Moving from bytes to words to semantics

S.K. Venkatesan
TnQ Books and Journals,
8, Second Crescent Park Road,
Gandhinagar, Adyar, Chennai 600 020, India
skvenkat@tnq.co.in

Abstract

Starting from several bytes of ASCII or Unicode strings one can construct a typeset page readable by the community that understands that script. Unfortunately, it still remains unreadable by the larger community of people who don't understand the script. Instead, if this page had been coded at the level of a semantic word, with each word denoting a unique semantic identity, with sufficient markers (the curly bracket nesting being one such example) for grammar and flow, then it would be able to display itself in each language without ambiguity. The eccentricities of ligatures, capitalization, joining of letters could then be handled accurately. Hyphenation, for example, could then be based not on patterns but on the semantics of the word. For example, hyphenation in English tends to depend on whether the word is a noun or a verb. In this work we discuss the possible atomic words (atoms of course have their own protons, electrons, etc.) of a language and the semantic markups that could lead us to such an ideal.

> "Somehow I too must find a way of making things; not plastic, written things, but realities that arise from the craft itself. Somehow I too must discover the smallest constituent element, the cell of my art, the tangible immaterial means of expressing everything."
>
> — Rainer Maria Rilke

Introduction

At a superficial level, it is tempting to identify a language by the script (glyphs) it uses, or, in more modern terms, by the Unicode values of the characters used in the document. Both methods would fail. The first one would fail for scripts that are common to many language users and the second will fail miserably, as Unicode fonts are rarely used by word-processors and typesetters. Many of the e-mail transactions in Indian languages are done through their phonetic Latin script equivalents. Even the complex Chinese language can be phonetically written using the Latin script in the Pinyin system. There was an earlier attempt at Latinizing the Chinese language known as the Wade–Giles system, but due to its shortcomings the new Pinyin system was formed.

Human speech, especially its root-words and structure, is of a language formed much before the introduction of writing systems. Some languages like Vietnamese and Malay–Indonesian changed to their currently used Latin scripts only recently. My native language, Tamil, for example, has a Brahmic script that is similar to Sanskrit but there are even some claims that in its root-word and structural formations it is akin to the primitive Sumer, Elamite and Mande languages. Although the Japanese and Korean[1] languages use the Chinese Han characters (ideographs), the languages themselves have more root-word or structural similarities with the Dravidian languages than with the Chinese languages. For ease of pronunciation the Japanese language uses a smaller subset of simplified Han characters as its alphabet.

Chinese Han script is a good writing system — the characters can carry meanings beyond the spoken language limit, and reading ideographs is faster than reading phonograms such as alphabets, because ideographs directly indicate the meaning while phonograms are changed to pronunciation first and only then the meaning is recognised. People with dyslexia find it difficult to read phonograms but they can understand ideographs easily. It is clear that the Chinese ideographs can have wider applications. Most importantly, the Han script has managed to remain the script that links all the languages

[1] The Korean language was originally written using the Chinese characters; it is now mainly written in Hangul, the Korean writing system, optionally incorporating Hanja to write Sino-Korean words.

of China, creating a single language identity.

The Chinese languages are isolating languages in which the word order, with the help of distinct particles, creates the structure and meaning of the sentence.[2] In the Latin script, a similar artificial isolating language, `lojban`, was created and has generated considerable interest [1]. It is possible to associate an EBNF (a kind of SGML-DTD structure) to `lojban` that makes the language parseable, with words that have unambiguous semantic meaning.[3] `lojban` has clearly demonstrated that an attempt at precision does not make the system rigid. In fact, such attempts surprisingly add to the richness of the language. The root-words in `lojban` have also been carefully selected so as to maintain certain cultural neutrality by including elements from Turkish, Chinese, English, Indian, Russian, Spanish, French, Japanese, and German.

This paper here draws much of its inspiration from the `lojban` effort, and is an attempt to bring `lojban` within the context of the TEX paradigm. Unlike both Chinese and `lojban` we make no attempt at a speech level, conceding that area entirely to the natural languages.

Root-word formation in natural languages

Among the commonly used Dravidian group of languages of India, Tamil has managed to develop by inward growth, rather than by borrowing words from distinct languages such as Sanskrit. Relatively few languages in the world have remained isolated, and managed to avoid direct borrowing of words from other languages. In Europe, the Basque language has had such a self-sustained internal development, but Basque is an amalgamating language that is difficult to learn.

