The Concede Model for Lexical Databases

Tomaž Erjavec,* Roger Evans,† Nancy Ide,‡ Adam Kilgarriff†

* Dept. for Intelligent Systems, Institute Jožef Stefan
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 paper describes the approach we have taken in the Concede project, which aims to develop compatible lexical databases for six Central and Eastern European languages. Starting with sample entries from original presentation-oriented electronic representations of dictionaries, we transformed the data into an intermediate TEI-compatible representation to provide a common baseline for evaluating and comparing the dictionaries. We then developed a more restrictive encoding, formalised as an XML DTD with a clearly-defined semantic interpretation. We present this DTD and discuss a sample conversion from TEI, together with an application which hyperlinks a HTML representation of the dictionary to on-line concordancing over a corpus.

1. Introduction

The EU INCO-COPERNICUS project CONCEDE (Consortium for Central European Dictionary Encoding) aims to build structured lexical databases (LDBs) 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 (Dimitrova et al., 1998). 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 undesirable to lose useful information (content or structure) from the original representations.

To date, we have concentrated on dictionary samples comprising approximately 500 entries from each of the CONCEDE dictionaries, chosen on the basis of frequency lists derived from the MULTEXT-East corpus. A detailed description of the sampling process is given in (Erjavec et al., 1999).

This paper is structured as follows: Section 2 describes the TEI-encoded samples of the CONCEDE dictionaries and overviews the sizes and element usage in the samples. Section 3 introduces the CONCEDE LDB format and details the notion of inheritance in the LDB. Section 4 overviews a sample conversion from the TEI to the LDB format, giving the steps involved and the information content of the resulting LDB in comparison to the TEI version. Section 5 then utilises both the TEI and LDB versions to produce an HTML representation of the dictionary integrated with a corpus querying system. Finally, Section 6 gives conclusions and direction for further work.

2. The TEI Encoding of Dictionaries

The Text Encoding Initiative, TEI, provides SGML-based guidelines for encoding Print Dictionaries (Sperberg-McQueen and Burnard, 1994, Chapter 12) , 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 five dictionary samples[1] according to the TEI.dictionaries 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 TEI dictionary encoding scheme, to support richer element content models but this was done within the guidelines for such extensions. In this section we overview the resulting TEI encoded samples.

The most directly observable property in a TEI encoding is the size of the samples and the distribution of markup (SGML tags) to text (`#PCDATA`). The dictionary sample ⟨**body**⟩ elements were first (non-faithfully) converted to ISO 8879 Latin-2, where we substitute SGML entities for 8bit encodings, e.g., `č` or `×`. Table 1 summarises counts on these files and gives a comparison with the version stripped of markup.

Despite the small size of the samples, which contain about 1% of the complete dictionaries, the numbers in the Table are quite large. This underlines the view that dictionaries contain extensive and richly structured language data, which can be utilised by computational tools, and highlights the complexity of their information content: the

[1] A different strategy was followed for Bulgarian.

Lang	Entr's	TEI	Text	Markup	Elem's
cs	580	1,001	405	59.5%	40,456
et	723	2,366	1,858	21.5%	36,108
hu	520	960	374	61.0%	48,277
ro	511	1,456	703	51.7%	48,870
en-sl	615	420	160	61.9%	17,775
ALL	2.949	6,203	3,500	43.6%	191,486

Table 1: Size and Markup of TEI Samples

size of a complete TEI encoded dictionary will be in the range of 100MB, and contain several million elements.

The dictionaries samples use together 58 different elements. Most are taken directly from the TEI.dictionary tag set, with some modifications and additions. Table 2 summarises the tag usage of those elements that are used in at least two of the five samples; the first line gives the total number of elements used in the sample. Where the TEI element definition has been modified in the local extensions it is flagged by an asterisk.

