the presence of a bracket. If it is present, the macro `\hwordwithno` is executed, otherwise the macro `\hwordwithoutno` is used. Note that the indent is specified with an `\hbox`. Since this occurs at the very beginning of a paragraph, it is necessary to leave the vertical mode explicitly, using `\leavevmode`. The TEX primitive `\mark` enables us to automate the typesetting of the headwords at the top of each page, which shows the range of entries on a particular page. For this dictionary, which is set in two-column format, a version of Knuth's double-column output routine from Appendix E of *The TEXbook* has been used. These output macros make use of `\vsplit` to divide the page into two columns. The `\headline` macro is as follows:

```
\def\headline{\hbox to \pagewidth{%
    \tenbf\strut\hbox to 14pc{\firstmark\hfil}%
    \hfill\folio\hfill\hbox to 14pc{\hfil\splitbotmark}}}
```



**Abraham** 2 **aðast**

klausturs'; sbr. fær. *abbati*, fsæ. *ab(b)ot(e)*, d. *abbed*. To. < lat. *abbas* (þf. *abbatem*) < gr. *abbas* < sýrl. *abbā* 'faðir, munkur'. Sjá *abbadís*, en bæði þessi to. hafa borist inn í íslenskt mál með kristninni.

**Abraham** k. karlmannsnafn, komið úr hebr., eiginl. merking 'faðir fjöldans'.

**ábrúðig(u)r** l. † 'afbrýðisamur'; †**ábrýði** h., kv. 'afbrýði', sbr. *afbrúðig(u)r*, *afbrýði*, †*afbrygði* (v.l.). Sbr. nno. *åbruig* l., *åbry* h. (s.m.). Ekki er öruggt hvort hér er upphaflegra *á-* eða *af-*forskeyti; ef *á-* er eldra mætti vitna til fe. *onbregðan* 'bregða hart við, hrökkva upp' (en *on-* í fe. getur líka svarað til *and-*). Sé *af-* eldra ættu orðin að vera mynduð af so. *abbregðan eða *bregða af* 'breyta um' e.þ.h., sbr. *afbragð*, *afbrugðinn*. Upphaflegt *g* í síðara lið orðsins hefur fallið niður og valdið uppbótarlengingu, *-brugðigr* > *-brúðig(u)r*, *-brygði* > *-brýði*. Sjá *bregða*.

**ábrystir**, **ábrestir**, **ábristir**, **ábre(i)stur**, **ábrystur** kv.ft. (17. öld) 'sérstakur réttur gerður úr (hitaðri) broddmjólk'. Orðið er líkl. dregið af að *bresta* eða *brysta* (*brista*) um það er mjólkurhlaupið tók að bunga upp í miðju og springa; sbr. nno. *bresta*, d. *briste* 'skiljast (um mjólk)'. Óvíst er hvort rita skal orðið með *e* eða *i*, en frábrigðilegar myndir þess gætu bent til gamallar u-st. beygingar: *-brestr*: *bristir*, *brestu* (nf. og þf. ft.). Aðrir telja að forliður orðsins *á-* sé s.o. og *ær* kv. 'sauðkind', en síðari liðurinn *-brystir* eigi skylt við ísl. *broddur* 'broddmjólk' og þ. máll. *briestermilch* (s.m.) (sem er raunar ummyndun úr *biest(milch)*). Vafasamt. Sjá *bresta* og *brystingur*.

**Absalon** k. karlmannsnafn, biblíunafn ættað úr hebr. *Abshalom*, eiginl. merking 'guð faðir er friður og velsæld'.

**ábyrgur** l. 'sem ber að gæta e-s, svara fyrir e-ð'; sbr. fær. *ábyrgur* (s.m.). Af sama toga eru so. **ábyrgjast** 'tryggja, svara fyrir', sbr. fær. *ábyrgjast* og nno. *åbyrgjast* (s.m.), og no. **ábyrgð** kv. 'trygging, ...', sbr. fær. *ábyrgd* kv. (s.m.). Sk. ísl. *bjarga* og *borga*, fe. *borgian* 'ljá, fá að láni', *borg* 'trygging, ...', fhþ. *borgēn*, *por(a)kēn* 'tryggja gegn, fela til varðveislu', sbr. nhþ. *borgen* 'ljá', ne. *borrow* 'fá að láni'. Lo. *ábyrgur* er e.t.v. leitt af týndri forskeyttri so. *anburgian, sbr. fe. *onbyrgan* 'tryggja, ábyrgjast'.

