Stretching the Limits of Babel, an Ever Growing Package?

Johannes L. Braams
TEXniek

1 Introduction

This presentation starts with an overview of the history of babel. The current state of babel is described and some thoughts on the problem areas that need to be dealt with are presented. In the last part of the presentation I will discuss possible future directions of development.

Stretching the Limits of babel — Overview

1. Introduction
2. A brief history of babel
3. The current state of babel
4. (Un)solved problems
5. Future development

March 1997 — JLB — 2

2 A brief history of babel

The first ideas for developing a set of macros to support typesetting documents with TEX in languages other than English developed around the time of the EuroTEX conference in Karlsruhe (1989). Back then I had created support for typesetting in Dutch by stealing `german.tex` (by Hubert Partl c.s.) and modifying it for Dutch conventions. This worked, but I was not completely satisfied as I hate duplication of code. Soon after that I found that more 'copies' of `german.tex` existed to support other languages. This led me to the idea of creating a package that combines these kind of language support packages. It would have to consist of at least two 'layers': all the code the various copies of `german.tex` had in common in one place, loaded only once by TEX, and a set of files with the code needed to support language specific needs. During the Karlsruhe conference the name 'babel' came up in discussions I had. It seemed an appropriate name and I sticked to it.

Stretching the Limits of babel — A brief history of babel

A brief history of babel

- First ideas at EuroTEX'89 Karlsruhe
- First published in TUGboat 12–2
- Update article in TUGboat 14–1
- Presentation at EuroTEX'95

March 1997 — JLB — 3

After the conference I started to work on "babel, a multilingual style-option system for use with LATEX's standard document styles". The first release with support for about half a dozen languages appeared in the first half of 1990. In TUGboat volume 12 number 2 an article appeared describing babel. Soon thereafter people started contributing translations for the 'standard terms' for languages not yet present in babel. The next big update appeared in 1992, accompanied by an article in TUGboat volume 14 number 1. The main new features were that an interface was added to 'push' and 'pop' macro definitions and values of registers. Also some code was moved from language files to the core of babel. In 1994 some changes were needed to get babel to work with LATEX2ε. As it turned out a lot of problems were still unsolved, amongst which the incompatibility between babel and the use of T1 encoded fonts was most important.

In 1995 the concept of 'shorthands' was introduced. A 'shorthand' is a construction based on active characters, but the active character changes is definition according to its context. It can have an argument and it can get a definition on user level, language level or system level.

In 1996 support for languages that need fonts with a different encoding was (re)introduced. Before the advent of LATEX2ε babel contained support for typesetting Russian texts, using cyrillic fonts. This support had to be completely rewritten.

3 The current state of babel

3.1 Languages supported

Currently babel supports no less than 36 languages. The level of support for the various languages varies. For some languages the support is nothing more than a provision of the translation of (most) words that can be generated by LATEX. For other languages shorthands are defined to ease

the typing of texts or to support certain hyphenation tricks. For some languages a fontencoding switch or specific typographic conventions need to be supported. The support for the Greek language also provides a different enumerating scheme (`\greeknumerals`).

For a number of languages multiple variants are supported.

The languages directly supported by babel are shown in the following slide.

Stretching the Limits of babel — The languages supported by babel

The languages supported by babel

Afrikaans	English	Irish	Sanskrit
Bahasa	Esperanto	Italian	Scottish
Basque	Estonian	Kannada	Spanish
Breton	Finnish	Lower Sorbian	Slovakian
Catalan	French	Norwegian	Slovene
Czech	Galician	Polish	Swedish
Croatian	German	Portuguese	Turkish
Danish	Greek	Rumanian	Upper Sorbian
Dutch	Hungarian	Russian	Welsh

March 1997 JLB 4

Apart from these languages two separate distributions are known to exist that are based on babel and provide support for the Ethiopian and Ukrainian languages.

3.2 Language attributes supported

At the EuroTEX 95 conference a babel BOF was held. The discussion focused on the topic of what defines a language. n the end a list of language attributes was produced.

Stretching the Limits of babel — Language attributes

Language attributes

1. Hyphenation patterns and associated `\lefthyphenmin` and `\righthyphenmin`.
2. Captions and dates
3. Quotation marks
4. Typographic conventions

March 1997 JLB 5

For attribute 2 it was thought that perhaps several formats of dates might need to be supported for some languages.

Stretching the Limits of babel — Language attributes (ctd)

3

1. Fontencoding
2. Mathematics
3. Enumerating
4. Ligatures
5. Punctuation

March 1997 JLB 6

The font encoding is also referred to as output encoding. An example of the language (or nationality) dependency of mathematics (2) is that `\tan` needs to produce either *tan* or *tg*.

Stretching the Limits of babel — Language attributes (ctd)

4

1. Input encoding
2. Direction of writing
3. Hyphen split
4. Collating sequence (`\alph` etc.)
5. Conventions for emphasis

March 1997 JLB 7

Jiří Zlatuška has published an article about 'hyphen split' which I couldn't trace. The conventions for emphasis might possibly be better placed in a document class which implements publishing house conventions.