Element	**cs**	**et**	**hu**	**ro**	**en-sl**
58 =	31	20	32	33	18
⟨**entry**⟩	580	723	*520	*511	615
⟨**sense**⟩	*3323	5368	5607	*9665	891
⟨**form**⟩	*4221	723	1339	*5504	873
⟨**orth**⟩	2998	723	1333	5974	906
⟨**gramgrp**⟩	*1600	721	1558	2875	714
⟨**pos**⟩	1618	685	*1407	1344	838
⟨**def**⟩	*5804	6838	*6677	*8339	0
⟨**usg**⟩	1774	980	*3005	2448	852
⟨**eg**⟩	2851	5291	*5414	956	1873
⟨**q**⟩	*6994	0	6212	1023	0
⟨**quote**⟩	0	7816	0	0	1903
⟨**re**⟩	1023	0	780	*3478	0
⟨**xr**⟩	0	32	13	158	87
⟨**ref**⟩	2059	0	16	261	87
⟨**ptr**⟩	0	0	33	9	0
⟨**lbl**⟩	0	0	1217	1053	159
⟨**pron**⟩	18	0	11	32	256
⟨**etym**⟩	33	0	355	488	0
⟨**lang**⟩	38	0	194	495	0
⟨**mentioned**⟩	0	0	292	542	0
⟨**gloss**⟩	0	0	19	25	100
⟨**subc**⟩	0	15	473	858	0
⟨**itype**⟩	0	514	0	132	0
⟨**number**⟩	0	39	0	650	0
⟨**case**⟩	1006	0	0	71	0
⟨**gen**⟩	670	0	0	787	0
⟨**mood**⟩	99	0	0	132	0
⟨**per**⟩	95	0	0	134	0
⟨**tns**⟩	3	0	0	141	0

Table 2: Tag Usage

The elements in the first part of the table comprise the core tagset, which share a number of common characteristics. They are used by all the dictionary samples, except for ⟨**q(uote)**⟩, for which see below. They are also high-frequency elements, and thus represent a significant part of the information content of the dictionaries: knowledge extraction techniques could benefit most by concentrating on these elements. Finally, almost all sample-particular TEI modifications are applied to the core tagset. The modifications point to areas of the TEI.dictionaries base tagset which may need revision or clarification. For example, most of the dictionaries required modification of the TEI.dictionaries to allow ⟨**lbl**⟩ to appear in the class `%dictionaryTopLevel`. Another example is the complementary distribution of ⟨**q**⟩ and ⟨**quote**⟩: both tags are used for the same purpose, and the TEI.dictionaries DTD could give guidance on which is the preferred choice.

The last part of the table reveals the morphological complexity of the project languages, since all of the elements are devoted to qualifying inflectional characteristics of the (head-)words. This also represents the information most immediately useful for linking into a corpus tagged for morpho-syntax, or for harvesting dictionaries to increase the lexical coverage of analysers.

2.1. Local elements

Table 3 gives those elements for each language that appear only in the sample in question. As before, "*" marks modified elements, while "+" marks elements that have been added to TEI.dictionaries by local extensions.

The lines are sorted by usage, and often reveal dictionary specific information, e.g., the bibliographical references in Estonian or ⟨**(tr)ans**⟩ in the English-Slovene sample. More interesting are the new elements in local extensions.

The Estonian and English-Slovene opted for using 'vanilla' TEI.dictionaries, with no modification. This approach has the advantage that both parse according to the same DTD and can therefore both be parts of one SGML document; the disadvantage is that certain elements which are explicitly tagged in the other dictionaries appear only as the value of an attribute in the non-modified TEI encoding. For example, the Czech and the English-Slovene digital originals already encode the information that some structure is an idiom, and the conversions attempted to preserve this fact. The Czech sample extended TEI with ⟨**idiom**⟩, while the English-Slovene uses a more generic ⟨**sense**⟩ with the attribute **type='idiom'**. Attribute values present certain processing difficulties for further conversion, and have ramifications for the semantics of our final representation.