Tamil is an agglutinating language that is explicit and logical, with rules that are easy for children to learn. In Tamil, the formation of `brivla` (compound-words) from `gismu` (root-words) is quite logical and consistent. In Arabic and Hebrew, words are constructed by weaving vowels over root consonant patterns; in line with this in Tamil, the consonants are considered the 'true' (material truth) letters whereas the vowels are considered the ones that provide 'life' (spirit) to it.

It is also no accident that these inward-looking languages are also ones that belonged to matriarchal clan societies with higher in-breeding tendencies. The overthrow of these closed clan societies also meant the mingling and mangling of words used by these societies, leading to the present set of large complex words used in each language. It is however extremely difficult now to look back in time and reconstruct these morphologies in a coherent and consistent manner.

One inspired attempt is the attempt by Asko Parpola [2] who observed a link between the Dravidian languages and the Indus Valley script. The words fish (mIn), star (min-mini), lightning (min-al) identified as a fig-tree (al) with aerial roots from heaven through the astrological associations of Saturn, the slow moving dark planet with the Tortoise, the fish with a roof. The darkness indicating 'mai' associated with tortoise (á-mai). The words 'mal' darkness and 'vel' whiteness are associated with the deities Kannan and Murukan, this being the Tamil equivalent of the Yang and the Yin, at eternal mythological war with their opposites. There is also the undertone of overthrow of matriarchal Yin by the Yang. Like Tamil, perhaps the Chinese languages also have mystical beginnings that spring from tortoise shells, I-ching and soothsayers.

TEX as a paradigm for a new language formation

The industrial and scientific age has also introduced new sets of problems and solutions that require a drastically different outlook from that of the past. The TEX language has been supporting complex scientific symbols and macros making it an ideal platform for a fresh attempt to formulate a mechanism for a modern content-oriented language.

If you are familiar with TEX then these examples will make sense to you:

0. I go there (English)
= *Nan*(I) *ange*(there) *po*(go) (Tamil)

```
\go{0}{I}{there}
```

1. I went there = *Nan ange po-nen*

```
\go{1}{I}{there}
```

2. I am-going there = *Nan ange po-(ki)ren*

```
\go{2}{I}{there}
```

3. I am-going-to-go there = *Nan ange po-ven*

```
\go{3}{I}{there}
```

[2] Even the inflexional English language is showing some tendencies of becoming an isolating language.

[3] `lojban` uses only a subset (lower case) of basic Latin characters (specifically, the letters a to z excluding h, q and w), while uppercase letters are reserved for characters in words of foreign origin that require deviation from `lojban` phonology. We will follow this tradition here.

English, as it is practised today, doesn't have much of a future (tense[4]) for precision semantics, as it is loaded with ambiguity.

You can write Perl macros for writing the above three statements in TEX:

```
\perlnewcommand{\go}[4]
{
 my $smiti_0=$_[0]; my $smiti_1=$_[1];
 my $smiti_2=$_[2]; my $smiti_3=$_[3];
 if($smiti_0==0) { $klama=' go ';}
 if($smiti_[0]==1) { $klama=' went ';}
 if($smiti_0==2) { $klama=' am going ';}
 if($smiti_0==3) {
            $klama=' am going to go '; }
 if($smiti_2=~m/^[A-Z]/) {
          $klama=$klama . ' to ';}
 $text=$smiti_1. $klama . $smiti_2;
 return $text;
 }
\perlnewcommand{\po}[4]
{
 my $smiti_0=$_[0]; my $smiti_1=$_[1];
 my $smiti_2=$_[2]; my $smiti_3=$_[3];
 my $klama='po';
 if($smiti_0==0) { }
 if($smiti_[0]==1) { $klama.='nen';}
 if($smiti_0==2) { $klama.='(ki)ren';}
 if($smiti_0==3) { $klama.='ven'; }
 if($smiti_2=~m/^[A-Z]/) {$smiti_2.='ku';}
 $text=$smiti_1.' '.$smiti_2.' '.$klama;
 return $text;
 }
```

If you are a `lojban`ist then you would write:

```
\klama{#}{mi}{ta}
```

However, all this doesn't prevent you from talking non-sense:

```
\go{3}{I}{yesterday}
```

i.e., I am-going-to-go yesterday. The extra particle 'to' was not required here but it comes into play when we have a specific place, e.g. for

```
\go{0}{I}{Wuhan}
```

we have

I go to Wuhan (English) = Naan Wuhan-ukku po (Tamil)

Tamil being an agglutinating language, the particle 'to' (-ukku) modifies the ending of the place-noun.