**1 að**, †**at** fs. (ao.); sbr. fær. *at*, no. *ad*, nno. *åt*, sæ. *åt*, d. *ad*, fe. *æt*, fhþ. *az*, gotn. *at*, lat. *ad*; aðalmerkingin virðist vera 'í áttina til' e.þ.u.l. Stundum talið sk. fír. *ad* 'lög, siðvenja' < *ado- 'markmið', af ie. rót *ad- 'ákveða, tiltaka (sem markmið)'. E.t.v. eru ísl. *til* (fs.) og *-tili* k. af sama toga (s.þ.).

**2 að-**, †**at-** forskeyti; sbr. fe. *æt-*, fhþ. *az-*, gotn. *at-*, fír. *ad-*, lat. *ad-* og ísl. *að* fs. Oftast í eiginlegri merkingu 'til, hjá', sbr. *aðfall*, *aðför*, *aðsókn*; í sumum tilvikum helst hin forna mynd forskeytisins *at-* í nýmálinu, sbr. *athvarf*, *athöfn*, *atlot*, *atriði* o.s.frv. Oft er erfitt að skera úr hvort um gamalt forskeyti eða síðar forskeytta fs. er að ræða. Sjá *að* (1).

**3 að**, †**at** nhm.; sbr. d. *at*, sæ. *åt*, sama orð og fs. *að* (*at*), sbr. að forsetningar sömu merkingar í e. og þ., *to* og *zu*, eru einnig notaðar sem nhm.

**4 að**, †**at** st.; sbr. fær. *at*, nno., d. *at*, sæ. *att*; upphaflega sama orð og fn. *það*, †*þat*, upphafs-*þ*-ið hefur fallið burt; sbr. e. *that* og þ. *das(s)* (fn. og st.); (*ek veit þat, hann kømr > *ek veit (þ)at hann kømr*).

**-að** viðsk. í no. *unað* h. (s.þ.) < germ. *-aiða- eða *-ēða-, þ.e. afleiðsla með viðsk. *-ða- < ie. *-to- af ai/ē-sögn. Óvíst er hvort um sama viðsk. er að ræða í *volað* og *forað* (s.þ.). Örugglega annar uppruni í *hérað* (s.þ.).

**aða**, **öðuskel** kv. (17. öld) 'skel af kræklingsætt (modiola modiolus)'; sk. nísl. *öðlingur* 'sælindýr af kræklingsætt (modiola phaseolina)' og fær. *øða*, nno. *odskjel* 'öðuskel' og fær. *øðulingur* 'kræklingur (modiolaria nigra)'; *aða* er e.t.v. í ætt við *eðja* og nafngiftin af því dregin að skeldýr þessi fundust helst á sandleirum, sbr. *að sitja eins og aða í leiru*, sbr. og so. *aðast*. Önnur skyld orð væru þá d. *ajle* 'for', mlþ. *ad(d)ele*, fe. *adul* 'óhreinindi'. Ættartengsl þessarar orðsiftar eru að öðru leyti óljós. Sjá *aðast*, *eðja* og *öðlingur* (2); (*aða* e.t.v. stytting úr *öðuskel*).

**1 aðal** h. 'eðlisfar, höfuðauðkenni', sbr. forliðinn *aðal-* (2), *eðli*, *öðlast*. Sjá *aðall*.

**2 aðal-** forl. 'höfuð-, megin-', sbr. *aðalstarf*, *aðalatriði*. Sjá *aðal* (1) og *aðall*. *Aðal-* kemur og fyrir sem forliður mannanafna, sbr. *Aðalbjörn*, *Aðalgeir*, *Aðalbjörg* o.fl. og á þá líkl. við ættgöfgi. Sjá *aðall* og *eðal-* og *Al-* (2) í pn.