Modifications of existing TEI elements change the overall dictionary structure in various ways. It remains to be seen whether it is possible to construct a DTD that would be a parametrisation of TEI, while at the same time parsing the default TEI.dictionaries. Nevertheless, as SGML (unlike XML) applications require a DTD to process all the samples as one document, we made a DTD which validates all the samples. It specifies exactly the elements and attributes used in the samples, and identifies the `EMPTY` ones, but imposes no structure on the elements, i.e., it specifies their content models with the keyword `ANY`.

cs		et		hu		ro		en-sl	
+⟨**idiom**⟩	1365	⟨**bibl**⟩	2786	⟨**oref**⟩	6749	⟨**stress**⟩	556	⟨**tr**⟩	4323+9hu
⟨**num**⟩	1098	⟨**cit**⟩	2786	+⟨**expr**⟩	1915	⟨**colloc**⟩	139	⟨**trans**⟩	3127+9hu
⟨**label**⟩	589+2sl	⟨**gram**⟩	56	+⟨**exprgrp**⟩	1468	⟨**superentry**⟩	42+1sl	⟨**hom**⟩	168
+⟨**asp**⟩	292	⟨**hyph**⟩	12	+⟨**opt**⟩	1180	⟨**term**⟩	37		
+⟨**ant**⟩	202			*⟨**abbr**⟩	369	⟨**m**⟩	11		
+⟨**voice**⟩	63			⟨**ovar**⟩	81				
+⟨**deg**⟩	34			⟨**seg**⟩	20				
⟨**name**⟩	5								

Table 3: Sample Specific TEI Elements

2.2. TEI attributes

A significant part of the information content, esp. if using (non-modified) TEI, lies in the element attributes. The CONCEDE TEI samples use 10 different attributes; Table 4 shows how many times each occurs in the samples.

Attr.	cs	et	hu	ro	en-sl	Σ
type	5388	2481	3425	7532	2159	20985
n	4680	5222	4109	3387	1	17399
rend			28	1318		1346
orig			350		619	969
key					616	616
id			156	84		246
target			33	79		112
lang				37		42
prev			21			21
extent			16	3		19

Table 4: Attribute Usage

Most attributes result from the 'legacy preserving' approach, where dictionary-specific information is encoded in **type**, **rend** and **orig** attributes. This is potentially troublesome for further automatic conversion, since, e.g., , the **type** attribute distinguishes 83 different values in the different samples, with little overlap. Other attributes are used to count and provide identifiers (`IDs`) for various elements. The **lang** and **extent** attributes fall in neither of these categories.

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) (W3C, 1998), 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), (W3C, 2000).

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

For structural elements, we follow the observations in (Ide and Véronis, 1995) that certain underlying regularities exist in all print dictionaries (in particular, the use of a hierarchical organisation 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 three structural elements:

- ⟨**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).[2]
- ⟨**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.

The DTD itself and a fuller version of its documentation is available in (Kilgarriff, 1999).

3.1. Inheritance

We wish to say that the following information package

```
<struc>
  <content_tag>content-1</content_tag>
  <struc>sense-1-content</struc>
  <struc>sense-2-content</struc>
</struc>
```

[2]XML documents are also described as trees, where the "parent" of a given element is the element in which it is immediately enclosed. To avoid confusion, we use the term "tree" and the associated terminology to refer only to the structures outlined in section 3.

is informationally equivalent to the following three packages:

```
<struc>
  <content_tag>content-1</content_tag>
</struc>
<struc>
  <content_tag>content-1</content_tag>
  <struc>sense-1-content</struc>
</struc>
<struc>
  <content_tag>content-1</content_tag>
  <struc>sense-2-content</struc>
</struc>
```

We clarify the notion of informational equivalence as follows. We take a dictionary entry as a description of the events of the word being used. Each time the word is spoken or written, this is an event.[3] We treat a content element as a proposition which may be true or false of an event. Thus ⟨**pos**⟩*n*⟨**/pos**⟩ is true of just those word-events which have part of speech *n*.