I have tried to illustrate by a simple example how languages belonging to very distinct families can be simplified by similar macros. English, like Chinese has a Subject-Verb-Object (SVO) order, while Tamil has Subject-Object-Verb (SOV) order,[5] but our SenseTEX way of writing has made it Verb-Subject-Object (VSO), more like Hebrew, Arabic and Celtic. We can relax this:

```
\SenseTeX{I,\go,Wuhan}
```

The SenseTEX environment will be discussed further in the next section.

SenseTEX

If words can be understood in terms of their underlying meanings, then they can be cross-sectioned, as first pointed out by George Thompson [3], using synonymity and antonymity to a smaller class. Clusters of related adjectives can also be formed. It is also possible to associate a unique number to this synonymous class of words known as the sense number. With more effort the cob-web of words (the cob-web being a graph, loosely speaking) can be cut down to a tree. This can be done by imposing a hypernym–hyponym hierarchy on all words. As this tree travels from the root to the leaves, it traverses from the abstract generalizations (groupings) to the concrete word (from the heaven to the earth, but upside down with roots in the clouds, like the sacred banyan tree).

Speaking of heaven, a search for the word "coke" in any search engines would get results for all the word-senses (coca-cola, charcoal, cocaine, etc.). It is not possible to be word-sense-specific unless the document has sense numbers indicated. (Although recently a search engine `http://beta.previewseek.com`, made by a company based in London, claims to do just that, with of course the usual ambitious claims about patented technology etc., to deter other search engines from reverse engineering it.)

Our approach here is quite straightforward. It uses no proprietary technology nor anything dificult to understand. It just adds value (sense numbers) to TEX when the author needs to do just that.

The user must associate either a sense-number or `lojban` word to each word he uses within the SenseTEX environment, e.g.,

```
\SenseTeX{coke}
```

would not parse — unless a `sensetex.cfg` file has either a valid `lojban` word or a sense-number (as in WordNet) associated with `coke`. More than one word can be used in SenseTEX, e.g.,

```
\SenseTeX{I go Wuhan}
```

[4] For instance, "I will go there" indicates future tense but "will" can add an extra emphasis meaning "I am definitely going there". Moreover, if you just wish to say "I go there tomorrow" — it is not possible.

[5] According to some linguists, Chinese is showing some tendency towards becoming a SOV language.

if each word has a unique entry in `sensetex.cfg`. If you are adept enough to make Λ, Ω, XƎTEX, or ConTEXt work for you, then you can use your own language's script instead of Latin.

We now add the SenseTEX macros:

```
\SenseTeX{\go{#}{I}{Wuhan}}
```

which can be trimmed down to this:

```
\SenseTeX{{I,\go{#},Wuhan}}
```

We can add further sentences, such as

```
\SenseTeX{{I,\go{#},Wuhan},\then,
{I,\go{#},Shanghai}}
```

which could mean "First, I go to Wuhan and then to Shanghai".

Curly brackets can be used to break sentences to logical pieces and double backslashes can be used as paragraph breaks. One must remember that except for the macros the SenseTEX environment behaves just like TEX. For predicate-words like `\go` and `\then`, not just a sense-number (or `lojban` word) but also a SenseTEX macro must be added in the `sensetex.cfg` file.

How are these sense-numbers indicated in print? One way is to print them on top of each word in a smaller point-size (here the `lojban` word 'klama'):

klama
go

If having SenseTEX change the printed output is not desired, then the **hypertex** package can be used and sense numbers (or `lojban` words) can be embedded in hyperlinks that leads to the entries in the sense dictionaries, either in your local system or on the web, e.g.

```
\href{http://www.ctan.org
/macros/sensetex/lojban.htm#klama}{go}
```

Conclusion

SenseTEX is a small beginning. Its future, like everything else, depends on the extensive effort required to build the `sensetex.cfg` file and a sense number based dictionary. From this small beginning one can then perhaps navigate the turbid waters of syntax using macros.

References

[1] `http://www.lojban.org`; see also Hong Feng, "The marriage of TEX and `lojban`", *TUGboat* **23**(1), 46–48, 2003.

[2] Asko Parpola, *Deciphering the Indus script*, Second paperback edition, Cambridge University Press, 2003, ISBN 0-521-79566-4.

[3] `http://wordnet.princeton.edu`