TEI Encoding of the Hungarian Explanatory Manual Dictionary

TAMÁS VÁRADI varadi@nytud.hu

CSABA ORAVECZ oravecz@nytud.hu

Department of Corpus Linguistics,
Institute for Linguistic Research
Hungarian Academy of Sciences

Abstract: The paper presents findings of ongoing work to convert the Hungarian Explanatory Manual Dictionary to TEI compatible encoding as part of the CONCEDEproject. It discusses the need to restrict and extend the TEI guidelines to accomodate the special Hungarian lexicographic requirements. The proposed modifications to the teidict2.dtd are presented in detail. The paper highlights a number of encoding problems which arise from the difficulty to reconcile the lexical view of the dictionary with the editorial view. Conversion between the two representations are sometimes far from trivial and require manual intervention.

1 Introduction

The present paper reports on work carried out as the Hungarian contribution in the CONCEDE project (for an overview of the project see Erjavec et al., 1999). The paper is structured as follows: Section 2 describes the source data for the projected Hungarian lexical database, Section 3 introduces the encoding scheme adopted, Section 4 lists the proposed modifications to the teidict2.td and Section 6 carries some conclusions.

2 Description of the source data

The Hungarian Explanatory Manual Dictionary (HEMD) (Juhász et al., 1972) is a major academic work first published in 1972. It was intended to serve as an updated, more practical compendium to the seven volume Hungarian Explanatory Dictionary, which was completed in the 1950's. It is a general purpose comprehensive manual dictionary compiled by a team of lexicographers at the Institute for Linguistic Research of the Hungarian Academy of Sciences. The dictionary consists of cc. 70,000 headwords chiefly aimed at covering standard literary and colloqual Hungarian but also including a wide range of regiters which are thought to be part of standard Hungarian.

The entries typically have a fairly elaborate structure breaking down into the following fields: head-word, part of speech, grammatical information, sense definition, usage information, examples, phrasal constructions (including collocations, sayings, idomatic expressions, each with their own examples, usage and sense definitions), cross-references to other head-words, and etymologies. As Hungarian pronunciation is thought to be highly regular and generally presenting no problem to native speakers, pronunciation is only mentioned sporadically. The same strategy is followed with respect to grammatical information. Morphological information about how headwords are inflected are not given even for irregular cases. Instead, a concise table of verb and noun paradigms is provided in the introduction.

Both for its professional merits (depth and breadth of lexical coverage) as well as practical issues such as copyright and availability, the HEMD was chosen for conversion into a lexical database.

The HEMD is currently being revised, with the 2nd edition due to be completed by the end of 1999. The opportunity of the revision was used also to develop an electronic version of the material. However, lack of time and the lexicographers' preference for traditional paper-and-pen methodology resulted in the fact that the revision work and the transfer into machine readable form were carried out as separate processes. The revised version was prepared on paper and was handed over to a colleague for encoding. Fortunately, the electronic form of the revised version was planned with sufficent care. A provisional DTD was compiled and the text of the dictionary was to be encoded in SGML using the SGML editor WriterStation. It was planned that the DTD was to be regularly evaluated and when necessary modified to accomodate unforseen cases.

In the event, however, the encoding work was carried out in TEX through a suite of TEX macros. They were devised to be so close to the SGML format that two way conversion between TEX and SGML annotation proved a trivial task with the help of a file editor. Unfortunately, the envisaged regular updates of the provisional DTD have become less and less frequent until they have ceased to take place. Because the encoding in TEX did not involve any periodic checks on the validity of the notation, as would have been the case with the use of a proper SGML editor, the compliance of the data to the DTD became increasingly slack. It was decided, therefore, that rather than attempting to patch the provisional DTD, a new DTD would be applied with the data annotation scheme modified as necessary.

