NAME: Lluís de Yzaguirre i Maura
TITLE: Neutral lingware: a need for technologically weak languages
PAGES: 3
NOTE: none
ABSTR: Today there are lots of endangered languages. Increasing complexity and social relevance of linguistic technology reduces the chances for endangered languages to survive. Some changes in the way language engeeniering designs and builds its products can make easier to adapt them to other languages.
EMAIL: lluis.deyzaguirre chez upf.edu
ADDR: Institut de Lingüística Aplicada
Universitat "Pompeu Fabra"
J. Xirau, s/n
E-08071 Barcelona
Catalonia
TEL: 34-935-422-269
FAX: 34-935-422-321

Neutral lingware: a need for technologically weak languages

The role of technology in the future of languages

There is no doubt that languages with no written form are highly likely to die due to several reasons: they lack of social prestige, it is difficult to pass them on the new generations and so on. Similarly, language technology is today extremely important in the development of languages. Thus, those languages devoid of synthesisers or voice recognisers will lose ground inasmuch as they will not be used in device control and in text dictation, etc. In short, this will keep these languages away from professional uses: after a few years, computers will not be provided with keyboards because languages such as English, Russian or Spanish will be easily recognised.

The role of market rules in lingware production

Basically lingware production (i.e., any kind of computing programmes, data or physical complements responsible for using or transforming language knowledge) is carried out by private firms that, according to the business rules, seek to obtain benefits. In the case of languages with few speakers or speakers who cannot use or acquire these technologies, private firms do not even consider to adapt their lingwares, especially if the target language has a specific alphabet or is typologically very different from the source language.

Given this view, I am afraid that those programme-producer firms will not account for the 3000 people who live in the Aran Valley (in northern Catalonia) and speak a variety of Occitan. Nor they will cater for the about 30000 Sorbian people, who speak the most Western Slavic language (only 80 km away from Berlin), which is today about to die. Thus Unesco and EU should be aware of the fact that the likeliness of most of the current languages to die will be inversely proportional to its degree of quantity and quality in their language technology. Besides, it should be emphasised that speakers from languages devoid of technology are the vast majority in the world.

A particular case is represented by speakers from minor languages, who are mostly multilingual. They usually find a lingware in the dominant language (superstratum) but have many difficulties (or it is much more expensive) in doing so with regard to their own language.

Alternatives

Today there is an increasing number of standards which can interchange any kind of linguistic data: texts (written or oral), dictionaries, terminographical works, syntactic trees, etc. However there is a lack of public (i.e., non commercial) lingwares that can be adapted to languages without technology. Thus a spell checker for French (including lexical lists, proper names and abbreviations) could be easily adapted to other inflectional languages such as Occitan. Agglutinating languages need another sort of control strategy, although a number of typologically related ones could share it.

Neutral lingware

Neutral lingware will be those that, even in commercial applications, can be reused in several languages. To attain this neutral status, it is required that lingwares read their data from a public format. Furthermore, it also necessary that data are available for the requested language. In this respect Internet has been proved to facilitate these large tasks involving many people.

The vast majority of lingwares can be designed to become neutral. I deliberately disregard all the automatic translation systems due to commercial reasons, since one of the languages involved will be, usually, commercially interesting.

In all cases it would required that the set of characters are interchangeable and that both the keyboard and the ordering criteria can be adapted. Below there are a number of applications that can be neutrally designed:

- spell and style checkers: They can be shared by typologically related languages provided they allow for dictionary editing.

- programmes suitable for looking up monolingual and bilingual dictionaries. They should contain more than one alphabet in the case of bilingual dictionaries.

- morphological and syntactic analysers.

- information retrieval tools, such as neology and terminology detectors

- voice synthesisers. If they are programmed their neutral condition depends on their set of phonemes.

- voice recognisers. If a strategy similar to that of synthesisers were adopted, then voice recogniser will be easier to reuse.

PALIC: a particular case

PALIC is the morphological analyser developed at the IULA. It shows how a programme can be applied to many Romance languages. It is currently working for Catalan, Spanish and French. This is possible because it was designed in this aim and the specific information of each language is external to the programme. As a result, the development of each language will be less hard than the previous one. In the near future languages such as Galician, Portuguese and the variant of Occitant spoken in the Aran Valley could be introduced.

Voice recognition

Voice recognition is being done with particular developments for each language. The development of voice recognisers for languages devoid of technology could be ameliorated if there was a two-stage process: first, a universal stage, common to each group of languages and second, a particular one.

The universal stage would consist in a language-independent phoneme recognition device. The particular stage would consist in the conversion of phonemes to spelling (STT). Actually this is the inverse process of what we have done in the synthesis stage (TTS).

If there were a universal phonemic recognition system suitable for languages without technological resources, then it would only be necessary to develop the automatic spelling module.

Beyond the improving of less technologically developed languages, this strategy has additional benefits:
a) a language recognition module can be placed between the universal phoneme recognition and the automatic spelling in a given language. Thus, recognition can be multilingual
b) a system can become more robust for non-native users.

Conclusion

If it is true (as it seems) that along the XXIst century more than 3000 languages are doomed to die, then it is required to take immediate action so as to provide all languages with linguistic technology. This could be reached by developing public neutral lingwares. It also can be achieved if lingware producers commit themselves to apply their products to languages other than the major ones.

NOTE: I'm grateful to Jordi Morel for style checking

References

De Yzaguirre, L., A. Matamala y T. Cabré (2000a): “El lematizador "PALIC" del IULA (UPF)”, accepted in "XVII Congreso AESLA, panel sobre Lingüística de Corpus y Computacional", Barcelona, 2000, may.

De Yzaguirre, L., A. Matamala, C. Bach, N. Castillo y E. Ustrell (2000b): “AMBILIC, el desambiguador lingüístico del corpus del IULA (UPF)”, comunicación aceptada para el XVII Congreso AESLA, panel sobre Lingüística de Corpus y Computacional.

Unesko Etxea, "World Languages Report", which is currently being elaborated, as presented in Barcelona's Ordinary Meeting of the Advisory Committee for Linguistic Pluralism and Multilingual Education