Markup Enhancement: Converting CEE Dictionaries into TEI, and Beyond

Tomaž Erjavec
Tomaz.Erjavec@ijs.si
Department of Intelligent Systems, Jožef Stefan Institute
Jamova 39, SI-1000 Ljubljana, Slovenia

Nancy Ide
ide@cs.vassar.edu
Dept. of Computer Science, Vassar College
124 Raymond Avenue, Poughkeepsie, NY 12604-0520

August 24, 1999

Abstract

This paper describes the process of markup enhancement for six Central and Eastern European language dictionaries. We provide examples of the process for the English-Slovene dictionary, currently being produced by the Slovene publishing house DZS, and based on the Oxford-Hachette English-French dictionary. The TEI document type for Dictionaries is presented, followed by the process of cross-translating from original DZS SGML documents into a TEI.dictionary document and into HTML. We next discuss the development of a specialized DTD that can serve as a general model for lexical data, and provide some examples of its use.

1 Introduction

The EU project CONCEDE aims to build structured lexical databases derived from existing machine-readable dictionaries, for six languages: Bulgarian, Czech, Estonian, Hungarian, Romanian and Slovene. The project builds on the experience and resources of the MULTEXT-EAST project [2], which developed an annotated parallel corpus for the same six languages. The CONCEDE lexical databases will be integrated with this corpus; the combined results of the two projects should constitute an integrated multilingual resource of unprecedented value.

The Text Encoding Initiative (TEI, [10] provides standard SGML-based formats for a range of text types, also dictionary data [6]. CONCEDE aims to produce lexica that are compatible with the TEI scheme. Our initial plan was to extract a subset of the TEI encoding guidelines for print dictionaries. To this end, the initial step in the project was to convert 500 entries from dictionaries in each of the project languages into the TEI.dictionary base document type. The results for each dictionary were then compared, with special attemtion to problems and solutions adopted in each case [8].

The initial plan within CONCEDE was to derive a subset of the TEI guidelines for dictionary encoding as the target DTD for all six lexicons. However, after examination of the encoding problems encountered in the initial 500 entries, a more radical approach was adopted. We have designed a document type definition (DTD) based on the TEI DTD for dictionary entries that allows for the maximum of flexibility in the placement of dictionary elements, and at the same time preserves the structure of the original entry and the relations among elements implied by that structure. Because our ultimate aim is to produce lexicons suitable for use in language engineering applications, the underlying model for this DTD is based roughly on feature structures [9]. Feature structures have been heavily used computational linguistics to model grammatical information, and their applicability to structuring lexical information has been acknowledged (see, for example, [7]). Our use of this formalism should therefore provide compatibility between the lexicons produced in the project and the structure required for their use in natural language applications.

This paper overviews the design stages of the dictionary encoding process within CONCEDE. Section 2. gives some examples from the original encoding for one of the project lexicons, the Oxford-DZS English-Slovene dictionary, in the TEI.dictionary format. With multiple conversion routines, changing data, and the source encoding already in SGML, it was essential to choose a conversion process that would exploit the advantages of standardised encoding, rather than be hindered by them. Section 3. describes the conversions in more detail. Section 4 descibes the design of the CONCEDE DTD and provides examples of its use.

2 The TEI En-Sl Dictionary Entry

The Slovene language resource for the LDB is the Oxford-DZS English-Slovene dictionary, currently in production at the publishing house DZS, and based on the Oxford-Hachette English-French dictionary [1]. The digital format of the En-Sl dictionary is SGML, with a Document Type Definition based on the markup used by Oxford University Press. The DZS markup describes structural elements, e.g. <LP> for 'lemma for phrasal verb patterns', but also rendition information, e.g. <C> for 'comma'.

Five hundred entries from the DZS dictionary were initially converted from the source SGML document type into TEI, more specifically, into the base module for dictionaries *TEI.dictionary*, producing a more structured and standardised resource. Because the DZS dictionary is still in production, the conversion to TEI also functionally validates the data, making it possible to spot and correct errors before publication.

