Harmonizing Dictionary Digitization

A Practical Report from the Project "Old German Reference Corpus“
with an Outlook to Standardization

RELISH meets LOEWE

Rendering Endangered Lexicons Interoperable through Standards Harmonization

Roland Mittmann
mittmann@em.uni-frankfurt.de

Frankfurt am Main, 10 October 2011

The Project

- „Referenzkorpus Altdeutsch“
  („Old German Reference Corpus“)
- Target: Creation of an annotated corpus of all Old High German and Old Saxon texts, accessible online
- Lemma translations and morphological information to be automatically preannotated before manual editing
- Information retrieved from several dictionaries (mostly late 19th / early 20th c.) – one per text or text collection

Sample text in database

Select Displayed Annotation Levels

T00Manuskript_B	d ˜ ˜ m		K i l a u b u	i n	k o t	f a t ˜ ˜
T01Manuskript_R	A					A
T02Manuskript_W	d˜m		Kilaubu	in	kot	fat˜
T05Referenztext_W	deo	.	Kilaubu	in	kot	fater
T06Standard_B	d e u m	.	g i l o u b u	i n	g o t	f a t e r
T07Standard_W	deum	.	giloubu	in	got	fater
T08Lemma	deus		gilouben	in	got	fater
T09Uebersetzung	Gott		glauben (an); beipflichten, gelten lassen, annehmen	in, an, auf, zu, bei, unter, zwischen, vor, durch, mit, nach, kraft, von, über, aus, gegen(über); in bezug auf, gemäß, während, als, für, um ... willen	Gott	Vater; Abt
T10Sprache	lat		goh	goh	goh	goh
T11M1a_DDDTS_Lemma	NA	$.	VV	AP	NA	NA
T12M1b_DDDTS_Beleg	NA	$.	VVFIN	APPR	NE	NA
T13M2a_Flexion_Lemma	o_Masc		wk1a		a_Masc	er_Masc
T14M2b_Flexion_Beleg_1	o_Masc		wk1a		a_Masc	er_Masc
T15M2c_Flexion_Beleg_2	Sg_Acc		Ind_Pres_Sg_1		Sg_Acc	Sg_Acc
T16S1a_Satz	CF_U_M		CF_U_M			
T17Ts1_Seite_Ed_1	S27					
T18Ts2_Zeile_Ed_1	7		8			
T19Ms1_Seite	911321					
T21Flexionsklasse_L	o		wk1a		a	er
T22Genus_L	Masc				Masc	Masc
T23Flexionsklasse_B	o		wk1a		a	er
T24Genus_B	Masc				Masc	Masc
T26Modus			Ind			
T27Tempus			Pres			
T28Numerus	Sg		Sg		Sg	Sg
T29Person			1			
T31Kasus	Acc				Acc	Acc
T34S_Satzart	CF		CF			
T35S_Syntax	M		M			
T36S_Einleitung	U		U			

Structure of a lexicon entry

- Lemma
  - followed by its
    - morphological information (part of speech)
    - translation (partly)
- Semantic structuring (partly)
- Records
  - sorted according to
    - morphological categories
    - spelling
    - context
  - followed by a reference to their location in the text

gomman-barn *st. n. männliches Kind, masculinum: nom. sg.* 7, 2.
gomo *sw. m. im Compos.* brûti-gomo.
got *st. m. deus (dominus): nom.* 1, 1. 4, 14. 5, 9. 13, 14. 21, 7 (3) *etc.* (*zus.* 28 *mal*). got Abrahames (Isakes) 127, 4. got totero 127, 4. truhtin got Israhelo (unser) 4, 14. 128, 2. *voc.* got 118, 2. 3. got min 207, 2 (2). min got 233, 7. *gen.* gotes 82, 9. 90, 4. 126, 3. 244, 2; *vgl.* 4, 18.

Eduard Sievers (ed.): TATIAN, 1872

Preparing the digitization

- OCR programs not optimized for historical fonts
  → manual digitization into XML format

- Provision to digitizers:
  - XML-encoded sample page
  - list of elements, attributes, and attribute values
    - payment per character implies choice of short, but recognizable element, attribute and attribute value names
  - list of special characters (to avoid ambiguities)

Example

gomman-barn *st. n. männliches Kind, masculinum: nom. sg.* 7, 2.
gomo *sw. m. im Compos.* brûti-gomo.
got *st. m. deus (dominus): nom.* 1, 1. 4, 14. 5, 9. 13, 14. 21, 7 (3) *etc.* (*zus.* 28 *mal*). got Abrahames (Isakes) 127, 4. got totero 127, 4. truhtin got Israhelo (unser) 4, 14. 128, 2. *voc.* got 118, 2. 3. got min 207, 2 (2). min got 233, 7. *gen.* gotes 82, 9. 90, 4. 126, 3. 244, 2; *vgl.* 4, 18.

```
- <entry>
    <lem>got</lem>
    <pos>st. m.</pos>
    <trlat>deus (dominus)</trlat>
  - <case>
      <form>nom.</form>
    - <inst>
        <rec>1, 1</rec>
        <rec>4, 14</rec>
        <rec>5, 9</rec>
        <rec>13, 14</rec>
        <rec>21, 7 (3)</rec>
        <rec>etc.</rec>
      - <rem>
          <com>zus. 28 mal</com>
        </rem>
      </inst>
    - <inst>
        <expr>got Abrahames (Isakes)</expr>
        <rec>127, 4</rec>
      </inst>
    - <inst>
        <expr>got totero</expr>
        <rec>127, 4</rec>
      </inst>
    - <inst>
        <expr>truhtin got Israhelo (unser)</expr>
        <rec>4, 14</rec>
        <rec>128, 2</rec>
      </inst>
    </case>
  - <case>
      <form>voc.</form>
    - <inst>
        <expr>got</expr>
        <rec>118, 2</rec>
        <rec>118, 3</rec>
      </inst>
    - <inst>
```

Difficulties

- Sample page cannot cover all problems appearing
- Multiple hierarchies
  - e.g. footnotes (different hierarchical structure: page)
- Encoding of variants, cases of doubt, remarks, and references
- Special characters (misprints, rare occurrences)
- Not all semantic differences obvious
  - e.g. lemma translations into different languages (German, English, Latin) without declaration → language skills needed

➔ Continuous correspondence with digitizers required

Outlook: Standardization

- Digitization of dictionaries intended only for internal use within the project
  (→ short, idiosyncratic element names and hierarchies)
- May however serve as an example for non-standards-compliant digitized dictionaries
- Approach: mapping dictionaries to an existing standard
  - adaptation of hierarchical structures
  - adaptation of element, attribute and attribute value names

Standardization: More difficulties

- Special features of Old High German and Old Saxon dictionaries:
  - lists of occurrences
  - text excerpts containing records – without marking the record
  - comments anywhere, internal and external references
  - misprints

- Special features of digitized versions at hand:
  - lemma translations without declaration of language
  - read errors and uncorrected misprints

➔ To be taken into consideration

Thank you!
Dankōm iuwih!
Thankon iu!