From Machine Readable Dictionaries to Lexical Databases: the CONCEDE Experience

TOMAŽ ERJAVEC[1], ROGER EVANS[2], NANCY IDE[3], ADAM KILGARRIFF [2]

(1) Dept. of Intelligent Systems
Jožef Stefan Institute
Ljubljana, Slovenia

(2) ITRI
University of Brighton
Brighton, U.K.

(3) Dept. of Computer Science
Vassar College
Poughkeepsie, USA

Abstract

It is commonly held that machine-readable dictionaries play a key role in bootstrapping effective wide-coverage language-technology, especially in less well-resourced languages. However, while the linguistic knowledge they contain is clearly necessary for this goal, it is far from clear that the format it is presented in is sufficient to reach it. A crucial step in the deployment of such resources is to map them into lexical databases with standardised and well-understood structure and semantics. Furthermore, considerable additional benefits are obtained if such structure and semantics are shared with other linguistic resources. Achieving such a goal, however, is often not an easy task.

This paper describes how such a mapping was carried out in the CONCEDE project, for six Central and Eastern European Languages (Bulgarian, Czech, Estonian, Hungarian, Romanian, and Slovene) for which few wide-coverage lexical resources had previously been available. In a two-stage process, the machine-readable data for each language was first mapped into broadly compatible, TEI-compliant SGML representations, and then these representations were harmonised into a single XML scheme. The resulting framework offers a concise, flexible lexical database specification, with a demonstrable ability to cope with a diverse range of dictionary and language requirements, and lexical resources suitable for monolingual and multilingual application.

1. Introduction

The value of language resources is greatly enhanced if they share a common markup with explicit minimal semantics. Achieving this goal for lexical databases (LDBs) is difficult, as large-scale resources can realistically only be obtained by up-translation from pre-existing dictionaries, each with its own proprietary structure. Furthermore, proprietary dictionary data sets developed primarily to support the production of printed dictionaries (for example, for typesetting) are notoriously difficult to formalise, due to lack of a formal specification of the data, failure to conform to specifications when provided, or simply errors (of content, structure or simply typography). The EU project CONCEDE[1] constructed lexical databases from existing machine-readable dictionaries for

[1] *Consortium for Central European Dictionary Encoding* – INCO-COPERNICUS project no. PL96-1152. The support of the European Commission for this research is gratefully acknowledged.

six Central and Eastern European languages, five monolingual and one bi-lingual: Bulgarian, Czech, Estonian, Hungarian, Romanian and English-Slovene. One of the goals of the project was to deliver these databases as an *integrated* resource, sharing database structure between the six languages and complementing and enhancing the annotated parallel corpus for the same six languages developed under the MULTEXT-East project (Dimitrova et al., 1998). To achieve this, the databases were represented using a common markup scheme implemented in XML (W3C, 1998). This scheme needed to ensure that the same structural tags were used and given the same interpretations in the different databases, while incorporating the flexibility to allow entries to retain potentially useful information (content or structure) from the original sources that could not be analysed within the framework.

In this paper, we describe the approach that was taken to develop the CONCEDE resources, indicating some of the issues we encountered along the way, and giving examples of the resources developed. In the next section we review the overall project structure and methodology, some aspects of which have been published in more detail elsewhere. We then focus on the CONCEDE DTD, the formal model we developed to encode our lexical databases, and illustrate it with examples from the delivered LDBs. This is followed by a section on the relationship with the MULTEXT-East parallel corpus, and integrated CONCEDE/MULTEXT-East demonstrators, and an overall conclusion.

2. Project overview

The CONCEDE project proceeded by first obtaining the source dictionaries in digital format, then sampling – in two stages – entries from these dictionaries, performing, again in two stages, the conversion to the CONCEDE format and validating the results. This section describes this process.