**aðall** k. 'yfirstétt (einkum í lénsþjóðfélagi); eðli eða höfuðeinkenni; meginhluti e-s'; sbr. d. *adel*, fe. *æðel* 'tiginborinn', *æðele* 'göfugt ætterni', fhþ. *adal* 'ætt, ættgöfug fjölskylda', gotn. *aþala-* í pn. *Athalaricus*. Orðsiftin sýnist leidd af ie. *ato- 'foreldri, faðir', sbr. gotn. *atta*, fsl. *otĭcĭ* 'faðir', og merkja í öndverðu arfleifð frá ættfeðrum eða -mæðrum. Önnur orð af sama stofni eru *aðili*, *eðli*, *óða*, *óðal*, *öðlast* og *öðlingur* (1). Merkingin 'yfirstétt eða lénsaðall' í norr. málum er fengin að láni úr þ.

**Adam**, **Ádam** k. karlmannsnafn úr hebr. *Ādām* 'maður'. Aðrir telja að nafnið merki 'hinn rauðleiti'.

**ádamant**, **adamas**, **adímas**, **átímas** k. (um 1500) 'gimsteinn'. Sjá *demantur*.

**áðan** ao. 'fyrir skömmu'; **áður,** †**áðr** ao. 'fyrr'. Sbr. fær. *áðan(i)*, *áður*, nno. *ådan*, *åder*, fd. *adens*, fsæ. *aþans*, nsæ. *ijåns*; sk. fe. *ǣdre*, fsax. *ādro* 'undir eins', fhþ. *ātar* 'fljótur, skilningsskarpur'; líkl. einnig í ætt við lettn. *ātrs* 'bráður, fljótur til' og lith. *otrùs* 'ákafur'.

**aðast** s. 'hreyfast hægt, mjakast áfram'; e.t.v. sk. *aða* og *eðja* og þá tekið mið af hægfara straumi eða

Figure 3: A sample page from the etymological dictionary

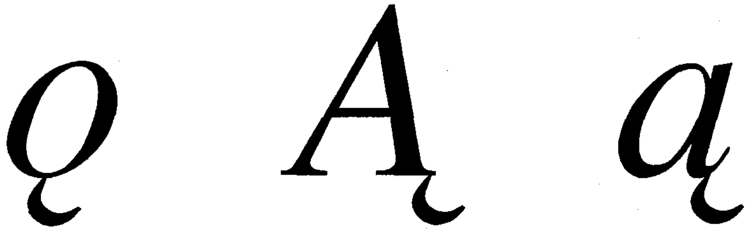


Figure 4: The placement of the ogonek accent under different letters

The `\firstmark` token marks the first headword on each page and the `\splitbotmark` marks the last headword on the page.

### 4.2 Accents

In an etymological dictionary, there are *n* different accents which need to be taken care of. TEX has an `\accent` primitive which positions accents over letters. Unfortunately the `\accent` primitive is limited to putting one accent over a letter. Often there is a need to put two accents over a single letter. For this it is necessary to write special macros.

One subtlety which TEX does not address directly is the fact that it is not always possible to position accents without knowing what letter it is put over. This holds, for example, for the acute accent over a **k**. Ordinarily, the accent primitive works fine for positioning an acute accent over letters (see e.g., **á**, **ś**), but when it comes to the **k**, the accent should not be positioned as in **ḱ**, but rather as in **ḱ**. A similar issue arises in regard to the positioning of the ogonek accent under letters. Thus it is usually placed under the middle of an **o** but at the right serif of the **A** and **a** (Figure 4).