The two direct descendants in the <BODY> of the TEI.dictionary base tagset we use are the <ENTRY> and <SUPERENTRY> elements. The former encodes entries from the original dictionary; the latter groups homonimous entries. An <ENTRY> can contain one or more syntactic or semantic <SENSE>s, where the syntactic ones can, in turn, contain their semantic senses. All three levels can contain (with certain restrictions) the form of the entry or sense, its grammatical information, translations, definitions, examples, and cross-references.

```
<entry key="beyond">
  <form><orth type='hw'>beyond</orth> <pron>bI"jQnd</pron></form>
  <sense orig='syn'>
    <gramgrp><pos>prep</pos></gramgrp>
    <sense orig='sem'>
    <trans><tr>onstran, onkraj, na drugi strani, preko</tr></trans>
    <eg orig='example'>
      <quote>beyond the city walls</quote>
      <tr>onstran mestnega obzidja</tr></eg>
...
```

```
<entry key="bias ply tyre" type='compound'>
  <form>
    <orth type='hw'>bias ply tyre</orth>
    <orth type='variant'>bias ply tire</orth>
    <usg type='label'>US</usg>
    <usg type='label'>GB</usg>
  </form>
  <gramgrp><pos>n</pos></gramgrp>
  <trans>
   <usg type='label'>Aut</usg>
   <tr>diagonalni pla&scaron;&ccaron;</tr>
  </trans>
</entry>
```

The situation is complicated by various usage indicators, which are encoded as elements or (TYPE) attribute values. Verbal entries have in the original format an especially rich structure, with the markup e.g. distinguishing idioms and phrasal verb patterns. This information has been again retained either in elements (e.g. an idiom block is encoded as a <SENSE TYPE="IDIOMS">.

In our document instance we tried to retain all the information from the original format, thus making the converison from the TEI.dictionary encoding back to the original possible at least in principle. However, this policy leads to rather heavy use of the TYPE attribute.

3 Dictionary Conversion

The cross-conversion from the original to TEI.dictionary means going from documents described by one SGML DTD into those of another. We spent considerable time choosing the right tool for this job, trying to balance ease of use, expressive power and availability. The program we settled on is OmniMark LE, the 'light edtition' of OmniMark$^{(R)}$, available from *http://www.omnimark.com/*. While Omnimark is a commercial product, the LE incarnation is available free of charge; LE is identical to the commercial version, except for the restriction that programs cannot have more than 200 'countable actions'. So far, this not been an obstacle; the current conversion from the DZS DTD into the TEI.dictionary DTD has 44 such actions.

The conversion then proceeded in identifying, in turn, the semantics of each of the DZS elements in TEI.dictionaries and implementing the mapping. The DZS DTD defines 46 elements, some of them with quite complicated content models. The conversion therefore did not proceed only from the DTD but also took into account the actual usage of patterns in the source dictionary.

Two types of conversions were necessary; the simpler one is a context-dependent renaming of elements, as in the following Omnimark SGML translation action:

```
element GR when parent is (O2 | O3)
    output "<lbl type='gram'>%c</lbl>%n"
```

More complex are conversions that need forward reference and, in a sense, add new structure to the document. An example is the <SUPERENTRY>

mentioned above: the original document marks homonims only inside the headword element, e.g. `<hw>like<hm>1</hm></hw>`. But the <SUPERENTRY> tag must be output before the start of the <ENTRY>. These cases can be solved by postponing the output until the necessary information becomes available.

We also implemented two conversions to HTML. The first is from the original, and tries to imitate the appearance of the printed dictionary, but with additional use of colors. The second is from the TEI.dictionary and formats the entry giving the descriptive names of tags, e.g. from the tag <GRAMGRP> we get 'Grammar group' in English and 'Slovnično gnezdo' in Slovene. Browsing on these formats gives, on the one hand, a feel for the finished product, and, on the other, an expanded, easy to understand standardised encoding. Both help in visualising the data and can be used as validation aids, both in the lexicographic process and in the task of LDB creation.