2.1 The source dictionaries

Sources for the dictionary data in the six languages were established in principle before the project began. Nevertheless the practicalities of accessing the data were not always straightforward. Several of the dictionaries were being produced in electronic form concurrently with the CONCEDE project, which meant that the availability of sample entries was constrained by the production process. In one case the entire dictionary was re-keyed because access to an electronic form proved problematic. The following table summarises the dictionaries used:

Language	Dictionary	Approx. Size
Bulgarian	Bulgarian Explanatory Dictionary	10,000 entries
Czech	Dictionary of Standard Czech for School and Public	24,000 entries
Estonian	Defining Dictionary of Standard Estonian	100,000 entries
Hungarian	Hungarian Explanatory Hand Dictionary	70,000 entries
Romanian	Dictionarul Explicativ al Limbii Roman	65,000 entries
English-Slovene	English-Slovene Dictionary Oxford-DZS	>300,000 entries (under development)

2.2 Encoding specifications

The starting point for the formal model of the dictionary to be used in CONCEDE was the Text Encoding Initiative (TEI) guidelines for dictionary encoding (Sperberg-McQueen and Burnard, 1994) in SGML. However, whereas these were designed essentially for descriptive purposes, so that existing dictionaries could be accurately described, the concern here was also prescriptive: to provide a model for how a lexical resource should be marked up if it is to be useful for language engineering applications. To this end, our objective was to have just one correct way to mark up any phenomenon, in any of the languages. This was an ambitious goal, as the input dictionaries were very different and there were very large numbers of phenomena on which consensus was, in principle, required. Within the resource constraints of the project, the CONCEDE model was fully developed only for those phenomena which were salient for language engineering and not infrequent. The model was embodied in the CONCEDE DTD (Erjavec et al., 2000), which is further discussed in Section 3. Over the duration of the project, XML displaced SGML as the formalism of choice for language resources of this kind. In recognition of this, a second XML version of the CONCEDE DTD was also developed. The two versions of the DTD are substantively identical, and we shall not dwell here on difference between them.

2.3 Lexicography

Lexicography in CONCEDE proceeded in two phases. In the first phase, a 500-word sample of each language was up-translated into TEI-based SGML. The emphasis in this phase was on understanding the issues for each language in the up-translation process, rather than seeking uniformity across the languages. For each language, the lexicography was an extensive iterative process of writing code, running it over the dictionary, examining the output, and refining the code. The six phase 1 databases were created and validated against their own individual TEI-based DTD specifications. The experience gained from this process was used to develop the CONCEDE DTD – a uniform representation capable of supporting the requirements of all six languages.

The second phase of lexicography first dealt with choosing the headword list to include in the final LDBs. Due to copyright and labour constraints the project did not envision converting the complete dictionaries, but rather choosing a representative sample of each dictionary and concentrating on that. The selection process, which was based on corpus evidence, is explained in detail in Erjavec et al. (1999). After the selection, the complete wordlist for each language was encoded according to the CONCEDE DTD. The encoding algorithms developed for phase 1 were adapted to deliver CONDEDE DTD-compliant output and the same iterative encoding process was undertaken for the larger headword list.

2.4 Conversion

The conversion from the TEI encoding to the CONCEDE LDB was performed differently for the different languages, but was in all cases fully or partially automatic. To give as an example the English-Slovene case, the converted LDB contains most of the content from the original, but omits about 10% of elements which we currently cannot yet exploit and which also have the most difficult (inconsistent) placement and scoping. The conversion process heavily exploited the fact that the input and output encodings are in (or at least compatible with) XML. This enabled us to utilise XML-aware tools, where each step of the conversion is validated against a (possibly intermediary) DTD, and errors analysed. The errors were corrected by upgrading the conversion from the original

digital format, or by manually correcting one of the intermediary CONCEDE documents; some were also left in as further work.

2.5 Validation

Validation of the CONCEDE lexical databases considered two aspects: 'form' and 'content'. The formal validation was a matter of ensuring that the databases were valid XML documents according to the CONCEDE DTD and XML declaration. Much of this validation was achieved through the use of automatic encoding methods for the LDBs. A global check of the final databases using formal validation tools ensured their overall XML integrity. At this stage certain structural aspects of the databases were also unified, e.g. the division of databases into header and body files, and consistent marking of identifiers on certain structural elements.

Regarding content validation, the iterative up-translation methodology requires the lexicographer to examine successive output in search of errors and so has a high degree of validation built in. This represents the limit of content validation carried out for the phase 1 lexical databases. For the final databases, a more objective validation process was undertaken, with each database scrutinised by partners from a different site. Due to resource limitations only a small sample of each database was cross-validated in this way, namely 30 words. This sample comprised 4 closed-class items, 8 nouns, 8 verbs, 5 adjectives, 5 adverbs, with half the open-class words having 3 or more senses, and including items with multi-word units.