For simple entries with no hierarchical or alternation structure, all the events are described by all the elements of the entry: for a single-sense word with ⟨**orth**⟩ *o*, ⟨**pos**⟩ *p* and ⟨**def**⟩ *d*, for all events of the word being used, we take the dictionary to be asserting that the orthography is *o*, the part-of-speech is *p* and definition is *d*.[4]

For every subdivision of an entry, an event may or may not be assigned to the subdivision, according to whether all the information items associated with the subdivision are true for the event. Thus a dictionary entry defines a partition (not necessarily exclusive or exhaustive) of the events for the word. For a simple two-sense word, we expect most events to be associated either with the one sense or the other, and all, by definition, to be associated with the top level of the entry. For simple cases, a satisfactory dictionary entry is one where most events are assigned to one and only one of the leaf nodes of the entry.

Two dictionary entries are informationally equivalent if they define the same partition of the set of events.

An alternative graphical notation takes nodes as ⟨**struc**⟩ elements, and arcs as inheritance links between ⟨**struc**⟩ (or ⟨**entry**⟩) and child ⟨**struc**⟩ elements. Content elements are shown at the nodes. Thus the two graphs in Figure 1 are informationally equivalent.

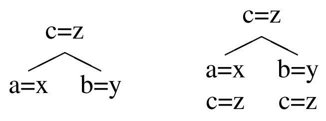

Figure 1: Two informationally-equivalent dictionary entries

[3]There are of course many problems with the identification of such events — if I copy a text, is that a new event or the same one again? Also written events do not have pronunciations, or spoken texts, orthographies, and so on; these issues lie outside the scope of the project.

[4]Inflectional morphology is addressed below.

3.2. Conjunctive/Disjunctive

Content elements occurring as children of a structure tag are interpreted conjunctively, i.e., all sibling information items are true of all the events assigned to that subdivision of the entry. Thus

```
<struc>
  <dom>nautical</dom>
  <pos>n</pos>
</struc>
```

asserts that all events specified in this part of the entry are both in the nautical domain, and nouns. The following

```
<struc>
  <def>def1</def>
  <geo>Canada</geo>
  <geo>Australia</geo>
</struc>
```

is anomalous (given the interpretation of ⟨**geo**⟩ as 'spoken in this geographical area') as an event cannot ordinarily be spoken in both Canada and Australia. Assuming the dictionary meaning was "people in Canada and Australia say this", the markup is incorrect.

Sibling ⟨**struc**⟩ elements are interpreted disjunctively and the appropriate markup would be

```
<struc>
  <def>def1</def>
  <struc>
    <geo>Canada</geo>
  </struc>
  <struc>
    <geo>Australia</geo>
  </struc>
</struc>
```

The notions of "conjunctive", "disjunctive", and "being true for a given event" do not apply to examples. For some further information types, "being true for a given event" will not be applicable to all events, e.g., a spoken event does not have an ⟨**orth**⟩ and a written event does not have a ⟨**pron**⟩.

There may be cases (in addition to examples) where two sibling elements have the same tag and the conjunctive reading is intended, as in the reading of

```
<struc>
  <domain>nautical</domain>
  <domain>law</domain>
</struc>
```

where the intention is to assert that the item is in the domain of nautical law.

3.3. Overwriting and cumulative inheritance

Where information is inherited, there is the possibility that an item of information is specified at two different nodes from which a node may inherit.[5] The question then is, from where do we inherit? The logical possibilities are that both values apply at the child; that neither do; or that just one does. In CONCEDE, we define the options as follows:

[5]To simplify the exposition, we include the node itself as one of the nodes from which it may inherit.

- Cumulative inheritance for elements that may take more than one value and are inherited and combined along the dictionary structure.
- Overwriting inheritance for elements that take only one value at a time. This implies that only one instance of the element may appear at a given node and that it is propagated along the dictionary structure unless and until a new value is specified for that element. In such a case, the new value "overwrites" the earlier one and is subsequently propagated to nested structures.

We block inheritance altogether for elements whose values are associated only with the structural node within which they appear, i.e., they are not propagated through the structure. Cross-references provide a good example, since they apply only to the level of description with which they are directly associated.

