The Marriage of TEX and Lojban

Hong Feng
Suite 3-3, 200 WuZhong Str.
Wuhan, Hubei Province
430040 China P.R.
hongfeng@gnu.org

Abstract

Lojban is an old artificial language which is ambiguity free, it can also be used as a tool to express Chinese text encoding in readable ASCII characters, and thus be applied in TEX typesetting system. Keywords: TEX, Lojban, Chinese.

When Prof. Donald Knuth invented the TEX system, Chinese was not considered as the default language to support, as TEX only accepts the readable 7-bit ASCII characters. Chinese, either the simplified, or the traditional encoding set, has many thousands of characters. For example, the GB2312-80 contains 6,763 simplified Chinese characters, which requires at least double-byte (16-bit) to represent one Chinese character (the encodings in the double-byte format are unreadable for people unless one checks the encoding table one by one!), and makes it difficult for TEX to typeset Chinese documents.

Various scenarios were presented to support Chinese and many relevant macros were developed in the past, such as those in LATEX and ConTEXt and the CJK package developed by Werner Lemberg, which is distributed with the TEX Live CD. The new Omega system tries to work directly with Unicode, which is a popular standard using 16-bit encoding. Despite the differences in the technical implementation details, all of them have assumed that Chinese characters are implicitly expressed in the fixed-length double-byte encoding which are not readable by people.

CTUG (Chinese TEX User Group) is trying another completely different approach to work out this problem. By discarding the man-made implicit assumption of the fixed character length in double-byte, CTUG imported Lojban to represent Chinese encoding in variable length but still in the readable ASCII set. This paper documents the idea in some detail, and points out the future tasks to do under the scenario.

Background Information about Lojban

Lojban (pronounced as LOZH-bahn), which stands for "Logic Language" in Lojban, is nothing new; actually it was presented as a constructed language in 1955 with the name "Loglan" by the project founder Dr. James Cooke Brown. It is based on the "Sapir-Whorf" hypothesis, which states that the structure of a language constrains thought in that language, and constrains and influences the culture that uses it. Over more than the past four decades, Lojban has become a mature artificial language. Here we highlight the main features of Lojban:

- Lojban is designed to be used by people in communication with each other, and now also possibly with computers.
- Lojban is designed to be culturally neutral. It is based on fully phonetic spelling, so people can learn to read Lojban on the fly.
- The regular grammar of Lojban is based on the principles of logic, which has an unambiguous grammar, and has successfully passed the YACC testing. This removes restrictions on creative and clear thought and communication.
- Lojban is designed as a simple language, with just 1,300 root words. Using these root words, it is possible to combine and form millions of words in a vocabulary with ease.

In essence, Lojban is quite close to Chinese grammar, thus a Chinese can quickly become a Lojban user. In the training seminar given by CTUG, practice has shown that some CTUG members could learn and grasp it within a week.

Chinese as Expressed in Lojban

Now, let's check how Chinese words are constructed. Overall, most Chinese linguists have agreed that Chinese has only four methods to construct a character: *XiangXing, ZhiShi, HuiYi* and *XingSheng.* Although it is hard to describe them in formal language, any Chinese character is constructed by one of these four methods, and the first method *"XiangXing"* is fundamental to the construction.

XiangXing means drawing a sign for a given meaning; thus Chinese is classified as an ideograph system in the language taxonomy.

According to researches into the signs recorded in ancient tortoise bones, the most original and frequently used signs are fewer than 500. Tortoise bones are the back shell of tortoises. Chinese people recorded the oracle on them. The signs of the oracle are the oldest Chinese characters we have discovered so far, and they are the origin of modern Chinese characters. And statistically, Chinese characters constructed by *XingSheng* (mostly based on *XiangXing* ideographs) occupy more than 90% of the modern Chinese character repertoire.

A character of *XingSheng* consists of two parts: one part indicates the pronouncation of the character, and the other part indicates the meaning of the character.

For example, my name in Chinese, Hong Feng, is expressed in two characters; each of them is constructed as a *XingSheng* character. Hong has two parts: the three points at the left side is the *Xing* part, and indicates the character is related to water; the right part is the *Sheng* part, pronounced as "gong", meaning the character has "ong" in the pronounciation. The character means large-scale, macro, magnificent, giant.

Feng has two parts too. The left part is also the *Xing* part, a *XiangXing* ideograph for mountains, and the right part specifies the pronouncation to be "feng". The character means the top of the mountain.

Thanks to the more than five thousand years of the simplification movement in the history, the grammar rules of the language are truly simple today. Chinese texts are very similar to an assembly line which we have seen in an automobile production workshop — Chinese characters are placed one by one without stop (i.e. without blank space left between them), quite like the TeX places characters in a box one after another to form a word, and words are placed one after another to form a sentence or a line, and lines are placed one after another to form a paragraph, and paragraphs are placed to be a page to the end.

