

ITU and UNESCO Global Symposium on Promoting the Multilingual Internet

Geneva, 9-11 May 2006

Paper Submitted by: Louis Pouzin
chairman, NLI consortium

For Discussion in Session: 9: Thematic Session on Multilingualism in Practice

When the initial internet mail protocols were defined, back in 1971 (some would say 1973), all commercial computers had 8-bit characters. Except Multics had 9-bit characters. Nevertheless, the internet mail protocols were based on 7-bit characters. That is ASCII (American Standard Code for Information Interchange), the English alphabet. The internet was only for the happy few who could write their name in English.

During the following 20 years, until 1992, there was not the slightest concern for other than the English alphabet. At this point the Web became a success story in the general public, and created pressure to exchange documents in various non ASCII formats (text, pictures, sounds). MIME (Multipurpose Internet Mail Extensions) were defined. However, applications and user awareness were lagging. It took a number of years for those technical capabilities to be commonly used. Even today, it is not unusual that people cannot open attachments, either due to ill formatting or insufficient expertise.

Still, domain names remained strictly ASCII. Due to strong international demand IDN (Internationalized Domain Names) were introduced in 2003. The major concern was not so much to make the web multilingual, but to keep the DNS (Domain Name System) as is, understand ASCII. Hence mixed alphabets, partly ASCII, partly not, and bizarre gimmicks with right to left languages.

This condensed reminder of the internet history intends to point out a few observations:

- the English dominated internet milieu resists as long as possible "foreign" languages;
- linguistic diversity is only perceived as a set of non English alphabets;
- keeping ASCII as pivotal coding is a prerequisite.

The human, cultural, political components of languages have been eliminated from the landscape. Technical wizards excel in technical solutions, not necessarily in user satisfaction.

When a communication system becomes a daily common tool for everyone, about 80 % of its uses are relations or transactions in the geographical vicinity of the users, in their own language. This is the premier requirement from a societal standpoint. It should also apply from an economic standpoint, unless the system is ill designed.

The lack of a multilingual user friendly interface for internet access has spurred the development of

native language tools, often dubbed "keyword" systems. The term may not be very conducive of a precise meaning, because most data base systems include a keyword based interface for searching. The internet variety comes in two modes. One is the well known search engine. The other is usually called "keyword lookup". A name in native language is converted directly into an internet reference, URL or IP. This one-to-one mapping is similar to the one used in mobile phones, where the user selects a name in an address book, then the phone dials the number.

Previous presentations in this session have very well reported on the success of keyword lookup systems in non English language countries, China, Korea, Turkey. They are the forerunners of a new generation of user friendly tools, matching the user's mindset rather than technical idiosyncrasies.

Evolution will not end there, of course. Once familiar with typed keywords users will presumably upgrade to "voice" keywords, then primitive statements. Insidiously semantics will creep into the dialog, with problems of ambiguities, prejudices, etc. A counterpart of user interface semantics is data semantics, that is, what information would be relevant to the user's concern. There would be a lot more to say in that perspective, but this would be much better presented by knowledgeable persons in linguistics, psychology, sociology and other related fields.

Internet multilingualism is nothing more, but nothing less, than human multilingualism. It should be treated as a societal phenomenon.