Mittmann

Title:	From the digitized glossary to the automatically pre-annotated text: Pre-processing the grammatical data for the Old German Reference Corpus
Short Title:	From the digitized glossary to the pre-annotated text
Autor:	Roland Mittmann
E-Mail Address:	mittmann@em.uni-frankfurt.de
Mail Address:	Roland Mittmann Johann Wolfgang Goethe-Universität Institut für Empirische Sprachwissenschaft Postfach 11 19 32, Fach 171 60054 Frankfurt Germany
Language:	British English

Abstract:

The project *Referenzkorpus Altdeutsch* (Old German Reference Corpus) aims to establish a deeply-annotated text corpus of all extant Old German texts. In order to reduce the effort required, we decided for an automated pre-annotation of the existing plain text corpus: After digitizing the respective glossaries, we retrieved the required information, enriched it with additional specifications and adapted it to the standards used for the corpus. This enabled us to automate the assignment of the appropriate annotation datasets to the word tokens in our corpus.

From the digitized glossary to the automatically pre-annotated text: Pre-processing the grammatical data for the Old German Reference Corpus

1. Introduction

While an automated annotation of modern language texts has almost become a matter of course, it still remains a desideratum for most historical languages. In many cases, the intention to change this is even complicated by a high degree of phonematical, morphological and spelling variation within the language. One example for this is Old German, consisting of Old High German and Old Saxon. Fortunately, since the late 19th century, a range of glossaries have been set up, each covering a subcorpus of Old German and listing all lemmata together with their corresponding attestations.[1] These data could be used to provide the word tokens in the texts with their lemmata and with further information on the lemmata and the records themselves.

This approach is used by the DFG-funded research project *Referenzkorpus Altdeutsch* (Old German Reference Corpus).[2] The project aims to produce a deeply-annotated corpus of all preserved texts from Old High German and Old Saxon, which date from ca. 750 to 1050 CE and comprise a total of 650,000 word tokens. The largest coherent subcorpora are the Old High German works of Notker Labeo and Otfrid of Weissenburg, an Old High German translation of the gospel harmony of Tatian the Assyrian and the Old Saxon gospel harmony now known as the *Heliand*. Edited versions of all texts exist in print; they have been digitized by the TITUS project.[3] The respective glossaries give the lemmata together with their translations and at least some of the morphological features specific to each lemma. The entries are completed by a list of attestations, followed by a reference to their location within the text. The glossaries have been digitized into an XML format (cf. Mittmann 2013), as depicted in Figure 1 by way of example.

[1] Heffner (1961), Hench (1890), Hench (1893), Kelle (1881), Sehrt (1955), Sehrt (1966), Sievers (1874), Sievers (1892) and Wadstein (1899).
[2] *http://www.deutschdiachrondigital.de*
[3] Thesaurus Indogermanischer Text- und Sprachmaterialien (Thesaurus of Indo-European Text and Language Materials), *http://titus.uni-frankfurt.de*

```
- <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>
```

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.

Figure 1: Detail from Sievers (1892), XML (l.) and original (r.) versions

2. Processing the glossary data

In the first step, all attested records of word tokens contained in a glossary are extracted with their corresponding lemma, part of speech, inflectional information and their location within the text. To this end, the glossary file is scanned line by line and the values are stored, and whenever a record appears, it is output into a file together with its corresponding qualities. If a record is given within its context (cf. Figure 2), it first has to be identified. While a scholar can easily recognize the word concerned within the context, the computer cannot do so – unless the record is identical to the lemma. Otherwise, all words of the phrase are checked whether their first letter is identical to the one of the lemma. If there are several of them, the same is done with the first two letters, and so on. For the contingency that words contain a past participle prefix, these are tentatively deleted and this option checked as well. If several possibilities remain, all are output. Failing this, the same process is repeated numerous times, using a pair of lists of graphemes or grapheme clusters that frequently correspond to each other within lemmata and records: All graphemes within the phrase covered by an entry in the second list are tentatively replaced by the corresponding ones in the first. If this still does not yield a result, the same is performed again with another pair of lists covering rarer correspondences, including, for example, verbal suppletion or inflectional forms of pronouns. The two pairs of lists are kept separate as the results of the application of the frequent possible correspondences should be checked first to avoid rare possible

correspondences being incorrectly applied. The lists are set up manually by examining missing or incorrect results.

```
<lem>uuësan</lem><pos>an. v.</pos>  [...]
<case><form>imp. sg.</form><inst><expr>ouh thu uuis obar fimf
burgi</expr> <rec>151, 6</rec>
```

Figure 2: Excerpt from digitized glossary file (cf. Sievers 1892: 491 and 495)

In the case of Figure 2, the record `uuis` can easily be recognized automatically as it is the only word token beginning with `u`. For the eventuality of, for instance, a word token `uuela` also being contained in the phrase,[4] the following – thus not only the third, but also the fourth – grapheme is checked as well: `uuel` cannot be matched to `uuës`, but `uuis` can be matched to it, as the replacement of `i` by `e` is contained in the lists. The lists would also help to attribute a word-initial `vu` or an inflected form `ist` to `uuësan`.

