TEI Lex-0: A Target Format for TEI-Encoded Dictionaries and Lexical Resources

Laurent Romary, Toma Tasovac

TEI Lex-0: A Target Format for TEI-Encoded Dictionaries and Lexical Resources

Laurent Romary, Inria

Toma Tasovac, Belgrade Center for Digital Humanities

Where we are coming from..

```
<entry>
    <def>Un animal sans queue ni tête</def>
    <hom>
        <form>I don't remember why but I need a variant here...</form>
    </hom>
    <gramGrp>
    <note>Oups, forgot to mention some grammatical constraints</note>
    </gramGrp>
    <form>
        <orth>maybe I could put the lemma here</orth>
    </form>
    <usg type="equiv">rabbit</usg>
    <xr type="translation"><ref>rabbit</ref></xr>
</entry>
```

Tightening the TEI dictionary chapter

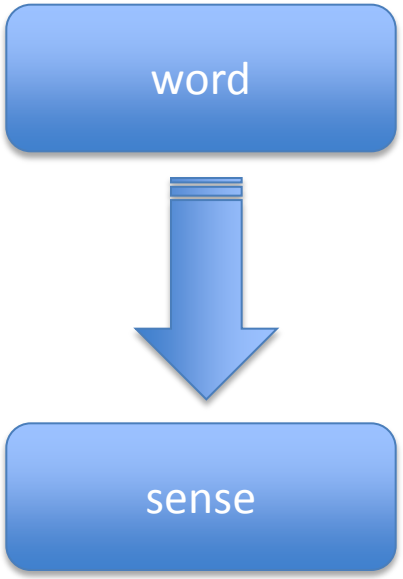
- Objective of TEI Lex-0
  - Improving consistent encoding of lexical entries across lexicographic projects
- Various use cases in mind
  - Target format (analysing and comparing)
    - Cf. TEI Analytics (MONK project)
  - Generic dictionary tools
  - Education (discussions – arguments are as important as schema)
- Position wrt current chapter
  - Provide further constraints; at times departing from the guidelines…
  - Not necessarily an editing/publishing format

TEI Lex-0 should be primarily seen as a format that existing TEI dictionaries can be univocally transformed to in order to be queried, visualised, or mined in a uniform way

Institutional setting

- Initial work
  - COST Action European Network of e-Lexicography (ENeL)
    - Working Group "Retrodigitised Dictionaries" (Toma Tasovac and Vera Hildenbrandt)
- Current
  - DARIAH
    - Working Group "Lexical Resources" (Laurent Romary and Toma Tasovac)
  - Support from H2020-funded European Lexicographic Infrastructure (ELEXIS)
- Further alignment with ISO 24613 (LMF), currently under revision

Enforcing the semasiological model




```
<entry>
    <form type="lemma">
        ...
    </form>
    <sense>
        <def>...</def>
    </sense>
</entry>
```

Overview of requirements

- General organisation of a dictionary entry
- Constraints on form and grammatical information
- Cross-references
- Embedded entries
- Usage
- Etymology

Simplifying the dictionary micro-structure

- Current situation
  - Containing vs. contained entries
    - <superEntry> – <entry> – <re>
  - Structured vs. unstructured entries
    - <entry> – <entryFree>
- The TEI Lex-0 vision
  - Representing all entry like objects as <entry>
    - Note: cf. ticket on <lbl> and <pc> in <entry>
    - Making more use of <dictScrap>
    - Making <entry> recursive

Recursive entry - example

```
<entry type="wordFamily">
  <form type="base">
    <orth>Haus-</orth>
  </form>
  <pc>,</pc>
  <form type="base">
    <orth>haus-</orth>
  </form>
  <pc>:</pc>
  <!-- possibly some shared usg information -->
  <entry type="wordForm">
    <form type="lemma">
      <orth expand="Hausaltar">-altar</orth>
      <pc>,</pc>
      <gramGrp>
        <gen value="masculine">der</gen>
      </gramGrp>
    </form>
    <sense>...</sense>
  </entry>
  <entry type="wordForm">
    <form type="lemma">
      <orth expand="Hausandacht">-andacht</orth>
      <pc>,</pc>
    </form>
        <!-- ... -->
   </entry>
   <!-- ... -->
</entry>
```