There are various options for stating that a particular sense is, e.g., used only in the US. The TEI name for geographical regions is ⟨**geo**⟩. The possibilities offered by the TEI.dictionaries include presenting it as an attribute-value pair on the ⟨**struc**⟩ tag

```
<struc geo="US"> ...
```

or as an attribute-value pair on a possibly empty ⟨**usg**⟩ element

```
<struc>
  <usg geo="US"> ...
```

or as the content of a ⟨**usg**⟩ element of type **geo**

```
<struc>
  <usg type="geo">US</usg> ...
```

or as the content of a GEO element

```
<struc>
  <geo>US</geo> ...
```

The choice made here interacts with the inheritance device. We want to be able to specify that a geographical specification on a subsense overwrites what might be stated at a main sense, but that, for example, register information at the subsense would not overwrite geographical information at the main sense. For it to be clear what to inherit, and what to overwrite, it must be clear what the mapping is from SGML elements, attributes, values and content to feature-value notation features and values.

For CONCEDE, we choose to equate element names to features, and element content to values. SGML attributes do not play a role in the inheritance. This choice is associated with a second choice: any two information-types where overwriting would be inappropriate become separate elements, and if it is appropriate that one item overwrites another, then they should be different content for the same element.

Thus ⟨**register**⟩ and ⟨**geo**⟩ are elements and the correct representation for *US* is

```
<struc>
  <geo>US</geo> ...
```

We have implemented the inheritance mechanisms specified above using the XSL Transformation language (XSLT), (W3C, 1999). A fuller description of our XSLT implementation is found in (Ide et al., 2000).

3.4. Multiwords

In many dictionaries, there are subentries, or definitions, or glosses, or further specifications of other kinds for the headword where it occurs within a certain phrase, idiom, collocation or other "multiword". (A multiword is simply an item that, for some purpose or other, might be treated as a lexical unit, but which is spelt with spaces in). The CONCEDE treatment is simply to introduce a new ⟨**struc**⟩ for the multiword containing an ⟨**orth**⟩ element with the multiword as its content along with any other specifications that the dictionary gives for the multiword.

This involves a measure of 'levelling': each dictionary will have its own taxonomy of multiwords, and will treat different classes differently. In general, the distinctions between categories such as "phrase" and "idiom" are hard to draw, particularly across languages, so we preferred not to use different tags where it would be so hard to be consistent in their use, across dictionaries and languages.

3.5. Morphology

Morphology is not treated properly in these proposals. Dictionaries standardly state what morphological paradigm a word belongs to and give fuller specifications only when there are irregularities, or where particular senses of the word are constrained to particular forms of the paradigm, or use non-standard forms. Thus inheritance and overwriting play an important role. But the inheritance can only be interpreted if there is a morphological engine which interprets a table of inflections to give the full paradigm. Where there is such an engine available, the treatment will be along the lines illustrated below:

```
<entry>
  <hw>sing</hw>
  <pos>v</pos>
  <alt>
    <brack><itype>base</itype>
       <orth>sing</orth>   </brack>
    <brack><itype>3rdsgpres</itype>
       <orth>sings</orth>  </brack>
    <brack><itype>prespart</itype>
       <orth>singing</orth></brack>
    <brack><itype>part</itype>
       <orth>sang</orth>   </brack>
    <brack><itype>pastpart</itype>
       <orth>sung</orth>   </brack>
  </alt>
  <struc> ... </struc>
</entry>
```

The paradigm will be associated with the ⟨**entry**⟩ node for the word by an ⟨**alt**⟩ link. The lexicon would then represent a large set of 'ground instances': in the simple case, one for each <meaning, morphological-paradigm-member> pair. Such a framework is required for it to be possible to interpret lexical information such as "meaning X does not occur with paradigm-member Y".

4. Sample Conversion into the Concede DTD

To date we have tested the CONCEDE LDB format on one of the TEI encoded samples, the bi-lingual English-Slovene dictionary. The translation is, to a great extent, automatic with little manual intervention.