3. Adapting the part-of-speech and inflectional information

Within the file, all part-of-speech and inflectional information is transferred into the standard of the *Deutsch-Diachron-Digital-Tagset* (DDDTS), a tagset developed by the project and built on the basis of the *Stuttgart-Tübingen-Tagset* (STTS) for modern German (cf. Schiller et al. 1999). The transfer is done by applying regular expressions. Automatically produced lists of all part-of-speech and inflectional information occurring in the glossary facilitate this task, although in the digitized glossaries these two types of information are not always clearly separated.

The information from the glossaries is enriched by manually added rules developed using the relevant grammars.[5] They identify, for example, exact inflectional classes of verbs and nouns from the shape of the lemma, whereas most glossaries only indicate a strong or a weak inflection. Finally, all records and their corresponding information are stored in a file, depicted in extracts in Figure 3.

[4] The difference between `e` and `ë` is disregarded here.
[5] Braune (2004) and Gallée (1993).

```
Lem          | Lem2    | Lem3    | PoS     | Form      | Expr | Expr2 |
Rec     | Lemma | DDDTS | Lemmabezug | Belegbezug | Flexion
[...]
uu&euml;san | uuësan | uuesan | an. v. | imp. sg. | uuis | uuis  |
151, 6 | VA     | VAIMP | irr¦st5    | irr¦st5    | Imp_Pres_Sg_2
```

Figure 3: Title and sample line from glossary data file

Figure 3 shows the data extracted from Figure 2, converted and enriched. The part-of-speech DDDTS tags `VA`/`VAIMP` (verb, auxiliary, imperative) reflect the information `v.` and `imp.`, the consideration as an auxiliary has been added lemma-specifically. The lemma-specific inflectional information `irr¦st5` also goes back to a manual amendment: The glossary lemma `uuësan` combines an irregular and a strong class 5 verb (cf. also Figure 4), the broken bar denotes alternatives. The declaration `an.` (anomalous), which would only yield `irr`, is thus ignored in this case. The second `irr¦st5` denotes the lemma-specific inflectional information of the record: Here, only `st5` would be correct, but this selection is left to the manual annotation. The record-specific inflectional information `Imp_Pres_Sg_2` is generated completely from `imp. sg.`; `Pres` and `2` are added automatically as there is an imperative only of the present tense, and in the singular, there is only one of the second person.

4. Unifying the lemmata

In the case of the Old High German texts, the various forms of each lemma and its translations are given in a unified form corresponding to the entries in Splett (1993), which covers the whole Old High German lexicon using a standardized orthography. Automatically generated lists of lemmata from each of the Old High German glossaries listed are expanded by giving the form and the translation found in Splett (cf. Figure 4). To map the glossary lemmata to the Splett lemmata, again, two pairs of lists are assembled. The first pair contains all replacement rules from the glossary lemmata to the Splett lemmata that apply mostly or always. The second one contains rules that have to be tentatively applied if the lemmata do not match – or to enable exceptions from the former rules. The composition of the lists is controlled by checking the alteration of the number of overall concordances when applying a rule. By this, a total weighted average of 84 % of all lemma concordances can be calculated for the seven Old High German glossaries. If there are several possible results, they are all output. In any case, the concordance list finally has to be precisely checked by hand to detect mistakes, especially “false friends” that look alike, but actually equate

to different lemmata. When the lemma matching is done, the Splett translations can be added automatically.

We deviate from Splett's practice in that ‹e› unaffected by umlaut is marked as ‹ë› and fricative ‹z› as ‹ʒ› in order to separate these pairs of phonemes according to the orthography used in, for instance, Braune (2004). To this, rules are set up for the application of the different graphemes that can be determined from the history of Old High German. The rules cover a total weighted average of 90 % (94 % for ‹e›/‹ë› and 77 % for ‹z›/‹ʒ›) of all cases, and in the course of a manual check of all concerned lemmata, an adaptation of the undecidable cases has to be performed.

```
uuësan sīn¦wësan 'sein, werden, geschehen, [...]¦sein, werden,
kommen, [...]'
```

Figure 4: Sample line from lemma concordance file (cf. Splett 1993: 815 and 1111)

Figure 4 shows the attribution of the glossary lemma `uuësan` to the Splett lemmata `sīn¦wësan`. In the first stage, the Splett lemma `wesan` is automatically retrieved from the dictionary, before it is altered to `wësan` as the ‹e› stands before a vowel ‹a› in the next syllable.[6] `sīn` is then added manually, for the forms of both verbs are subsumed under the same glossary lemma. After the translations are added, `ermattet, kraftlos` and `Sein, Grundlage` are deleted manually, as there are no equivalents in the glossary to the adjectival and substantival homographic lemmata `wesan` given by Splett (1993: 1111 and 1113).