4 The Concede DTD

As could be expected, the initial encoding experiment for the 500 entries revealed considerable variation among the structures and elements of the CONCEDE dictionaries. Although we originally intended to use the structured TEI scheme for encoding dictionary entries, several partners used the TEI "entry free" alternative, which allows for placement of dictionary elements at any point within an <ENTRY> tag. Despite these variations, certain underlying regularities exist in all of the dictionaries, in particular, the use of a hierarchical organization that enables the factoring of information over nested levels. Although the nesting arrangement of levels in the hierarchy is not consistent across dictionaries, the use of a hierarchy to avoid re-specification of common information is virtually universal (for a discussion, see [6]).

It has been shown that all of the levels in dictionary hierarchies potentially contain the same elements [6]. There is no need, therefore, to have a proliferation of structural tags (i.e., tags marking levels in the hierarchy) which would have the same definition in the DTD. Instead, the CONCEDE DTD includes a general entry division element, <STRUC> whose name is deliberately chosen to be neutral. The <STRUC> element is used to designate structural divisions within entries, such as divisions into homographs, etc., as well as to bracket associated sets of information, most notably of three kinds: information about the "forms" of the headword (pronunciation, inflected forms, hyphenation, etc.), grammatical information (part of speech, person, number, etc.), and sense information (grouped subsenses, etc., as well as information typically associated with a given sense (usage, domain, definition, translation, etc.). Any of a set of "atomic" dictionary tags can appear within the <STRUC> element, which include orth, pron, hyph, syll, stress, pos, gen, case, number, tns, mood, usg, time, register, geo, domain, style, def, eg, etym, xr, trans, itype (see [10] for a description of these tags). For example:

```
<struc type=entry>
  <orth>demigod</orth >
  <pron>'dEmI,god<pron/>
  <pos>n</pos></struc>
  <struc type=sense>
```

```
    <struc type=subsense>
      <def>a being who is part mortal, part god.</def></struc>
    <struc type=subsense>
      <def>a lesser deity.</def></struc></struc>
  <struc type=sense>
    <def>a godlike person.</def></struc>
</struc>
```

This basic structure defines a hierarchy that can be visualized as a tree, with a node corresponding to each level. Atomic tags indicate attributes (features) associated with that node; tag content provides the values for those features. Feature/value pairs at any node apply to all subtrees rooted at that node. Thus the encoding above can be rendered as follows:

```
[entry]  ---------|--- [sense]
orth : demigod    |       |
pron : dEmI,god   |       |---[subsense]
pos  : n          |       |    def : a being who is part...
                  |       |
                  |       |---[subsense]
                  |            def : a lesser deity
                  |
                  |--- [sense]
                       def : a godlike person
```

By traversing the tree either from the top-most node to a given terminal (or the reverse) and accumulating the information associated with each node visited during this traversal, all of the information associated with a particular *sense* of the head word (i.e., the word associated with the root node of the tree) is acquired. Thus, children of any node function as disjuncts in the feature structure formalism. We extend this formalism to allow for overriding, a frequent phenomenon in dictionary entries; in particular, when information incompatible with that given at a node higher in the tree is found (e.g., if a feature is respecified with a different value), only the information at the innermost node is retained.

The CONCEDE DTD also provides an <ALT> element that is used to designate alternates at any node. For example, the following renders a portion of the En-Sl example given in Section 2 using the CONCEDE DTD:

```
<struc type=entry key="bias ply tyre" etype='compound'>
  <orth>bias ply tyre</orth>
  <usg type='geo'>GB</usg>
  <alt>
    <orth>bias ply tire</orth>
    <usg type='geo'>US</usg>
  </alt>
  <pos>n</pos>
  <usg>Aut</usg>
  <trans>diagonalni pla&scaron;&ccaron;</trans>
</struc>
```

