Roland Mittmann – Goethe-Universität Frankfurt am Main – mittmann@em.uni-frankfurt.de

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

The approach used by the DFG-funded research project ‘Referenzkorpus Altdeutsch’[1] (Old German Reference Corpus) shows how the pre-annotation of a historical plain text corpus can be largely automated: The respective glossaries are digitized, the required information is retrieved, enriched and adapted to the standards used for the corpus, so that the word tokens in the corpus can be automatically assigned the appropriate annotation datasets.

The project aims to produce a deeply-annotated corpus of all preserved texts from the oldest stages of German (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[2] project. The respective glossaries, many of which are themselves over a hundred years old, typically 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)[3].

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, it first has to be identified. Whilst a scholar can easily recognize the word concerned within the context, the computer cannot – 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. In case 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 1:** Excerpt from digitized glossary file (cf. Sievers 1892: 491 and 495)*[4]

In the case of Figure 1, 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,[5] 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`.

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. 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

[1] http://www.deutschdiachrondigital.de

[2] Thesaurus Indogermanischer Text- und Sprachmaterialien (Thesaurus of Indo-European Text and Language Materials), http://titus.uni-frankfurt.de

[3] Mittmann, Roland (2013): Digitalisierung historischer Glossare zur automatisierten Vorannotation von Textkorpora am Beispiel des Altdeutschen. In: Hoenen, Armin/Jügel, Thomas (eds.): 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). Berlin: Gesellschaft für Sprachtechnologie und Computerlinguistik (GSCL). http://www.jlcl.org/2012_Heft2/3Mittmann.pdf

[4] Sievers, Eduard (1892): Tatian. Lateinisch und althochdeutsch mit ausführlichem Glossar. 2nd edition. Paderborn: Schöningh.

[5] The difference between `e` and `ë` is disregarded here.

enriched by manually added rules developed using the relevant grammars that 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 2.

```
Lem | Lem2 | Lem3 | PoS | Flex | 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 2: *Title and sample line from glossary data file*

Figure 2 shows the data extracted from Figure 1, 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 3), 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.

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)[6], 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 3). 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 weighted total 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)[7]. 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 weighted total 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 3: *Sample line from lemma concordance file (cf. Splett 1993: 815 and 1111)*[6]

Figure 3 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, before it is altered to `wësan` as the ‹e› stands before a vowel ‹a› in the next syllable.[8] `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)[6].

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 be later 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)[4] will be correctly pre-annotated solely as shown in Figures 2 and 3 – and not also as an uninflected form of the adjective *wīs* ‘wise’ (cf. Sievers 1892: 503)[4].

[6] Splett, Jochen (1993): Althochdeutsches Wörterbuch. Berlin: de Gruyter.

[7] Braune, Wilhelm (2004): Althochdeutsche Grammatik. Band I: Laut- und Formenlehre. 15th edition, ed. by Ingo Reifenstein. Tübingen: Niemeyer.

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