With the conversion from TEI to the CONDEDE DTD we obtained the final lexical databases in the CONCEDE format. The LDB distribution comprises a directory containing the LDB DTD and the six lexical databases encoded in XML. Each database is stored in two files: the header and the body of the LDB. HTML versions of the databases have also been derived. The sizes of the final databases (project target size and actual delivered size) are given in the following table:

Language	Target size	Delivered size
Bulgarian	4500	2830
Czech	3000	10623
Estonian	1500	7115
Hungarian	2500	6852
Romanian	4500	4600
English-Slovene[2]	500	500

3. The CONCEDE DTD

The unified CONCEDE DTD (Erjavec et al., 2000) aims to provide a minimal model, with as few elements as possible, each with an unambiguous, clearly-defined interpretation. We identified an inventory of TEI elements capable of representing all the content elements in the source dictionaries, and fixed their TEI interpretations. These elements are: <orth>, <pron>, <hyph>,

[2] For logistical reasons, CONCEDE was only able to undertake a proof-of-concept implementation for English-Slovene.

<syll>, <stress>, <pos>, <gen>, <case>, <number>, <tns>, <mood>, <usg>, <time>, <register>, <geo>, <domain>, <style>, <def>, <eg>, <etym>, <xr>, <trans>, <itype>. For structural elements, we followed Ide and Véronis, (1995) and introduced a simple general scheme involving three structural elements to capture inheritance, alternation and general grouping:

<struc> represents a node in the tree. <struc> elements may be recursively nested at any level to reflect the structure of the corresponding tree. <struc> is the only element in the encoding scheme that corresponds to the tree structure; all other elements provide information associated with a specific node (i.e., the node corresponding to the immediately enclosing <struc> element).

<alt> alternatives may appear within any <struc>. The use of this element corresponds to the shorthand often used in dictionary entries, where two equally applicable sets of information apply to the entire sub-tree, as where there are two possible spellings and two or more meanings, and either spelling can be coupled with any meaning.

<brack> is a general-purpose bracketing element to group associated features.

A major feature of the DTD is the specification of the information content of the nodes in the dictionary entry hierarchy. In particular, we specify a notion of inheritance, and distinguish monotonic from overwriting inheritance. In brief, information is inherited down the <struc> tree, with sister <struc>s being viewed disjunctively. Many-valued attributes, such as <domain> accumulate values though inheritance. Single-valued attributes have an overwriting semantics: a value will be inherited from the nearest ancestor <struc> (including this <struc> as first 'ancestor') – see (Erjavec et al., 2000) for further details.

Figure 1: An English/Slovene entry

```
<entry id="ensl.44">
  <hw>although</hw>
  <pos>conj</pos>
  <pron>O:l"D@U</pron>
  <struc type="sense" id="ensl.44.3">
    <trans>
      <alt type="orth">
        <orth>&ccaron;eprav</orth>
        <orth>&ccaron;etudi</orth>
      </alt>
    </trans>
    <struc type="eg">
      <eg><q>they're generous, although poor</q></eg>
      <trans><orth>radodarni so, &ccaron;eprav revni</orth></trans>
    </struc>
  </struc>
  <struc type="sense" id="ensl.44.4">
    <trans><orth>vendar</orth></trans>
    <struc type="eg">
      <eg><q>I think he's her husband, although I'm not sure</q></eg>
      <trans><orth>mislim, da je njen mo&zcaron;, vendar nisem prepri&ccaron;ana</orth></trans>
    </struc>
  </struc>
</entry>
```

Figure 1 shows an example entry encoded using the CONCEDE DTD. This example is from the English-Slovene bi-lingual dictionary and is the entry for the English headword **although**. The entry introduces the headword, part of speech and (English) pronunciation, and then splits into two <struc> elements corresponding to two distinct senses. The first sense provides a translation into Slovene with two alternative orthographic forms, followed by a sub <struc> giving an example in English and its Slovene translation. The second <struc> corresponds to the clausal sense of 'although', and provides a similar structure but without any orthographic alternatives.

Figure 2 shows a further example from the Romanian dictionary and is the entry for the headword **tinerețe** (youth). Due to its length the example contains only the first sense, but what is important is its treatment of morphological information: the two forms of the noun (which together with the headword determine its complete inflectional behaviour) are in an <alt> relation, while the grammatical / form information is grouped together with <brack>.