This corresponds to the following structure:

```
[entry]  . . . . . . . . . . . . . [alt]
```

```
orth : bias ply tyre                orth : bias ply tire
geo  : GB                           geo  : US
pos  : n
usg  : Aut
trans: diagonalni pla&scaron;&ccaron;
```

The dotted line indicates that [alt] is not a child of the node labelled [entry], but rather that it provides a set of alternative information. When traversing the tree to gain information about a specific use, if the [alt] information is utilized it overrides the corresponding feature/value pairs at the [entry] node. This is equivalent to providing two separate constructs:

```
[entry]
orth : bias ply tyre
geo  : GB
pos  : n
usg  : Aut
trans: diagonalni pla&scaron;&ccaron;
```

and

```
[entry]
orth : bias ply tire
geo  : US
pos  : n
usg  : Aut
trans: diagonalni pla&scaron;&ccaron;
```

We are still in the process of finalizing the CONCEDE DTD. However, this general overview provides an outline of its major features. When complete, the CONCEDE DTD will be incorporated into the Corpus Encoding Standard (CES) [4], [5].

5 Conclusion

This paper describes two stages in the development of an encoding scheme suitable for information extracted from everyday dictionaries that is intended ultimately for use in language engineering applications, in the context of the CONCEDE project. We have utilized the TEI guidelines as an intermediate DTD for encoding six Central and Eastern European language dictionaries. We have developed a DTD intended to provide the target structure for the data extracted from these dictionaries, which will render it maximally compatible with other natural language resources. To verify this possibility, CONCEDE aims to integrate its TEI dictionaries with information from an aligned and morphosyntactically annotated English original of Orwell's '1984' [3] concordancing at *http://nl2.ijs.si/corpus/* and its translations into the six project languages.

Acknowledgements

This work was supported in part by the EU project Copernicus Concede and by the associated grant from the Ministry of Science and Technology of Slovenia.

References

[1] Marie-Hélène Corréard and Valerie Grundy, editors. *Oxford-Hachette French Dictionary.* 1994.

[2] Ludmila Dimitrova, Tomaž Erjavec, Nancy Ide, Heiki-Jan Kaalep, Vladimír Petkevič, and Dan Tufiş. Multext-East: Parallel and Comparable Corpora and Lexicons for Six Central and Eastern European Languages. In *COLING-ACL '98*, pages 315–319, Montréal, Québec, Canada, 1998.

[3] 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. URL: http://ceres.ugr.es/ rubio/elra.html.

[4] Nancy Ide and Greg Priest-Dorman. *The Corpus Encoding Standard*. 1996. URL: http://www.cs.vassar.edu/CES/.

[5] Nancy Ide. Corpus Encoding Standard: SGML Guidelines for Encoding Linguistic Corpora. In Antonio Rubio, Natividad Gallardo, Rosa Castro, and Antonio Tejada, editors, *First International Conference on Language Resources and Evaluation, LREC'98*, pages 463-70, Granada, 1998. ELRA.

[6] Nancy Ide and Jean Véronis. *Encoding Dictionaries*, pages 167–180. Kluwer Academic Publishers, Dordrecht, 1995.

[7] Nancy Ide, Jacques Le Maitre, and Jean Véronis. Outline of a Model for Lexical Databases. *Current Issues in Computational Linguistics: In Honour of Don Walker. Linguistica Computazionale* IX, X, pages 283-320, Pisa, 1995. [reprinted from *Information Processing and Management*, 29, 2, pages 159-186]

[8] Adam Kilgarriff, editor. *Concede Deliverable 2.2: Dictionary Encoding Schemes*. University of Brighton, 1999.

[9] Stuart Shieber. *An Introduction to Unification-based Approaches to Grammar*. CSLI Lecture Notes Series, University of Chicago Press, Chicago, 1986.

[10] C. M. Sperberg-McQueen and Lou Burnard, editors. *Guidelines for Electronic Text Encoding and Interchange*. Chicago and Oxford, 1994.