5. Linking the information to the TITUS text

A subsequent program then links the pre-processed glossary data file and the lemma concordance file to the TITUS text. This program matches every word of the text with the records in the glossary data file. If the numbering of the record locations is identical in TITUS and in the glossary, a one-to-one assignment is possible. Otherwise, all corresponding datasets are assigned, and all but one will later be discarded manually. Thus, the word token *uuis* in the phrase "Themo quad her: ouh thu uuis obar fimf burgi." (Tatian 151, 6; cf. Sievers 1892: 227) will be correctly pre-annotated solely

[6] The rare cases of ‹e› instead of ‹ë› before ‹a›, as in *reda* 'speech', are most effortlessly re-adapted by hand.

as shown in Figures 2 and 3 – and not also as an uninflected form of the adjective *wīs* ‘wise’ (cf. Sievers 1892: 503).

6. Further procedures

The manual adaptation of the pre-annotation is subsequently performed with the software ELAN, developed by the Max Planck Institute for Psycholinguistics at Nijmegen, the Netherlands. Linde (2013) describes the challenges of this task which implies multifaceted case-by-case decisions. Figure 5 shows a detail from the ELAN file containing the automatically pre-annotated and the manually adapted text.

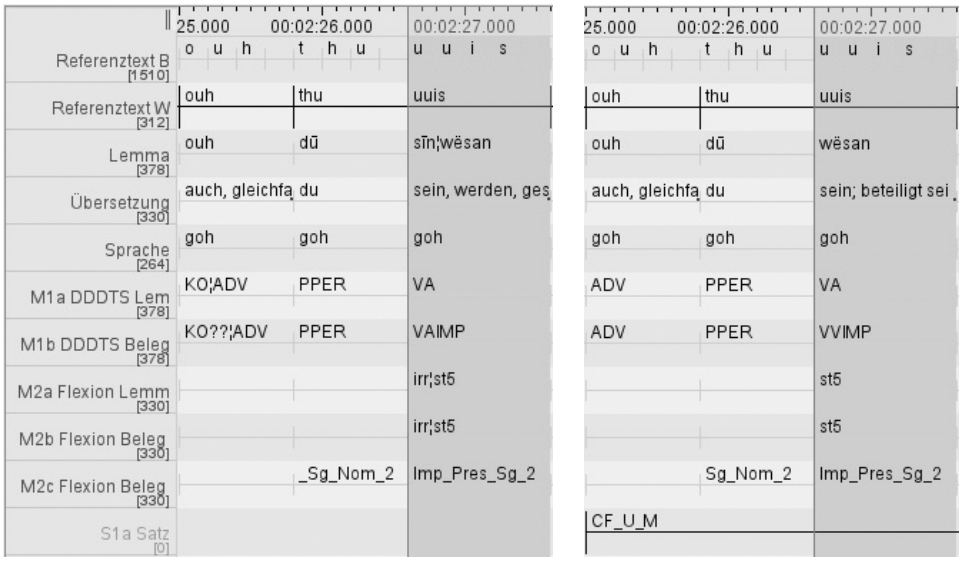

Figure 5: Detail from ELAN file before (l.) and after (r.) manual annotation

The manually adapted data are automatically checked for annotation mistakes. This is done by comparing the attested records to inflected forms of the lemmata that are created using the respective part-of-speech and morphological information (cf. Mittmann, forthcoming). The completed files are then transferred to the ANNIS database, hosted by the University of Potsdam. Figure 6 shows the corresponding result of a database query via *www.deutschdiachrondigital.de*.

edition	ouh	thu	uuis	obar	fimf	burgi
lemma	ouh	dū	wësan	ubar	fimf	burg
translation	auch, gleichfalls	du	sein; beteiligt sein	über	fünf	Stadt, Burg
posLemma	ADV	PPER	VA	AP	CARD	NA
pos	ADV	PPER	VVIMP	APPR	CARD	NA
inflectionClassLemma			ST5		I	I_FEM
inflectionClass			ST5		I	I_FEM
inflection		SG_NOM_2	IMP_PAST_SG_2		FEM_PL_ACC_0	PL_ACC
clause	CF_U_M					
document	T_Kapitel(151)					

Figure 6: Detail from query result via *www.deutschdiachrondigital.de*

7. Conclusion

More than a hundred years ago, philologists diligently compiled the first glossaries on the Old German subcorpora, striving to enable others to look up the lemmata of specific word forms in order to find their attestations in the medieval texts. Today, their work proves to be of even greater use than the editors could ever have imagined: In a digitized form, the glossaries prove perfectly appropriate to serve as the source for a comprehensive annotation of the historical documents. A manual amendment of the automatically pre-annotated data may indeed remain indispensable, but in most cases, a one-to-one assignment of the lemmatical, part-of-speech and morphological annotation to the word tokens does considerably facilitate the process of accomplishing the deeply-annotated text corpus.