The resulting LDB contains most of the content from the original, but omits about 10% of elements which we currently cannot yet exploit and have, at the same time, the most difficult (inconsistent) placement and scoping. The En-Sl sample was thus converted to a well-formed and valid CONCEDE LDB which, however, preserves only the more basic dictionary information.

The conversion process 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 or patterns not taken into account by the program. 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 the final LDB format as further work.

The first step in the conversion of the TEI source to the CONCEDE LDB encoding involves simplification and renaming of the TEI encoding. We used the following transformations on the TEI sample:[6]

- Suppress `#PCDATA` in ⟨**form**⟩, ⟨**gramgrp**⟩, ⟨**sense**⟩, ⟨**trans**⟩, ⟨**eg**⟩
  A number of (mostly) structural TEI elements allow text to appear in them directly; this text is usually punctuation and graphical marks, whose meaning has, for the most part, been explicated in the markup. The conversion suppresses such `#PCDATA`, simplifying them to element content only.

- Remove tags for ⟨**superentry**⟩, ⟨**form**⟩, ⟨**gramgrp**⟩, ⟨**xr**⟩
  The TEI nests elements deeper than the flatter LDB format; this step reduces the spurious embedding of elements

- Rename tags ⟨**orth type=hw**⟩ → ⟨**HW**⟩, ⟨**tr**⟩ → ⟨**ORTH**⟩, ⟨**hom**⟩ & ⟨**sense**⟩ → ⟨**STRUC**⟩, ⟨**quote**⟩ → ⟨**Q**⟩, ⟨**ref**⟩ → ⟨**XR**⟩
  Apart from renaming tags, this step also introduces the **TYPE** attribute and assigns it a value, e.g., **idiom** or **hom**.

- Suppress elements ⟨**usg**⟩, ⟨**lbl**⟩, ⟨**label**⟩, ⟨**gloss**⟩
  While (equivalents of) these elements exist in the CONCEDE DTD, they exhibit the most complicated placements and scopings; they can appear before, after or in the middle of an element. The encoding also exhibits a presentation oriented format, difficult to resolve accurately. At the same time, these elements contain information that we currently do not yet exploit computationally.

- Copy over tag ⟨**trans**⟩
  This element retains its semantics.

- Copy attributes **ID** (and **LANG**)
  This information is important to keep a link between the TEI elements and those of the LDB. Arguably, the ID attribute would better be translated into some other type of reference.

The above transformation produces a document with the correct tags and mostly correct nestings but which is not yet semantically well-formed; it lacks the ⟨**ALT**⟩ elements. At this point, an element can still contain, interspersed within it, a number of elements with the same general identifier, i.e. tag name. This is not allowed in the LDB model, where alternatives must be nested in ⟨**ALT**⟩.

To this end the second step of the conversion orders sibling elements ⟨**ORTH**⟩, ⟨**PRON**⟩, ⟨**POS**⟩, and ⟨**Q**⟩ and groups them into ⟨**ALT**⟩. While most elements are indeed order independent there are exceptions, e.g., irregular inflections of English headwords have alternating sibling ⟨**ORTH**⟩ and ⟨**PRON**⟩ elements. This step therefore loses information from the TEI encoding.

