Developing TEI-Conformant Lexical Databases for CEE Languages

Tomaž Erjavec
Tomaz.Erjavec@ijs.si
Department of Intelligent Systems, Jožef Stefan Institute
Dan Tufiş
tufis@valhalla.racai.ro
RACAI, Romanian Academy of Science
Tamás Váradi
varadi@nytud.hu
Linguistics Institute, Hungarian Academy of Sciences

31. december 1998

1 Introduction

The present paper reports on ongoing work in the INCO-COPERNICUS project CONCEDE (Concortium for Central European dictionary Encoding). The paper is structured as follows. After setting out the aims of the project, it will discuss the approach and the work programme adopted to achieve its aims. The paper will highlight on the issue of headword selection as one key methodological problem that the project had to face. In conclusion, the expected impact of the project will be illustrated.

2 Aims and rationale

CONCEDE has the following five strategic goals:

1. delivery of lexical databases for each of the six participating CEE languages

2. testing and extenstion of the TEI dictionary guidelines to accomodate CEE languages

3. transfer and development of dictionary encoding and management tools to the CEE languages

4. development of tools to support multilingual, integrated access to a lexical and corpus database for the six CEE languages

5. transfer of expertise in dictionary encoding and lexical database design from EC to CEE partners

The CONCEDE project will develop lexical databases, in a general-purpose document-interchange format, for the languages of the six participating CEE partners: Bulgarian, Czech, Estonian, Hungarian, Romanian and Slovene. As far as fundamental differences between the languages and possibly lexicographic practices allow, all LDBs will have the same structure. The size of the LDBs developed will vary (due to resourcing constraints) from around 500 lemmas to 4500 with an average of 2500 lemmas.

3 Background and approach

For the major Western European languages, there now exist a range of lexical resources which are useable by Language engineering systems. By making such resources available for CEE languages, CONCEDE will support a corresponding development of the language engineering industries in those countries.

3.1 Resource development

Obviously, electronic lexical resources owe a great deal to traditional lexicography. It is largely the information in paper dictionaries which has populated the lexical databases now in use and indeed with one exception most CEE CONCEDE partners will make recourse to some machine readable dictionaries. Up-translating an electronic version of the paper dictionary to a lexical database can be undertaken semi-automatically [1]; [3]. This is a process that the EU partners have direct experience in, which will be applied to the CEE languages.

3.2 Resource standardisation

One of the key observations arising out of the accumulated experience in the field is the value of standardisation. For obvious practical constraints, the TEI dictionary encoding guidelines were developed for and validated against Western European languages. By adopting and, where necessary, adapting the TEI standard to the CEE languages CONCEDE will not only offer the practical benefit of (re-)usable lexical resources but will also test and validate the wider applicability of the TEI guidelines.

4 Work programme

To achieve its goals CONCEDE is following a two-stage approach. In the first stage, the standardisation issues will be resolved and pilot LDBs of a representative sample of around 500 words for each language developed as a basic proof of concept. In the second phase, these dictionaries will be enlarged to medium-scale resources of the envisaged size. The full paper will contain a discussion of the steps involved in the process.

5 Headword selection

One issue of common interest not only to the CONCEDE partners but, we believe, in a wider context as well is the question of finding a suitable small-scale, balanced sample of the lexicon of a language. We designed a language independent methodology used it on the CESANA encoded parallel corpus “1984” by Gerge Orwell, which was one of the deliverable in the MULTEXT-Eastproject.[2] (CESANA encoding specifies for each wordform, among other things, its associated grammatical information and the corresponding lemma). A balanced sample for parts of speech means that the part of speech (POS) distribution of the sample has to reflect the corresponding distribution of the different parts of speech in the corpus.

A simplistic approach applying the formula in (1) would have the disadvantage that it would be slanted against certain parts of speech that have few lemmas but these lemmas occur very frequently in the corpus.

$$n_{POS} = \frac{N_{POS}}{N_L} * n_L \tag{1}$$

where

N_{POS} = number of lemmas in the corpus of a given POS
N_L = number of all lemmas irrespective of their part of speech
n_L = sample size, i.e., the number of lemmas to be chosen
n_{POS} = number of lemmas in the sample of a given POS

To remedy the above shortcoming, we have applied a statistical method that involves dividing the corpus into a sequence of test samples of fixed number of lemmas (500 in our case). The test samples reflect the usage degree of a certain part of speech in the language context. The proposed formula that gives the number of part of lemmas of a given POS in a sample is displayed in (2).

$$n_{POS} = \frac{\sum_{i=1}^{n_T} n_{POS}^i}{n_T} \tag{2}$$

where
n_{POS} = number of lemmas in the sample of a given POS
n_L = number of test samples
n_{POS}^i = number of lemmas of a given POS in the i^th test sample

The advantages of this method are the following:

- the POS distribution in the test samples reflect the structural POS distirbution of the language. This fact is supported by the stylistic and author coherence of the text.

- the number of POS lemmas chosen does not depend on the whole number of lemmas in the corpus. In other words, it is independent of the size of the corpus.

Selection of the headwords was carried out in to terms of POS and frequency of occurrence in the corpus. Lexemes were divided into three frequency ranges (high, medium, low) established and for each POS the sum of high, medium and low frequency lemmas had to conform to the n_{pos} value as computed above. Full details of how the headword list was compiled will be given in the final version of the paper.

6 Conclusion

By providing much-needed lexical resources in a standard reusable way, CONCEDE is expected to have a major impact on the language engineering sector of the region. By selecting words on the basis of their frequency in naturally occurring texts for the languages, rather than by some artificial notion of which words might be useful, CONCEDE will make the lexical databases maximally useful for real applications.

A further important feature of the LDBs arises from their aligned nature and relationship to the MULTEXT-Eastcorpus. Although the lemmas for each language have been chosen independently, they will be drawn from corpora that are aligned with each other. This means that the majority of words in the LDB for one lanugage will be correlated with words in the corresponding LDB for the other languages. This mapping may not be one-to-one, it nevertheless can be expected to support a range of bilingual and multilingual applications, such as concordance-based translation among the six languages and English.

References

[1] B. K. Boguraev and E. J. Briscoe, editors. *Computational Lexicography for Natural Language Processing.* Longman, Harlow, 1989.

[2] Tomaž Erjavec and Nancy Ide. The MULTEXT-EAST corpus. In Antonio Rubio, Natividad Gallardo, Rosa Castro, and Antonio Tejada, editors, *First International Conference on Language Resources and Evaluation, LREC'98*, pages 971–974, Granada, 1998. ELRA.

[3] Nancy Ide, J. Le Maitre, and J. Véronis. Outline of a model for lexical databases. *Information Processing and Management*, 29(2):159–186, 1993.