These facts bring up the question of coding. The character set of the Computer Modern fonts is really quite limited since it only uses 128 character positions out of the 256 possible. Adobe PostScript fonts have many more characters (around 300) and the Bitstream fonts even more. In particular, many accented characters are part of the standard Adobe and Bitstream fonts and it would seem natural to use those. This makes it necessary to define the characters in a manner similar to that adopted by Knuth for the mathematical symbols. Thus we would have control sequences such as `\aacute`, `\oogonek`, `\ubreve`, etc. In fact, in many ways this is probably a better choice than using the accent coding. In our case, for instance, where we want to be able to use both the Adobe fonts and Knuth's Computer Modern — changing between them with a simple switch — a fragment of our font coding macros runs on along the following lines:

```
\newif\ifcmfonts
...
\ifcmfonts
 \message{*** Computer Modern fonts used ***}
  \font\tenrm=imr10  % im... Icelandic versions of Computer Modern
  \font\ninerm=imr9
  \font\eightrm=imr8
...
  \def\aa{\accent23a}
  \def\L{\leavevmode\setbox0\hbox{L}\hbox to\wd0{\hss\char32L}}
...
\else
  \message{*** PostScript fonts used ***}
  \font\tenrm=Timesro at 10pt
  \font\ninerm=Timesro at 9pt
  \font\eightrm=Timesro at 8pt
...
  \def\aa{\char7} % Yes, our encoding is a bit peculiar!
  \def\L{\char3}
...
\fi
```

This makes it possible for us to define the characters without regard for the particular fonts we are using. Here a generic or logical approach to the coding of characters is adopted rather than the typographically-oriented coding of the *TEXbook*. The TEX coding scheme is natural for use with the Computer Modern fonts. When extending TEX to use other fonts it becomes less than ideal and thus it is appropriate to adopt another coding scheme which is not font-based.[1]

At one point I attempted to use the Bitstream fonts, but they turned out to be useless for our purposes since the implementation of the Bitstream-to-TEX conversion program from Personal TEX makes it impossible to freely arrange characters in the fonts. Thus it is not possible to use the Icelandic hyphenation table if using the Bitstream fonts.

Sometimes the plain macros are not up to the typesetting of floating accents. There is no TEX primitive which puts accents underneath letters. Knuth has defined some macros in plain which are used for this. To put a dot underneath a letter the macro \d is used and the macro \b is used to put a bar underneath a letter. The former macro is defined in plain.tex as:

```
\def\d#1{\oalign{#1\crcr\hidewidth.\hidewidth}}
```

These macros work nicely for putting a dot and a bar underneath straight l̲ẹtters but are not adequate for *italic l̲ẹtters.* Re-definition of these macros along the following lines makes it possible to use these macros both for straight l̲ẹtters and *italic l̲ẹtters.*

```
\def\d#1{\ifnum\fam=\itfam
  \oalign{#1\crcr\hidewidth\kern-0.1em.\hidewidth}%
\else
  \oalign{#1\crcr\hidewidth.\hidewidth}\fi}
\def\b#1{\ifnum\fam=\itfam\oalign{#1\crcr\hidewidth%
    \kern-.3em\vbox to.2ex{\hbox{\char22}\vss}\hidewidth}%
    \else\oalign{#1\crcr\hidewidth%
    \vbox to.2ex{\hbox{\char22}\vss}\hidewidth}\fi}
```

These macros only work for the italic family. If they are to be extended to slanted letters, it will be necessary to introduce yet another conditional testing for membership in the \slfam.

TEX is not able to put two accents over a single letter. These are quite frequent in the etymological dictionary and, unfortunately, always occur in the italic font. After having tried some fiddling around with kerns and such things, which did not produce what I felt were adequate results, I looked at the definition of the \accent primitive in the listing for the TEX program (Knuth 1986:462–463). The positioning of accents is "straightforward but tedious" according to Knuth. And further:

> Given an accent of width $a$, designed for characters of height $x$ and slant $s$; and given a character of width $w$, height $h$, and slant $t$: We will shift the accent down by $x - h$, and we will insert kern nodes that have the effect of centering the accent over the character and shifting the accent to the right by $\delta = \frac{1}{2}(w - a) + h \cdot t - x \cdot s$.