3 Selection of an encoding scheme

As one of the basic boals of the CONCEDE project is to devise and validate a TEI-conformant encoding standard, an obvious choice of an initial formalism was to consider the guidelines of the TEI Dictionary Working Group. However, the price of wide converage and generality of TEI is that it is unnecessarily large and not restrictive enough for a particular domain. Accordingly, the corresponding subset of the TEI DTD, the teidict2.dtd and the teidict2.ent entity set was adopted as a base from which to develop an encoding scheme more tailored to the data at hand. Because the dictionary data was already well structured and supplied with markup, we decided to use the <entry> elements instead of the looser <ENTRYFREE> type. Even so, it was found that the definition of <ENTRY> needed both restriction and extension at the same time.

4 Proposed modifications to the teidict2.dtd

4.1 Introducing the <expr> and <ExprGrp> elements

A frequent feature in HEMD is the use of longer phrases, sentences that are adduced to throw light on how the entry word is used in context. Such items may come with their own usage and style labels, definitions and sometimes etymology. HEMD distinguishes three types among them i.e. sayings, pseudo-sayings, and similes. The element <EXPRGRP> serves to contain several <EXPR> elements, which share their <TYPE> attribute. They have a more elaborate structure than what could be encoded within <USG type="coll"> tags, on the other hand, they are more constrained than <RE> as they do not feature on the same level in the hierarchy as <ENTRY>. Hence we felt the need to set up a separate element to deal with this material.

Example:
ablak ...
Sz: *gúny: nem teszi (ki az) az ~(á)ba:* ezzel ugyan nem fog dicsekedni!; *az ~on dobja v. hajítja v. szórja ki a pénzt:* fölöslegesen költekezik.
window ...
Sz: *sarcastic: does not put (out) in (the/his) ~:* this is certainly nothing to show off with; *throws/casts/hurls money through the ~:* spends money unnecessarily.

```
<exprGrp type="SZ">
 <expr><usg>gúny:</usg>
   <eg rend="POST colon">
   <q>nem teszi <hint>ki az</hint> <oref><hint>á</hint>ba</q>
   </eg>
   <def>ezzel ugyan nem fog dicsekedni!;</def>
 </expr>
 <expr>
   <eg rend="POST colon">
   <q>az <oref>on dobja <lbl>vagy</lbl> hajítja
      <lbl>vagy</lbl> szórja ki a pénzt:</q>
   </eg>
   <def>fölöslegesen költekezik.</def>
 </expr>
</exprGrp>
```

4.2 Extending the content model or attribute list of some tags

4.2.1 Adding <usg> and <lbl> to the content model of the <def> tag

ad ... **3.**
<Ruhadarabot> vkinek a testére húz, segít.
give ... **3.**
<Piece of clothes> sy his body pull, help.

```
<def>
 <usg type="hint">Ruhadarabot</usg> vkinek a testére húz, segít.
</def>
```

alatt ... **8.**
vmely feltétel, ürügy ~ : (helyesen:) feltétellel, ürüggyel
under ... **8.**
some condition, pretext ~: (correctly:) under some condition, pretext

```
<def>
 <usg type="acc">
```

```
   <lbl>helyesen</lbl> feltétellel, ürüggyel
 </usg>.
</def>
```

4.2.2 Adding <type> attirubute to the <def> and <pos> tags

In Hungarian lexicographic tradition, sometimes a definition is given in terms of describing the situation when the particular word or phrase is used. In such cases, the <DEF> is marked with a <TYPE> attribute to indicate that this is not a genuine definition.

a . . . **3.** . . . <A figyelmet vmely hangra irányító szóként.>
a . . . **3.** . . . <Used as a word to draw attention to a sound.>

```
<def type="rep">A figyelmet vmely hangra irányító szóként.</def>
```

Some suffixed forms are defined in <DEF> tags of similar attribute by explicating the composition of the form concerned. E.g.

ahhoz . . . Az *az* mutnévm ragos alakja
ahhoz . . . Inflected form of the demonstrative pronoun *az*

```
<entry><lemma>ahhoz</lemma><sense><def type="rep">Az <mention>az
</mention>mutnévm ragos alakja</def></sense>
```