For a large number of the attributes above, examples can be found in babel. The attributes 1 and 2 were the very basis of the system and are supported for all the languages in the babel distribution. For a large number of languages some support is available for non-standard quotation marks (3) and specific typographic conventions (4).

The attributes 1 through 4 are less common, but do occur for some languages. The attributes 4 through 5 do not currently occur in any language definition file. People have been trying to get support for Hebrew typesetting working over the past couple of years. For this they need attribute 2 to be supported. The work that I am aware of sofar has shown that bidirectional typesetting needs extensive changes in LaTeX itself which can not easily be done from the outside of LaTeX.

3.3 Document elements supported

In a document various elements can be identified which should possibly inherit the language attribute.

Stretching the Limits of babel — Document elements

Document elements

- the main text
- table of contents (and of figures, tables)
- running headers (and footers)
- floating objects (marginpars, figures, tables)

March 1997 JLB 8

In the original babel system only the main body of the text would be influenced by the setting of a language switch. During the recent history of babel the tables of contents etc. and the running headers have been added. They now inherit the language attribute which is valid at the time an

entry in the table of contents is generated. In the process of adding this support the setting of the language attribute has also been added to LaTeX's auxiliary files.

3.4 Shorthands

From the start of babel some language definition files have contained code to make some characters 'active'. In the early years this only happened to the double quote ("), which was a rather safe choice as both Don Knuth and Leslie Lamport had stated in their books that it should not be used in texts. Nevertheless this active character caused problems as it also has a function to indicate to TeX that a hexadecimal number follows.

With babel release 3.5 in 1995 the concept of 'shorthands' was introduced. A shorthand is basically an active character, possibly followed by a second character.

Stretching the Limits of babel shorthands

shorthands

- A shorthand consists of an active character, possibly followed by an argument
- Shorthand characters do not change `\catcode`
- Shorthand characters are written out unexpanded

March 1997 JLB 9

The difference with earlier releases of babel was that from then on active characters remain active throughout the document. They do not change `\catcode` other then in controlled situations, such as a verbatim environment. The only thing that changes is their definition.

When necessary, shorthand characters are made to expand to a non-active copy of themselves. Another aspect of shorthands is that when they are written out (to an `.aux` file for instance) they do not get expanded.

Currently quite a large number of characters are used as shorthand characters, as can be seen in the following slide.

Stretching the Limits of babel Overview of shorthands

~ System, Basque, Catalan, Estonian, Galician, Sanskrit, Spanish
: Breton, French, Russian, Turkish
; Breton, French, Russian
! Breton, French, Russian, Turkish
? Breton, French, Russian
" Basque, Catalan, Danish, Dutch, Estonian, Finnish, Galician, German, Polish, Portuguese, Russian, Sanskrit, Slovene, Spanish, Swedish, Upper Sorbian
‘ Catalan (optional)
’ Catalan, Galician, Spanish (optional)
^ Esperanto
= Turkish
_ Sanskrit

March 1997 JLB 10

4 Unsolved problems

Stretching the Limits of babel Unsolved problems

Unsolved problems

1. floats inheriting language attributes
2. bidirectional typesetting
3. Multiple input encodings and hyphenation
4. Non standard input encoding

March 1997 JLB 11

In the current babel a number of problems remain unsolved. The most important ones are shown in the slide above.

Problem 1 still needs to be researched. It seems obvious that a floating object should inherit the language attribute from the 'surrounding text'. This probably means that a way has to be found to pass this information to the floating object.

On problem 2 Rama Porrat in Israel has done quite a lot of work. As far as I know she didn't quite succeed in creating a `hebrew.sty` though. She did find out however that LaTeX needs to be changed in a number of points to fully support right-to-left typesetting. LaTeX is by nature a package which was developped in an environment where virtually everybody uses solely left-to-right typesetting.

With multilingual documents the authors will no doubt find out that the hyphenation is done by TeX on a paragraph basis. This means that the hyphenation algortihm uses the `\lccodes` which are in effect at the end of the paragraph for the hyphenation process of the entire paragraph. In some cases this might leed to wrong hyphenation of single words or phrases in a different language for which other `\lccodes` are needed.

With LaTeX2$_\varepsilon$ the (still experimental) package `inputenc` came along. This package makes all the 'special' characters active and defines them to expand to LaTeX's 'internal' encoding. For some languages people have claimed that for them the 'special' characters need to be of the category code 'letter'.

5 Future development

Possible future development

- Extending the support with more languages?
- Integrating the Ω version of babel with the 'normal' babel distribution
- Integrating concepts of babel in LaTeX?

As people keep sending me contributions for 'new' languages the number of languages supported by babel will keep growing. Perhaps the support for some languages will be distributed seperately from the core babel distribution (which would ease the task of maintenance of babel somewhat).

Currently a special version of babel has been developed by Yannis Haralambous for multilingual support with Ω. Both versions of babel should be reunited for the sake of easier maintenance.

The experience gained with the development and maintenance of babel over the past years is very valuable in the development of the language support module for LaTeX3. Some of the *concepts* that have been developped for babel will form input for the development of this part of LaTeX3.