Well, I thought, this is what I need for the positioning of double accents. And so I decided to implement this in a TEX macro called \dblacc. The idea is to first put one accent over a letter and then use Knuth's formula as if it were a single character needing one accent. The macro \dblacc needs to play with a number of variables. The names of these variables are similar to the ones Knuth uses in the equation above:

```
\newdimen\xheight % the accents are designed for the x-height
{\it\xheight=\fontdimen5\the\font} % x-height for the italic font
\newdimen\Shift \newdimen\A
```

[1] I should note that I have not implemented this scheme completely and letters which can only be set with a floating accent (e.g., doubly accented letters) are still coded with an accent-based coding.

```
\newdimen\X \newdimen\W \newdimen\H \newdimen\HT
\newdimen\XS \newcount\slant \newdimen\lk \newdimen\rk
```

The macro itself is defined as follows:

```
\def\dblacc#1#2#3{\leavevmode\setbox1=\hbox{#2#3}%
%   #1 topmost accent, #2 first acc, #3 character
\H=\ht1\W=\wd1\setbox0=\hbox{#1}\A=\wd0
\ifnum\fam=\itfam\slant=4
  \HT=\H\divide\HT by \slant
  \XS=\xheight\divide \XS  by \slant
\else\slant=0
  \HT=\H\multiply\HT by \slant
  \XS=\xheight\multiply\XS by \slant
\fi
\lk=\W\advance\lk by -\A\divide\lk by 2\advance \lk by \HT
\advance \lk by -\XS\rk=\A\advance\rk by \lk
\Shift=\xheight\advance\Shift by -\H
\kern\lk\lower\Shift\hbox{#1}\kern-\rk\unhbox1\relax}
```

TeX is only able to handle integer arithmetic. In the TeX program, the slant parameter of the font is used to position the accent. Here a brute force approach is used and the `\slant` is set to 4, which is then used in division to equal multiplication by 0.25, which is the value of the slant parameter both in `cmti9` and Times-Italic.

Now, I must admit that this does look rather complicated and I felt that a simpler method could be found. I was aware of Peter Olivier's macros for setting double accents (Olivier 1988) but these do not work for the italic fonts. After having made the `\dblacc` macro, I saw Christina Thiele's macro `\diatop` (Thiele 1987). This macro does a pretty good job but does not assign correct width to the overall construction for doubly-accented letters and so is not suitable for running text (this could probably be easily fixed). However, I did run a small experiment timing the `\dblacc` and `\diatop` macros. After ascertaining that the former runs faster I sort of lost interest in redefining the macro! Anyway, the `\dblacc` macro has performed pretty well in this project and so there has not been a pressing need to change to something which perhaps is somewhat simpler.

### 4.3 PostScript

The book is typeset with TeX using PostScript to drive the typesetter. This approach was taken so that it would be possible to get high-resolution typesetting on the Linotronic 300. Using PostScript also demanded the use of Adobe typefaces since we did not have the Computer Modern faces for use with PostScript at the high resolution offered by the Linotype machine. Even so I doubt that we would have used the Computer Modern faces since they are not particularly well suited to typesetting in narrow columns. The lowercase alphabet length of cmr9 is 118.0124 pt while the corresponding figure for Times-Roman at 9 points is 107.48698 pt (the dictionary is set using fonts at 9 pt). For the bold fonts there is an even greater difference in that cmbx9 has a lowercase alphabet length of 136.1019 pt while Times-Bold is 114.50696 pt.

This difference of length shows up clearly in the number of `Overfull \hbox` messages when typesetting with these fonts. TeX has a parameter called `\tolerance` which enables the user to specify how much glue is allowed to stretch and shrink between words. Normally `plain` TeX sets `\tolerance` equal to 200 but this, as pointed out by Knuth (1984:28–29, 96), is much too strict for narrow column setting.

The following table shows some experimental runs with `\tolerance` equal to 500, 1000 and 5000 using either Adobe fonts (Times-Roman, Times-Italic, and Times-Bold) at 9 points or Computer Modern fonts (`cmr9`, `cmbx9` and `cmti9`). These experimental runs were done on the articles comprising the letter `s` totaling 4,318 paragraphs and running to 240 pages. It is evident that the number of over- and underfull boxes is dependent on both the fonts used and the setting of the `\tolerance` parameter. From these experiments it was decided to set the `\tolerance` to 1000 during the processing of the book.

| Fonts | Box type | \tolerance | | |
|---|---|---|---|---|
| | | 500 | 1000 | 5000 |
| Adobe | overfull | 397 | 309 | 22 |
| Times | underfull | 1 | 2 | 177 |
| Computer | overfull | 1037 | 463 | 71 |
| Modern | underfull | 3 | 3 | 421 |

Another approach, which I feel would be worth trying, would be to make a special version of Computer Modern designed for narrow settings. Knuth (1989) has recently given a fascinating example of the way he changed the parameters of the Computer Modern fonts for the Concrete fonts. Something similar can no doubt be done to make the fonts suitable for narrow columns.