These two steps produce the sample encoded in the CONCEDE LDB. To exemplify, we give below one entry in the input TEI and output CONCEDE encodings.

```
<entry key="bezant">
  <form>
    <orth type='hw'>bezant</orth>
    <orth type='variant'>bezzant</orth>,
    <orth type='variant'>byzant</orth>
    <pron>"beznt</pron>,
    <pron>bI"z&amp;nt</pron>
  </form>
  <gramgrp><pos>n</pos></gramgrp>
  <sense>
    <trans>
      <tr>bizantinec</tr>,
      <tr>bizantinski zlatnik</tr>
    </trans>
  </sense>
  <sense>
    <trans><usg type='label'>Archit</usg>
      <tr>medaljon</tr>
      <gloss>ornament v obliki okrogle
plo&scaron;&ccaron;e</gloss>
    </trans>
  </sense>
  <sense>
    <trans><usg type='label'>Herald</usg>
      <tr>zlat krog</tr></trans>
  </sense>
</entry>
```

```
<ENTRY>
  <HW>bezant</HW>
  <ALT type=PRON>
    <PRON>"beznt</PRON>
    <PRON>bI"z&amp;nt</PRON>
  </ALT>
  <ALT type=ORTH>
    <ORTH type=variant>bezzant</ORTH>
    <ORTH type=variant>byzant</ORTH>
  </ALT>
  <POS>n</POS>
  <STRUC TYPE=sense>
    <TRANS>
      <ALT type=ORTH>
        <ORTH>bizantinec</ORTH>
```

[6]For clarity the TEI tags are written in lower/mixed case, and the Concede tags in upper case.

```
      <ORTH>bizantinski zlatnik</ORTH>
    </ALT>
  </TRANS>
</STRUC>
<STRUC TYPE=sense>
  <TRANS>
    <ORTH>medaljon</ORTH>
  </TRANS>
</STRUC>
<STRUC TYPE=sense>
  <TRANS>
    <ORTH>zlat krog</ORTH>
  </TRANS>
</STRUC>
</ENTRY>
```

In the current version the conjunctive ⟨**BRACK**⟩ elements have not been used, but would have to be if labels were included into the conversion. Appropriate usage of ⟨**BRACK**⟩ would also prevent the loss of information as e.g., exhibited above, where the two variant spellings should the grouped with their respective pronunciations.

4.1. Information content of the LDB

It is difficult to compare and even define the information content of a dictionary and a lexical database, but quantities offer an approximation. A comparison between the filesizes of the TEI and CONCEDE encoded sample shows a reduction of 11% in filesize, and almost the same reduction in the total number of elements used. Table 5 gives tag-counts for the TEI encoded and the CONCEDE encoded En-Sl sample.

TEI		Concede	
⟨**pron**⟩	256	⟨**HW**⟩	615
⟨**pos**⟩	838	⟨**ORTH**⟩	4613
⟨**trans**⟩	3127	⟨**PRON**⟩	255
⟨**eg**⟩	1873	⟨**POS**⟩	838
⟨**quote**⟩	1903	⟨**TRANS**⟩	3043
⟨**ref**⟩	87	⟨**EG**⟩	1868
⟨**xr**⟩	87	⟨**Q**⟩	1899
⟨**hom**⟩	168	⟨**XR**⟩	87
⟨**sense**⟩	891		
⟨**tr**⟩	4323		
⟨**gramGrp**⟩	714		
⟨**gloss**⟩	100		
⟨**usg**⟩	852		
⟨**lbl**⟩	159		
⟨**label**⟩	2		
15	17774	**8**	15846

Table 5: Tagcounts on TEI and CONCEDE LDB

In comparison with the TEI dictionary encoding, the CONCEDE LDB format is somewhat less informative, and contains a percentage of conversion errors. However, it has a much simpler content model and a defined inheritance structure, making it easier for applications to exploit the dictionary information.

5. Sample application

The CONCEDE project also addresses the integration of machine-readable dictionaries and lexical databases with corpora. As a demonstration of using the TEI and CONCEDE LDB formats of the English-Slovene sample, we have converted some TEI dictionary entries into HTML, which are hyperlinked, via the LDB, to an on-line concordancing system.[7]

The corpus used in the experiment is the English-Slovene parallel part of the MULTEXT-East corpus, which contains approx. 100.000 words in each language. The corpus is aligned at the sentence level and tokenised, where each word is annotated with its lemma and PoS tag. The on-line query system has as its corpus processing backend the CQP system (Christ, 1994), which incorporates a powerful query language that allows querying for all of the above annotation.

