

Language-specific encoding in multilingual corpora Requirements and solutions

Jost Gippert, Frankfurt a/M

The paper discusses methods of providing language-specific encoding which is an essential feature to be considered in the preparation of multilingual corpora. The necessity of marking the different languages contained in such corpora can easily be demonstrated by looking at word doublets such as French *haut* "high" and German (*er*) *haut* "he strikes" (or even *Haut* "skin") which when encoded uniformly and without marking, can hardly be distinguished as such.

The first part of the paper concentrates on an evaluation of Unicode as a basis of encoding and its limits. Given that Unicode provides a script-based, not a language-based encoding, the use of Unicode will not be sufficient in any case for the purpose of distinguishing language-specific data. Even for scripts such as Greek or Georgian which are primarily used for just one language each, additional marking will be necessary because the scripts mentioned have been in use for other language varieties such as Gaulish or Svan. Another shortcoming of Unicode consists in the treatment of combinations of letters with diacritics. For many of these (e.g., *ä*, *à*, *š*), Unicode provides two ways of encoding, viz. as so-called "precomposed characters" (where the combination is encoded as a single unit) and as a combination of the base letter and its diacritic(s) treated separately (i.e., *a* plus " etc. encoded separately). As there are no "rendering engines" yet that might reidentify such sequences with the intended diacritic unit, the treatment of diacritics has to be considered right from the beginning when developing Unicode-based corpora. In addition, the question is raised whether an extension of the set of "precomposed characters" would be feasible and helpful. The attempt of the TITUS project to prepare such an extension will be introduced.

In the second part of the paper, the encoding strategy of the WordCruncher retrieval system will be demonstrated and discussed. This system which has been applied successfully for multilingual corpora within the TITUS project (cf. <http://titus.uni-frankfurt.de/texte/texte.htm> and <http://titus.uni-frankfurt.de/texte/tituswc.htm>) has the shortcoming of joining language marking to font marking. A possible way out of this problem can consist in converting the markings into an SGML/XML based tagging system, the properties of which will have to be established according to the specific

requirements of the languages involved and with a view to the requirements of linguistic retrieval.