The Concede model for Lexical Databases

Tomaž Erjavec,[*] Roger Evans,[†] Nancy Ide(3),[‡] Adam Kilgariff[†]

[*] Dept. for Intelligent Systems, Jozef Stefan Institute
Jamova 39, SI-1000 Ljubljana, Slovenia
tomaz.erjavec@ijs.si
[†] Information Technology Research Institute, University of Brighton
Brighton, U.K.
{Roger.Evans, Adam.Kilgarriff}@itri.brighton.ac.uk
[‡]Dept. of Computer Science, Vassar College Poughkeepsie, USA
ide@cs.vassar.edu

Abstract

The value of language resources is greatly enhanced if they share a common markup with an explicit minimal semantics. Achieving this goal for lexical databases is difficult, as large-scale resources can realistically only be obtained by up-translation from pre-existing dictionaries, each with its own proprietary structure. This issue is a central concern in our work in the CONCEDE project, which aims to develop compatible lexical databases for six Central and Eastern European languages. This paper describes the approach we have taken in CONCEDE. Starting with sample entries from original presentation-oriented electronic representations of dictionaries, we discuss how we first transform the data into an intermediate TEI-compatible representation, and from there into a more restrictive shared encoding, formalised as an XML DTD with a clearly-defined semantic interpretation.

1. Introduction

The EU INCO-COPERNICUS project CONCEDE (Consortium for Central European Dictionary Encoding) aims to build structured lexical databases derived from existing machine-readable dictionaries for six Central and Eastern European languages: Bulgarian, Czech, Estonian, Hungarian, Romanian and Slovene. One of the goals of the project is to deliver these databases as an integrated resource, complementing the annotated parallel corpus for the same six languages developed under the MULTEXT-EAST project (**?**). To achieve this, the databases must as far as possible share a common markup scheme, using the same tags and giving them the same interpretations. At the same time, it is important not to lose useful information (content or structure) from the original representations.

We have addressed this problem by considering in the first instance samples of 500 entries from each of the dictionaries. The samples were chosen on the basis of frequency list derived from the MULTEXT-EAST corpus, and the sampling process has been described in detail in (**?**). We adopted a three-stage approach to the encoding, as explained in the following sections.

2. Encoding the Dictionaries in TEI

The Text Encoding Initiative (TEI, (**?**)) provides SGML-based guidelines for encoding of vaious kind of language data. TEI also defines a base tag set for Print Dictionaries, i.e. for "encoding human-oriented monolingual and polyglot dictionaries (as opposed to computational lexica, which are intended for use by language-processing software)". The encoding of dictionary samples in a TEI perscribed manner was the first step in moving towards the Concede LDB.

The original dictionaries came in a variety of legacy formats, from Word to SGML. The conversion involved many special-purpose filters, and decisions on how to represent given information in TEI.dictionaries. At this stage, the guiding principle was to preserve or further detail the information found in the original digital format. In some cases, it was necessary to introduce extensions to the standard encoding scheme, to support richer element content models but this was done within the guidelines for such extensions.

The paper will give examples of markup and extensions and the rationale behind them.

Here insert stuff from http://nl.ijs.si/CnC/cnc-DR4.1.html

3. The Concede DTD

With the information now in a standard format, we were in a position to develop a single DTD to cover all the dictionaries. We have used XML (Extensible Markup Language) [4], due to its emergence as the de facto standard for data representation, and in order to take advantage of facilities developed within the XML framework (e.g., the Extensible Style Language (XSL) [5]).

Our guiding principle was to provide a DTD with as few elements as possible, each with an unambiguous, clearly-defined interpretation. This task breaks naturally into two parts: content and structural elements.

For content elements, we identified an inventory of TEI elements (orth, pron, hyph, syll, stress, pos, gen, case, number, tns, mood, usg, time, register, geo, domain, style, def, eg, etym, xr, trans, itype) capable of representing all the content elements in the source dictionaries (not necessarily 1-1), and fixed their TEI interpretations.

For structural elements, we follow the observations [6] that certain underlying regularities exist in all print dictionaries (in particular, the use of a hierarchical organization that enables the factoring of information over nested levels) and that all levels in dictionary hierarchies potentially contain the same elements. Therefore, we adopt a simple general scheme involving just three structural elements:

- ¡struct¿ introduces nested structure of any sort, inducing tree structured elements down which information is

inherited by default - ¡alt¿ introduces alternative elements within a ¡struct¿ at any level - ¡brack¿ provides for grouping of information other than as a sub-structure (for example, in order to associate a cross-reference with some elements only of a ¡struct¿, or to group elements of like kind, such as syntactic elements).

Each of these element types has a simple semantics in terms of grouping and inheritance of information, which is implemented using the XSL Transformation language (XSLT).

The paper will provide examples and the DTD.

4. Converting Dictionaries into the Concede DTD

Apart from hand converting sample TEI dictionary entries into the CONCEDE DTD encoding, we have also performed a largely automatic conversion of the complete 500 sample from the English-Slovene dictionary. The conversion produced a valid and well-formed CONCEDE lexical database, but at the cost of preserving only shallow dictionary information. In particular, some elements (usg, lbl, label, gloss) are suppressed in the target. These elements exhibit more complex scopings in regard to other elements, and are often inconsistently encoded.

The conversion heavily exploited the fact that the input and output encodings are in SGML. This enabled utilising SGML-aware tools, where each step of the conversion is validated against a (possibly intermediary) DTD, and errors analysed. Errors can be caused by encoding inconsistencies, which are corrected in the TEI source, or patterns not taken into account by the program.

The paper will provide examples of the mapping, and the kind of problems that arise.

Include http://nl.ijs.si/CnC/cncjsi-DR3.1.html ?

Include http://nl.ijs.si/telri/Bratislava/cnc-mte.html ?

5. Citation Format

REFERENCES

[1] Ludmila Dimitrova, Tomaz Erjavec, Nancy Ide, Heiki-Jan Kaalep, Vladimr Petkevic, and Dan Tufis. Multext-East: Parallel and Comparable Corpora and Lexicons for Six Central and Eastern European Languages. In COLING-ACL '98, pp. 315-319, 1998.

[2] Tomaz Erjavec, Dan Tufis, Tamas Varadi. Developing TEI-Conformant Lexical Databases for CEE Languages. COMPLEX'99, pp. 205-209, 1999.

[3] C. M. Sperberg-McQueen and Lou Burnard (eds.) Guidelines for Electronic Text Encoding and Interchange. Chicago and Oxford, 1994.

[4] Bray, T., Paoli, J., Sperberg-McQueen, C.M., eds. (1998). Extensible Markup Language (XML) Version 1.0. W3C Recommendation. http://www.w3.org:TR/1998/REC-xml-19980210

[5] Clark, J., ed. (1999). XSL Transformations (XSLT). Version 1.0. W3C Recommendation. http://www.w3.org:TR/xslt

[6] Nancy Ide, Jean Veronis. Encoding Dictionaries. In "The Text Encoding Initiative: Background and Context" Kluwer, Dordrecht, pp. 167–180, 1995.

6. References