Sometimes the <POS> tag carries a similar indirect definition.
de- (idegen szavak előképzője)
de- (prefix in foreign words)

```
<gramgrp>
  <pos type="rep">idegen szavak előképzője</pos>
</gramgrp>
```

4.2.3 Restricting the contents of the <eg> element

In the teidict2.dtd the element <%N.EG> included <%M.DICTIONARYPARTS>. It was allowed to recur in any position and could introduce a whole bag of unnecessary elements. It almost looks like a bug in the original design and was eliminated in our DTD.

```
<!ELEMENT %n.eg; - O          (%n.q; | %n.quote; | %n.cit;)+
                     +(%m.dictionaryParts | %m.formPointers)>
```

4.3 The revised dtd

Technically, the modifications were implemented in comformity with the TEI guidelines as described in Chapter 29 of the TEI manual (Sperberg and Burnard, 1994). In conclusion we display the proposed modifications to the teidict2.dtd in full.

```
<!--          concede.dtd  -->
<!-- The declarations below define new extensions -->
<!ELEMENT %n.exprGrp    - -  (%n.expr;)+                          >
<!ATTLIST %n.exprGrp;        %a.global;
                             %a.dictionaries;
          type               (space|Sz|Szh|Szj)  'space'          >

<!ELEMENT %n.expr;      - -  (%n.q | %n.usg; | %n.lbl; | %n.def;
                             | %n.eg; | %n.etym;)*                >
<!ATTLIST %n.expr;           %a.global;
                             %a.dictionaries;                     >

<!-- The following declarations define revised tags -->

<!ELEMENT %n.def;       - O  (%paraContent | %n.usg; | %n.lbl;)*>
<!ATTLIST %n.def;            %a.global;
                             %a.dictionaries;
          TEIform            CDATA               'def'
          type               CDATA               #IMPLIED         >

<!ELEMENT %n.pos;       - O  (%paraContent;)                      >
<!ATTLIST %n.pos;            %a.global;
                             %a.dictionaries;
          TEIform            CDATA               'pos'
          type               CDATA               #IMPLIED         >

<!ELEMENT %n.usg;       - O  (%paraContent | %n.usg)*             >
<!ATTLIST %n.usg;            %a.global;
                             %a.dictionaries;
          TEIform            CDATA               'usg'
          type               CDATA               #IMPLIED         >

<!ELEMENT %n.eg;        - O  (%n.q; | %n.quote; | %n.cit; | %n.usg)+

                                                 +(%m.formPointers)
                                                                  >
<!ATTLIST %n.eg;             %a.global;
                             %a.dictionaries;
          TEIform            CDATA               'eg'             >
```

5 Encoding problems

In a few cases the encoding presented some difficulty not because of the lack of notation devices but on the contrary because TEI offered various alternative ways to encode the same phenomenon. We set out below the particular solutions we adopted.

1. Pointers and cross referencing
   We have decided to implement the cross-reference to a numbered homonym by forming a

unique id of the homonym out of the orthographic form and the homonym number and include as an attribute of the <FORM> tag. The hardwired number is local to the entry concerned and is independent of any id number the entry itself may assume. This arrangement is expected to make any possible updating of the cross-references fairly straightforward.

áll[1] ...
stand[1] ...

```
<entry type="hom" n='1'>
   <form id='áll.1' type="lemma"><orth>áll</orth></form>
```

állandó ...
[←*áll*[1]]
standard ...
[←*stand*[1]]

```
<etym><mentioned><ptr target='áll.1'></mentioned></etym>
```

2. Optional and alternate forms

   In order to save space, the HEMD frequently resorts to putting optional information in parentheses and to merging alternative forms into a single disjunctive expression.