Reducing the content model of <entry>

- Allowed in <entry>
  - <form>, <sense>, <entry>, <etym>, <gramGrp>, <usg>, <xr>, <pc>, and <dictScrap>
- Not allowed in <entry>
  - <def>, <hom>, and <cit>
- Additional features
  - Mandatory @xml:id
  - Mandatory @xml:lang to indicate the object language
  - Encouraging using a @type on <entry>
    - Ongoing discussion to determine a coherent set of values

Refining the use the representation of forms

- Cf. Banski, Bowers and Erjavec (2017) at eLex
  - Inheritance of grammatical information
  - Multiple parts of speech in dictionary entries
  - Representation of lemmas and inflected forms
  - Representation of paradigm
  - Representation of variants (orthographic, phonetic, dialectal)

Structured lexical references (xr/ref) - 1

- The current mess
  - <gloss> for the provision of simple (non refined) translation equivalents of the head word
  - <usg type="synonym”> for synonym references
  - cit[@type="translation"]/quote for translation equivalents in bilingual or translation dictionaries
  - <oRef>, <pRef> for the resolution of “~" headword placeholders in quotations and other dictionary text
  - <xr>, <ref> as a general cross-referencing mechanism
  - <ptr> as a pointer to another location
  - <link>
  - <mentioned> in the etymology section
  - <term> for mentions of technical terms

Structured lexical references (xr/ref) – 2

- Towards a more unified and more constrained mechanism for lexical references
  - existing lexical entity in some dictionary or lexicon
  - lexical objects without a target reference
- Based upon <xr>/<ref>

```
<entry>
  <form type="lemma”>
    <orth>dog</orth>
  </form>
  <sense>
     <xr type="hyperonym”><ref xml:lang="en” type="entry">mammal</ref></xr>
  </sense>
</entry>
```

  - xr/@type mandatory (default: ‘lexical’)
    - indication of the lexical relation between the headword of the entry and the object referred to
    - lexical, synonym, hyponym, hyperonym, antonym , meronym (etc.), translationEquivalent, cf
  - ref/@type may be an indication of the target object category (‘entry’, ‘sense’)
  - Recommend use of ref/@xml:lang, ref/@target, ref/@notation

Usage information

- a label which can be attached at various points in the entry hierarchy in order to signal e. g. restrictions in terms of geographic regions, domains of specialized language or stylistic properties for the particular lexical item that it is attached to
  - label-like descriptors (often abbreviated) and as fuller narrative expressions
    - E.g. <usg type="style" norm="expletive">Schimpfwort</usg>
    - <usg type="hint">(рекла сељанка на њиви за време врућине)</usg> (“(said by a peasant woman in the field in hot weather)”)

Providing more coherence to usg/@type

- usg/@type is made mandatory
- usg/@norm is encouraged
- Reference works of Svendsen (2009) and Atkins and Rundell (2008)
  - Cf. usual linguistic notions of diachronic, diatopic, diastratic etc.
- Dropping values that are superfluous in the current guidelines given other TEI lex choices
  - lang, gram, syn, hyper, colloc, comp, obj, subj, verb
- New recommended values
  - temporal, domain, sociocultural, meaningType, frequency, attitude, normativity, hint

Implementation

- ODD specification available on
  - https://github.com/DARIAH-ERIC/lexicalresources
- Change documentation via GitHub tickets
- Pushing request to the TEI guidelines when we think what we do is of general interest

We've got a ticket to ride…

- Numerous evolutions that the group has initiated or to which members of the group
  - Historical reminder: generalisation of <gramGrp> as container of grammatical information (2011-2012)
  - model.entryPart.top for <pc> and <lbl>
    - Reconciling the lexical and the editorial view
  - Generalising the use of @notation (orth, pron, hyph, stress, syll, pRef, oRef)
  - Improving <etym>: att.typed and recursivity
  - <entry> in <def> (chr-emil)
  - Deprecating oVar and oRef; before a deeper reform is set in place…
  - Towards a model.sensePart class for <sense>
  - …

What's next on our plate

- Publishing the TEI Lex recommendations!
- Disseminating and training
- Improve our vision:
  - Recursive entries
  - Final values for various @type attributes
  - Leftovers: e.g. Frequency and source corpus

MERCI - ХВАЛА

DARIAH-EU
Digital Research Infrastructure
for the Arts and Humanities

elexis european lexicographic infrastructure

Co-funded by the Horizon 2020 programme
of the European Union