References

Braune, Wilhelm. 2004. *Althochdeutsche Grammatik. Band I: Laut- und Formenlehre. 15th edition, edited by Ingo Reifenstein.* Tübingen: Niemeyer.

Gallée, Johan Hendrik. 1993. *Altsächsische Grammatik. 3rd edition, edited by Heinrich Tiefenbach.* Tübingen: Niemeyer.

Heffner, R.-M. S. 1961. *A Word-Index to the Texts of Steinmeyer. Die kleineren althochdeutschen Sprachdenkmäler.* Madison: The University of Wisconsin Press.

Hench, George Allison. 1890. *The Monsee Fragments.* Straßburg: Trübner.

Hench, George Allison. 1893. *Der althochdeutsche Isidor.* Straßburg: Trübner.

Kelle, Johann. 1881. *Glossar der Sprache Otfrids.* Regensburg: Manz.

Linde, Sonja. 2013. "Manuelle Abgleichung bei automatisierter Vorannotation: Das Tagging grammatischer Kategorien im Referenzkorpus Altdeutsch". *Altüberlieferte Sprachen als Gegenstand der Texttechnologie / Ancient Languages as the Object of Text Technology (= Journal for Language Technology and Computational Linguistics – JLCL, Vol. 27 – 2/2012)* ed. by Armin Hoenen & Thomas Jügel. Berlin: Gesellschaft für Sprachtechnologie und Computerlinguistik (GSCL), 53-64. http://www.jlcl.org/2012_Heft2/4Linde.pdf

Linde, Sonja & Roland Mittmann. 2013. "Old German Reference Corpus. Digitizing the knowledge of the 19th century. Automated pre-annotation using digitized historical glossaries." *New Methods in Historical Corpora (=Korpuslinguistik und interdisziplinäre Perspektiven auf Sprache / Corpus Linguistics and Interdisciplinary Perspectives on Language – CLIP, Vol. 3)* ed. by Paul Bennett, Martin Durrell, Silke Scheible & Richard J. Whitt. Tübingen: Narr, 235-246.

Mittmann, Roland. 2013. "Digitalisierung historischer Glossare zur automatisierten Vorannotation von Textkorpora am Beispiel des Altdeutschen". *Altüberlieferte Sprachen als Gegenstand der Texttechnologie / Ancient Languages as the Object of Text Technology (= Journal for Language Technology and Computational Linguistics – JLCL, Vol. 27 – 2/2012)* ed. by Armin Hoenen & Thomas Jügel. Berlin: Gesellschaft für Sprachtechnologie und Computerlinguistik (GSCL), 39-52. http://www.jlcl.org/2012_Heft2/3Mittmann.pdf

Mittmann, Roland. Forthcoming. "Automated quality control for the morphological annotation of the Old High German text corpus. Checking the manually adapted data using inflected forms of the lemmata". *Proceedings of "Historical Corpora 2012" (= Korpuslinguistik und interdisziplinäre Perspektiven auf Sprache /*

Corpus Linguistics and Interdisciplinary Perspectives on Language – CLIP), ed. by Ralf Gehrke et al. Tübingen: Narr.

Schiller, Anne et al. 1999. *Guidelines für das Tagging deutscher Textcorpora mit STTS. (Großes und kleines Tagset).* http://www.sfb441.uni-tuebingen.de/a5/codii/info-stts-en.xhtml

Sehrt, Edward. 1955. *Notker-Wortschatz.* Halle: Niemeyer.

Sehrt, Edward. 1966. *Vollständiges Wörterbuch z um Heliand und zur altsächsischen Genesis.* Göttingen: Vandenhoeck & Ruprecht.

Sievers, Eduard. 1874. *Die Murbacher Hymnen.* Halle: Buchhandlung des Waisenhauses.

Sievers, Eduard. 1892. *Tatian. Lateinisch und althochdeutsch mit ausführlichem Glossar. 2nd edition.* Paderborn: Schöningh.

Splett, Jochen. 1993. *Althochdeutsches Wörterbuch.* Berlin: de Gruyter.

Wadstein, Elis. 1899. *Kleinere altsächsische Sprachdenkmäler.* Norden/Leipzig: Soltau.