The typesetting process used standard TeX (with one exception) running with an Icelandic hyphenation table. The change from standard TeX relates to the hyphenation where one change was made to the TeX code. In section 902 of the program listing (1986:380), Knuth declares that "TeX will never insert a hyphen that has fewer than two letters before it or fewer than three after it". This is less than ideal for Icelandic where it is very common to hyphenate before the second last letter of a word. So slight changes were made to the code to accomplish this. Otherwise, the TeX program is standard. In particular, I don't think using Multilingual TeX would have been to any advantage in this case since dozens of languages and dialects are referenced in the dictionary, many of them extinct and no doubt getting hold of hyphenation patterns for these would have been pretty difficult![2] Correcting the overfull boxes was therefore done by hand. This, however, did not turn out to be a particularly onerous task.

Using PostScript with TeX is really quite straightforward. It is of course necessary to supply the requisite `tfm` files. These can be rather easily generated from the AFM files provided with the Adobe fonts. For this I used the `aftopl` program on the UNIX TeX distribution and changed it so that it would recognize the font encoding I had adopted. The `aftopl` program makes `pl` files which can then be changed to `tfm` files with the `pltotf` program. I have used ArborText's `dvips` driver to generate the PostScript code from the `dvi` files. This has all worked quite satisfactorily.

The only thing which leaves something to be desired is the possibility for previewing the typeset pages. The IBM RT has a screen with a resolution of 118 points to the inch. It has an excellent and very fast previewer, enabling the user to jump to any page in a 100-page section of the dictionary almost instantaneously. No standard Adobe files exist for this resolution. So I tried making some up using the Bitstream fonts, e.g., Dutch for Times-Roman. It turned out of course that one foundry's Times-Roman is not another's. The widths of the characters are not comparable so what should be a nicely justified text comes out quite ragged on the screen. This should come as no surprise and was presumably one of the main motives behind Knuth's development of the Computer Modern family, namely the need to generate a consistent set of fonts for use on devices with very different resolutions.

However, though I see a need for bringing up a set of correct PostScript screen fonts, the need is not pressing since the preview is only used to check for widow lines, overfull boxes and such things.

Though PostScript has a reasonable character set, some characters are missing which are needed in this project. These I have made up using the Fontographer font editor, a PostScript font editor running on a Macintosh. Fontographer is quite different from METAFONT. A character is made by drawing curves and lines on the Macintosh screen. Fontographer uses Bézier curves like METAFONT but it has no understanding of "meta-ness", so each character has to be drawn on its own with the user attending to the overall aspects of the design. The output of Fontographer is a PostScript file which can either be downloaded to the printer or prepended to the PostScript file containing the text of the dictionary itself. Actually, only the latter approach seems to work on the Linotronic.

All things considered, I would like to stress that using PostScript fonts with TeX has turned out to be much easier than I had at first imagined. It is of course true that PostScript has some limitations in that it always operates from a single design size. This leads for instance to small caps letters which are obviously of a lower quality than a specially designed small caps font. Also, the quality of the

[2] With a purely bilingual dictionary the advantages of using Multilingual TeX are obvious.

letters at small point sizes leaves something to be desired, but this is not a problem for the setting of a dictionary which almost exclusively uses 9pt fonts. Those setting mathematics are of course aware of the limitations of the PostScript fonts for mathematics.

## 5. Other Projects

I have already mentioned the dictionary of verbs which will be the major work undertaken over the next few years. Considerable time has been spent on the database side of this project (which is now being ported from MS-DOS to UNIX), and also on the typesetting aspects. We have also embarked on a study of older Icelandic dictionaries. Some of these will be republished by the Institute, freshly typeset using TEX. The first three volumes are now underway: an Icelandic-Latin dictionary from 1683, an Icelandic-Danish-Latin dictionary from 1814, and a Danish-Icelandic dictionary from 1819.