Figure 2: A Romanian entry

```
<entry id="ro.2">
  <hw>TINERE&Tcedil;E</hw>
  <stress>TINER`E&Tcedil;E</stress>
  <alt>
    <brack>
      <gram>nom_substantiv_sing_indef</gram>
      <orth>tinere&tcedil;e</orth>
    </brack>
    <brack>
      <gram>nom_substantiv_pl_indef</gram>
      <orth>tinere&tcedil;i</orth>
    </brack>
  </alt>
  <pos>substantiv</pos>
  <gen>fem</gen>
  <struc id="ro.2.6" n="1">
    <def>Perioad&abreve; din via&tcedil;a omului &icirc;ntre copil&abreve;rie &scedil;I  maturitate.</def>
    <struc id="ro.2.6.2" type="Sec">
      <def>Perioada de &icirc;nceput a existen&tcedil;ei unui animal sau a unui copac.</def>
    </struc>
  </struc>
…
```

As mentioned, the complete lexical databases were by the end of the project converted to the CONCEDE DTD. In order to give an impression of the information content of lexicons, we give, in Appendix A, the tagcounts of the six LDBs.

4. The MULTEXT-East "1984" corpus

The CONCEDE project had strong connections to the MULTEXT-East project (Multilingual Text Tools and Corpora for Eastern and Central European Languages – see Dimitrova et al. (1998)). Most of the partners were the same in both projects, and the parallel aligned corpus developed in

MULTEXT-East were used in CONCEDE to provide the data for the sampling of the word-lists in CONCEDE.

In addition, CONCEDE supported a second release of the MULTEXT-East corpus. In the years since the MULTEXT-East resources were released, they have served as models for reference corpora and used in a number of experiments, e.g. in evaluating part-of-speech tagger performance, developing new taggers and lemmatisers, automatic extraction of bi- and multi-lingual lexicons and studies on multilingual sense disambiguation. Over this time, a number of errors and inconsistencies were discovered in the data and specifications. These errors were subsequently corrected, but because the work was done at different sites and in different manners, the corpus encodings had begun to drift apart. The CONCEDE project offered the possibility to bring the versions back on a common footing, in order to deliver an integrated corpus and dictionary resource. Thus, the corrected "1984" corpus was normalised and the primary data re-encoded according to the TEI guidelines and, largely, to the XML recommendation.

This "CONCEDE" version of the MULTEXT-East corpus has been released as part of the CONCEDE deliverables. This version, further discussed in Erjavec (2001), contains: the revised and expanded MULTEXT and EAGLES based morphosyntactic specifications, in print form and as (over 5000) TEI feature structures; the morphosyntactic lexicons, totalling at least 15,000 lemmas per language; and the corrected and TEI encoded "1984" annotated corpus, with about 100,000 words per language. The corpus includes 2-way and 7-way sentence alignments in CES (Corpus Encoding Standard).

Two demonstrators of this combined MULTEXT-East/CONCEDE resource were also developed during the project. In the first demonstrator, the "MULTEXT-East Sampler", which provides chapter 1 of "1984" rendered in HTML in all six languages, was extended to include links from words to the corresponding CONCEDE LDB entries, also rendered in HTML. This provides a convenient way to explore some parts of the CONCEDE LDBs in context, and carry out comparisons of encoding and coverage between the different CONCEDE languages. The second demonstrator is a bi-lingual concordancer, which uses the CONCEDE English-Slovene bilingual LDB together with the aligned English and Slovene MULTEXT-East corpora. It allows the user to browse the LDB entries, and obtain on-the-fly generated bi-lingual corpus examples about the particular elements of an entry. This demonstrator takes advantage of the CONCEDE DTD inheritance mechanism, by constructing corpus queries based on all the information available via the ancestors of a queried element.

5. Conclusions

This paper has presented the CONCEDE lexical databases and the XML encoding model used to structure them. We have described the approach that was taken to develop the CONCEDE resources, from source dictionaries, through encoding of a sample, development of a unified DTD to conversion and validation of the final lexical databases. The formal LDB model of CONCEDE was presented in more detail, also giving examples from the LDB. Finally, the CONCEDE edition of the MULTEXT-East parallel corpus was presented, and two demonstrators built on the combined resource were briefly described.