   - optionality

ablak ... **1.** Épületen, járművön a világosság és a levegő bebocsátására való (zárható) nyílás.
window 1. On a building or a vehicle, a (closable) hole to let the light and air in.

```
<def next=ablak.1.3 id=ablak.1.2>
 Épületen, járművön a világosság és a
 levegő bebocsátására való </def>
<def prev=ablak.1.2 opt=y next=ablak.1.4 id=ablak.1.3>zárható</def>
<def prev=ablak.1.3 id=ablak.1.4> nyílás.</def>
```

   - compressing alternate forms into a single expression

     While it may be argued that the optional information in the above example may be left unencoded as part of the text of the definition (at least as a first approximation), the alternatives in the example below need to be spelt out in order to derive the explicite forms the expressions take. Sometimes the generation of all possible alternatives requires manual intervention as not all the resulting forms are well formed or because it is difficult to establish how to spell out the compressed expression.

ad ... **11.** ... *Vmire v. vminek ~ja magát v. a fejét:* vmire szánja, ill. vminek átengedi magát
give ... **11.** ... *To sth o. as sth ~ oneself o. one's head:* gives himself in to sth

```
<exprGrp><expr>
 <eg rend="POST colon">
  <q>Vmire <oref>ja magát</q>
  <q>Vmire <oref>ja a fejét</q>
  <q>Vminek <oref>ja magát</q>
  <q>Vminek <oref>ja a fejét</q>
 </eg>
 <def>vmire szánja, <lbl>ill</lbl> vminek átengedi magát.</def>
</expr></exprGrp>
```

ad … **14.** … *áldását -ja:* a) *vkire:* megáldja; b) *vmire: átv is* jóváhagyja;
give … **14.** … *one's bless ~:* a) *sy:* bless; b) *sth: fig too* concede;

```
<exprGrp><expr>
 <eg rend="POST colon">
  <q>áldását <oref>ja vkire:</q>
 </eg>
 <def n="a">megáldja;</def>
 <eg rend=''POST colon''>
  <q>áldását <oref>ja vmire:</q>
 </eg>
 <usg>átv is</usg>
 <def n="b">jóváhagyja;</def>
</expr></exprGrp>
```

3. Abbreviated keywords in definitions and examples

   Another frequent space saving device is the abbreviation of the headword to its intial letter when it is used in the body of the definition.

   ablak …
   ~nyílás … Épület falában ablaknak, ill. ablak helyett hagyott ny.
   window …
   ~opening … In the wall of buildings an o. for or instead of a window.

```
<def>
 Épület falában ablaknak,
 <lbl>ill</lbl> ablak helyett hagyott
 <abbr orig="ny.">nyílás</abbr>
</def>
```

6 Conclusions

Our work so far has suggested that preparing a lexical database from a machine readable version of a paper edition of the dictionary is far from being a trivial task. Even when the source material is available in some sort of annotated form, it is likely to be too closely following the space saving notational expediences that dominate the content and layout of the paper edition. Our source dictionary in particular is very rich in such compression techniques (e.g., abbreviations, parenthetical and disjunctive expressions etc.).

For a strictly lexical (as against typographical) view of the dictionary, it is sometimes difficult to tease out the component that inherently carry content from those that serve are dictated by space constraints. Should disjunctive phrases within the body of definitions be resolved, for example? The problem is compounded by the fact that often the interpretation of these space saving notational devices require a degree of human intelligence that is difficult to automate. Furthermore, in a number of cases the mechanical explication of alternatives lead to non-existent or ill-formed expressions, a situation which again calls for manual editing.

The conflict between the typographical and the lexical views of the dictionary that we find particularly acute in the case of HEMD, may be eased by the introduction of levels of granularity of encoding[1] much like the encoding levels in the CES guidelines. This would make it possible to develop a successive approximation to either end of the two different encoding strategies.

References

T. Erjavec, D. Tufis, and T. Váradi. Developing TEI-conformant lexical databases for CEE languages (CONCEDE project). In *In this volume*, 1999.

J. Juhász, I. Szőke, G. O. Nagy, and M. Kovalovszky, editors. *Magyar Értelmező Kéziszótár*. Akadémiai Kiadó, 1972.

C. M. Sperberg and L. Burnard, editors. *Guidelines for Electronic Text Encoding and Interchange*. Chicago and Oxford, 1994.

[1]We are grateful to Nancy Ide for this idea.