The latter dictionary shows off some of TEX's capabilities quite nicely. The dictionary is Danish-Icelandic, although we are primarily interested in the Icelandic vocabulary. A list of all the Icelandic words (with reference to the appropriate headword) will be included with the book. TEX automatically writes these words (which have been specifically marked in the dictionary) to a file and a special program then takes care of sorting and merging these entries which are then input to TEX again for typesetting.

## 6. Some Lessons

I don't think it will be necessary to explain to this audience why we have found TEX to be eminently suitable for lexicographic work. The typesetting is unquestionably of the highest order. Our experiences with TEX over the last couple of years have taught us many lessons. The most important of these is perhaps the following: When coding a manuscript, always code it at the most abstract level possible. This was not our approach when we embarked on the etymological dictionary. This was partly due to the fact that we were preparing a file for a typesetter. The virtues of logical or generic coding are many, as pointed out by Lamport (1988) for example, and Knuth (1989:31–32) has an interesting example of this relating to the different use of text numerals and mathematical numerals. The value of logical coding is apparent in the making of dictionaries where we are dealing with text which is relatively highly structured. By using logical coding it is relatively straightforward to use the same manuscript for typesetting as for input to a database system. This of course, is one of the ideas behind the SGML standard. TEX by itself does not force any particular style of coding on the user, but it does enable the use of generic coding, and I feel that this should be used to the fullest extent possible, especially perhaps in lexicographic work, where it would considerably ease the process of putting printed dictionaries on-line (Alshawi, Boguraev and Carter 1989). Of course, when it comes to the actual, final typesetting, it is not possible to completely bypass typographical coding. Thus, in order to get rid of widow lines and other such typographical blemishes, it is necessary to use purely typographic command such as `\looseness`. These commands are, however, few and can be easily isolated.

One advantage of this approach is that by relatively simple re-definitions of macros, it is possible to print completely different proofs of the same text, with cross-references or grammatical information highlighted in special ways. This has been tried and found to be highly useful.

This approach has now been consistently adopted for other works now being coded in TEX at the Institute. This holds for the historical dictionary of verbs mentioned earlier, as well as the series of reprints of older dictionaries. My aim, through these diverse types of dictionaries, is to produce a reasonably comprehensive macro package for the typesetting of dictionaries, a package which will enable the user to describe the logical structure of the text while TEX takes care of the formatting.

## Bibliography

Alshawi, Hiyan, Bran Boguraev, and David Carter. "Placing the Dictionary On-Line." Pp. 41–63 in *Computational Lexicography for Natural Language Processing*. Bran Boguraev and Ted Briscoe, eds. London: Longman, 1989.

Knuth, Donald E. *The TEXbook*. Reading, Mass.: Addison-Wesley, 1984.

Knuth, Donald E. *TEX: The Program. Computers and Typesetting*, Vol. B. Reading, Mass.: Addison-Wesley, 1986.

Knuth, Donald E. "Typesetting Concrete Mathematics." *TUGboat* 10:31–36, 1988.

Kuhn, Sherman. "On the Making of the Middle English Dictionary." *Dictionaries: Journal of the Dictionary Society of North America* 4:14–41, 1982.

Lamport, Leslie. "Document Production: Visual or Logical." *TUGboat* 9:8–10, 1988.

Olivier, Peter J. "Publishing 'Exotic' Documents with EXOTEX, A New Macro Package." Paper presented at TEX88, Exeter University, July 1988.

Pind, Jörgen, "The Computer Meets the Historical Dictionary." *Nordisk DATAnytt* 16(10):41–43, 1986.

Pind, Jörgen. *Bókin um Macintosh*. Reykjavík: Mál og menning, 1987.

Thiele, Christina. "TEX, Linguistics, and Journal Production." Pp. 5–26 in *Conference Proceedings, Eighth Annual Meeting of the TEX Users Group*. Dean Guenther, ed. *TEXniques* 5. Providence: TEX Users Group, 1988.

## TeX Users Group

**Stanford University, August 13–24, 1984**
**Terman Engineering Center Auditorium and The Graduate School of Business**