In the TeX system, if you have a new sign which is not defined yet, then you could design the glyph of the new sign in METAFONT (or in the PostScript language) in a box, and give the box a name (as a new control sequence) to the METAFONT (or to the PostScript) program; after doing that, you could use the new sign with TeX, as if it were one of the built-in readable ASCII characters.

Such cases have happened many times in TeX history. For example, the Euro currency was put into use on January 01, 2002, but the currency symbol was made available much earlier for TeX by two NTG members, who designed the symbol in METAFONT (see MAPS Number 27), so now you could express the Euro in a TeX document directly by `\symbol[euro]`.

Likewise, Chinese characters can be handled in the same way, and once we give each sign in a box a name in Lojban (which also means we discarded the fixed double-byte Chinese encoding, instead, we use variable length of the readable ASCII characters to represent the Chinese, then TeX can be regarded as a native formatter for Chinese immediately!

If we design carefully, we can build a one-to-one mapping table between the existing Chinese encoding set (GB, Big5, Unicode or whatever) and Lojban the expression set, which makes the conversion easy to handle by scripts in Perl or whatever. As Lojban expressions are in readable ASCII, they can be edited using any editor (such as GNU Emacs) even on a simple text-only terminal.

Marriage of TeX and Lojban

As we have seen above, it is possible to encode Chinese by using Lojban as the meta-language. This is the key step to get marriage of TeX and Lojban to happen.

It is necessary to review how TeX defines a control sequence. In TeX, π is defined as `\pi`, likewise, supposing we defined the glyphs of Chinese figures (from zero up to nine) in control sequences in Lojban respectively like this:

Chinese	Lojban
zero	\no
one	\pa
two	\re
three	\ci
four	\vo
five	\mu
six	\xa
seven	\ze
eight	\bi
nine	\so

Now, we can typeset the Chinese number "two zero zero two" this way in TeX: `\re\no\no\re`; the backslash characters won't add too much burden for people to read, and by designing a new macro, it is possible to remove them like this:

`\chinese{re no no re}`.
`TUG''two zero zero two''` (English and Chinese combined together for TUG2002) can be represented in `TUG\chinese{re no no re}`.

TEX and Lojban have agreed to marry.

Tradeoffs and Benefits

The tradeoff of the marriage is obviousl: it adds one step to convert the current Chinese double-byte encodings into Lojban expressions, though we can let a computer do the job automatically, which can be counted as a trivial matter.

Perhaps the true tradeoff would be the requirement for the TEX user to understand more or less about Lojban, thus enlarging the learning curve. Though it is not indispensable, the more one knows Lojban, the higher efficiency would be obtained during the typesetting. If one can read and write in Lojban, then it is possible to typeset Chinese directly in TEX (without the conversion mentioned as above).

Now let's check the benefits. We can just use the word "tremendous" to describe so many benefits we will have under this scenario. Chinese expressed in Lojban is in human readable ASCII characters, which are supported in almost any computer nowadays. This means, all existing programming languages can be used to support Chinese directly too. There are also other benefits to describe the XML's meta data for Chinese in Lojban (again, it is unambiguous in grammar), which goes beyond the scope of this paper.

To-do tasks

To make this scenario become reality, we have several major tasks to do:

- Define a Chinese-Lojban dictionary. As mentioned above in section 2, there are approx. 500 ideographs to define, and it also requires definition of three other methods in Lojban. Lojban is suitable to describe the logic relations because it is designed as a logic language.
- Lojban specification comes with a dictionary which contains ca. 1,300 root words, so it just requires some time and care to build the mapping relation to finish the task.
- Define free, high quality fonts of Chinese. In practice, at least four fonts are required. Now this is a part of the MNM Project (MNM's Not Millions). There are many non-free Chinese fonts available, so commercial publishers can purchase the non-free fonts, and we can add the fonts by applying this scenario.
- Mapping the Chinese fonts adds value to the control sequences in Lojban as key in a hash table. Once the hash table is ready, we could build it into the TEX source tree. The system will be ready to use.

Conclusion

This paper only explained a very small part of the power of Lojban — how to typeset Chinese in TEX using another approach. By importing the idea of re-encoding Chinese in variable length strings of human readable ASCII codes, instead of fixed length in double-byte, the TEX system without any modification can support Chinese typesetting directly.

About the Author: Hong Feng is the founder and the Chairman of the Chinese TEX User Group. He is also the co-founder of the FSF-CHINA Academy. He can be reached by hongfeng@gnu.org

References

[1] Donald E. Knuth. *The TEXbook.* Addison-Wesley, Reading,Massachusetts, 1984.

[2] Hans Hagen. *The euro symbol.* MAPS, Number 27, 2002.

[3] http://www.lojban.org. *Lojban Reference Grammar.*