For the experiment we identified 29 dictionary entries, whose headwords do in fact appear in the corpus. We then produced a Web rendering of these entries where it is possible to click on elements, causing the retrieval of associated bi-lingual concordances. The queries are automatically constructed according to the (inherited) information available for the element in question.

The envisaged application of such an intergration of a dictionary representation with corpus evidence is to either use it in the process of making the dictionary, or to give to the end-user of the dictionary a means to further supplement dictionary examples.

The process of converting to the HTML dictionary/concordance representation first involved choosing suitable anchor elements for querying; these are ⟨**orth**⟩/⟨**HW**⟩&⟨**ORTH**⟩, ⟨**pos**⟩/⟨**POS**⟩, ⟨**tr**⟩/⟨**ORTH**⟩, ⟨**quote**⟩/⟨**Q**⟩. The elements were given IDs in the TEI encoding, which persist into the LDB format.

Queries are then constructed for each anchor element, taking into account the information in the anchor as well as that inherited from superordinate anchors.[8] The type of query is dependent on the anchor element. Headwords, ⟨**hw**⟩*hw*⟨**/hw**⟩, are translated into the query `[lemma="hw"]`, meaning "find the lemma string *hw* of a token in the (default) English part of the corpus". Anchoring to the part-of-speech, ⟨**pos**⟩*pos*⟨**/pos**⟩ is somewhat more complicated: `[lemma="hw" & ice="ice(pos)"]`. Here the *pos* is mapped from the dictionary typology, e.g., *vtr* into the ICE (International Corpus of English) tagset, used in our '1984' corpus, e.g., `V(.*,.*)`. Anchoring on orthography (of, say an idiom) takes into account the inherited headword and, to some extent, *[sb]/[sth]* patterns. So, for example, ⟨**orth**⟩*catch up with [sb]*⟨**/orth**⟩ translates into `[lemma="catch"] "up" "with" ".+"`. Finally, the translations (⟨**orth**⟩ when parent is ⟨**trans**⟩) trigger the queries also on the Slovene part of the corpus. For example, *razred* as a translation for *category* translates into `[lemma="category"] :SL [lemma="razred"]`.

After such a query is constructed for each potential anchor, the query is run off-line to ensure hits in the corpus. If concordances are found, the query URL is hyperlinked

[7] This page is available at *http://nl.ijs.si/telri/Bratislava/cnc-mte.html*

[8] We use only overwriting inheritance in this experiment.

to the HTML rendering of the TEI element bearing the appropriate ID. In our 29 dictionary entries there were 2870 potential anchor elements, of which 337 produced matches in the corpus.

The HTML encoding, in spite of the small size of the corpus, demonstrates how to exploit the LDB data and presents a method of visually combining corpus searches with information encoded in dictionaries.

6. Conclusions

The paper described the TEI-encoded dictionary samples of the CONCEDE project and the CONCEDE LDB format and its notion of inheritance. It also detailed a sample application, where the TEI dictionary entries are converted to the LDB and the TEI and LDB formats are used to integrate a representation of the dictionary with a corpus querying system.

Further work....

7. References

Christ, Oliver, 1994. A modular and flexible architecture for an integrated corpus query system. In *Proceedings of COMPLEX '94: 3rd conference on Computational Lexicography and Text Research*. Budapest, Hungary. CMP-LG archive id 9408005.

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

Erjavec, Tomaž, Dan Tufiş, and 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.

Ide, Nancy, Adam Kilgarriff, and Laurent Romary, 2000. A Formal Model of Dictionary Structure and Content. In *Proceedings of EURALEX'00*. Stuttgart.

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

Kilgarriff, Adam, 1999. Generic encoding principles. CONCEDE Project Deliverable 2.1, University of Brighton, UK.

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. URL. Http://www.w3.org/TR/1998/REC-xml-19980210.

W3C, 1999. XSL transformations (XSLT) version 1.0. URL. Http://www.w3.org/TR/xslt.

W3C, 2000. Extensible stylesheet language (XSL) version 1.0. URL. Http://www.w3.org/TR/xsl.