CONCEDE has met its main objectives as originally envisaged. The primary objective was *"to deliver medium-sized TEI-conformant lexicons"*, suitable for Language Engineering use, for the six

languages. This goal has been achieved and in several cases well exceeded in its own terms, although to some extent what counts as medium-sized may have moved on since the project was proposed. Secondary objectives were *"spreading expertise"* and *"validating TEI-DWG guidelines"*. All the partners were fully engaged in the encoding, lexicography and validation processes, so that the project has clearly spread expertise in lexical encoding well beyond its traditional anglo-centric base. The project has also played a key role in the ongoing development of standards for dictionary/LDB representation. It established limitations to the suitability of the TEI-DWG guidelines for language engineering purposes and proposed a more appropriate model, which builds on the TEI work, taking most of its tags from the TEI tagset, but is more constrained in what it allows. The resulting CONCEDE DTD is currently being incorporated into the XCES standard as the recommended DTD for encoding LDBs in CES. Most importantly, perhaps, the project has succeeded in providing foundational resources for work in Language Engineering in these six languages, for morphological, grammatical, semantic or other research, or as the basis for development of more commercial applications.

References

Dimitrova, Ludmila, Tomaž Erjavec, Nancy Ide, Heiki-Jan Kaalep, Vladimír Petkevič, Dan Tufiş, 1998. Multext-East: Parallel and Comparable Corpora and Lexicons for Six Central and Eastern European Languages. In COLING-ACL '98. Montreal, Quebec, Canada.

Erjavec, Tomaž, Dan Tufiş, Tamas Varadi, 1999. Developing TEI-conformant lexical databases for CEE languages. In Proceedings of the 4th International Conference on Computational Lexicography, COMPLEX'99. Pecs, Hungary.

Erjavec, Tomaž, Roger Evans, Nancy Ide, Adam Kilgarriff, 2000. The CONCEDE Model for Lexical Databases. In Second International Conference on Language Resources and Evaluation, LREC'00. pp.355-362, ELRA, Paris.

Erjavec, Tomaž, 2001. Harmonised Morphosyntactic Tagging for Seven Languages and Orwell's 1984. In. 6th Natural Language Processing Pacific Rim Symposium, NLPRS'01 pp. 487-492, Tokyo.

Ide, Nancy and Jean Veronis, 1995. Encoding Dictionaries. In The Text Encoding Initiative: Background and Context, Kluwer Academic Publishers, pp. 167-180.

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

W3C, 1998.*Extensible markup language (XML) version 1.0.* http://www.w3.org/TR/1998/REC-xml-19980210.

Appendix A. Tag counts of the CONCEDE LDBs

Element	Bg	Cs	EnSl	Et	Hu	Ro	Element	Bg	Cs	EnSl	Et	Hu	Ro
<alt>	831	4870	1805	4	643	8151	<na>	-	131	-	61	7024	-
<aspect>	-	2714	-	-	-	-	<number>	158	6172	-	175	-	63
<brack>	12	6506	-	-	10851	12425	<oref>	-	-	-	-	16631	-
<case>	-	4977	-	-	-	-	<orth>	3454	14739	9435	7119	12510	12798
<def>	8923	11832	-	22670	23395	13101	<per>	34	593	-	-	-	-
<degree>	-	149	-	-	-	-	<pos>	1063	10655	808	6635	7421	6591
<eg>	8946	26702	3974	30598	-	-	<pron>	-	152	538	2	14	581
<entry>	**2830**	**10623**	**500**	**7115**	**6852**	**4600**	<q>	10360	26702	3974	32957	11931	-
<etym>	356	480	-	-	1281	5311	<source>	273	-	-	11639	-	-
<gen>	1310	4117	-	-	-	3657	<stress>	-	-	-	-	-	5118
<gloss>	-	5939	-	3135	61	-	<struc>	10543	19001	6600	23084	27208	10460
<gram>	1292	115	-	-	283	10348	<subc>	1225	10	-	52	2803	1220
<hw>	2830	10623	500	7115	6852	4600	<tns>	16	427	-	-	-	-
<hyph>	-	-	-	1094	-	-	<trans>	-	-	6005	-	54	118
<itype>	40	-	-	3831	-	852	<usg>	1562	3240	-	4767	9039	3262
<lang>	393	485	-	-	705	3747	<voice>	-	500	-	-	-	-
<m>	1	-	-	-	533	592	<xptr>	-	-	-	-	206	-
<mood>	-	675	-	-	-	-	<xr>	111	8453	127	235	28	-