Roland Mittmann

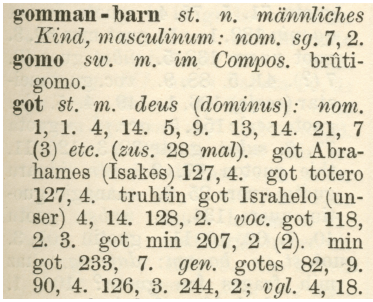
Digitalisierung historischer Glossare zur automatisierten Vorannotation von Textkorpora am Beispiel des Altdeutschen

Um Wörter und Wortformen innerhalb von Texten auffindbar zu machen, waren im vordigitalen Zeitalter Glossare unerlässlich. Heute lassen sich ihre Daten automatisiert mit den zugehörigen Texten zusammenführen, um die Texte so mit weiteren Informationen anzureichern. Für die dazu notwendige Digitalisierung der Glossare ist angesichts des historischen Druckbildes und der oft nicht eindeutigen Informationsauszeichnung ein manuelles Vorgehen am zielführendsten. Je nach Strukturierung des Glossars und nach Art und Überlieferungsdichte des behandelten Textes ergeben sich dabei unterschiedliche Herausforderungen und Probleme. Diese werden am Beispiel der Digitalisierung der Glossare zum Althochdeutschen und Altsächsischen dargestellt.

1 Die Problemstellung

Das digitale Zeitalter hat die technischen Voraussetzungen dafür, automatisch durchsuchbare Textkorpora zu erstellen, mit sich gebracht. Während die Digitalisierung der bloßen Texte heutzutage meist kein Problem darstellt, bleibt weiterhin die Frage, wie sich diese unannotierten Korpora unter möglichst geringem Aufwand mit zusätzlichen Informationen versehen lassen.

Zu vielen altüberlieferten Texten sind im 19. und 20. Jahrhundert Glossare erstellt worden, mit Hilfe derer es möglich ist, zu einem bestimmten Lemma dessen belegte Wortformen zu eruieren und die Stellen im Text zu ermitteln, an denen diese Wortformen erscheinen (vgl. Abbildung 1). Diese Glossare könnten somit also eine ergiebige Wissensbasis darstellen, wenn es durchführbar ist, im Zuge der Automatisierung eine Suche in anderer Richtung vorzunehmen: Hierzu müsste man von jedem einzelnen im Text belegten Wort ausgehen, diesen Beleg im Glossar wiederfinden und bei Mehrdeutigkeit zur eindeutigen Zuordnung auch die Position im Text abgleichen, dann die Angaben zu morphologischen Eigenschaften der Wortform auslesen und schließlich auch das Lemma sowie wiederum die zum Lemma angegebenen morphologischen Informationen extrahieren. All diese Angaben ließen sich nun dem Wort im Text zuordnen. Suchabfragen könnten sich dann nicht nur auf die Wortformen selbst, sondern auch auf die zu den Wörtern angegebenen Eigenschaften beziehen, sodass es möglich wäre, morphologische sowie teilweise auch stellungsbezogene syntaktische Aspekte in die Suchabfrage einzubeziehen. Der nach der Verknüpfung von Texten und Glossaren noch benötigte manuelle Annotationsaufwand sollte sich damit im günstigsten Fall auf eine bloße Kontrolle der ausgegebenen Daten beschränken können.



gomman-barn *st. n. männliches Kind, masculinum: nom. sg.* 7, 2.
gomo *sw. m. im Compos.* brûtigomo.
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.

Abbildung 1: Auszug aus einem Glossar (Sievers, 1892, S. 343) mit Belegform (z.B. *gotes*) und Belegstelle (z.B. 82, 9) zum Lemma *got*

Dieser Ansatz ist im Forschungsprojekt ‚Referenzkorpus Altdeutsch' auf die althochdeutschen und altsächsischen Textdenkmäler angewendet worden. Hier existiert zu jedem Teilkorpus (mindestens) ein Glossar, das die zur Auszeichnung benötigten Informationen liefern konnte.[1] Während die Texte selbst bereits digitalisiert vorlagen,[2] musste für die Digitalisierung der Glossare erst ein Vorgehen entwickelt werden.

Das DFG-finanzierte Projekt ‚Referenzkorpus Altdeutsch' hat zum Ziel, ein tief annotiertes Korpus aller überlieferten Texte der beiden ältesten Sprachstufen des Deutschen (Althochdeutsch und Altsächsisch, etwa 750 bis 1050 n. Chr.) zu schaffen. Das 650.000 Zeichen umfassende Korpus setzt sich ebenso aus interlinearen Übersetzungen lateinischer Texte wie aus freien Übersetzungen, Adaptationen und gemischten deutsch-lateinischen Texten zusammen. Hinzu kommen einige wenige vollständig in einer altdeutschen Sprache verfasste Texte, vor allem Zaubersprüche.

Im Folgenden wird zunächst das grundsätzliche Vorgehen bei der Digitalisierung beschrieben, bevor anschließend auf die Wiedergabe der Datenstruktur der gedruckten Glossare im Detail eingegangen wird. Den Abschluss bilden ein Abschnitt über spezifische Probleme, die sich beim Digitalisieren ergeben haben, sowie ein Ausblick auf eine mögliche Weiterverarbeitung der digitalen Glossare.

2 Das Vorgehen

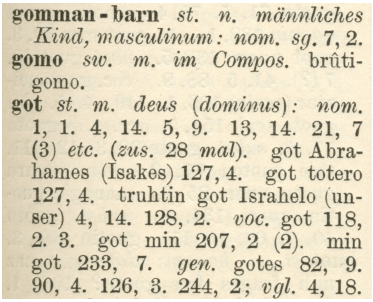
Die Glossare wurden jeweils in ihrer Gänze digitalisiert, da der Aufwand zur Auswahl der tatsächlich benötigten Teile unverhältnismäßig groß gewesen wäre und auch die Gefahr bestanden hätte, Wichtiges auszulassen. Zudem können die digitalisierten Glossare auf diese Weise auch anderen Verwendungen zugeführt werden. Da sich OCR-Programme als ungeeignet für die Drucktypen des späten 19. und frühen 20. Jahrhunderts erwiesen, wurde die Digitalisierung vollständig manuell durchgeführt.[3] Im Gegensatz zu anderen Digitalisierungsprojekten zu Publikationen aus jener Zeit stand hier eine korrekte

Wiedergabe und Unterscheidung verschiedener Typenformen und Drucktechniken in derart hohem Maße im Vordergrund, wie sie automatische Verfahren bislang nicht unter vertretbarem Aufwand erlauben. Zudem war auf diese Weise auch eine unmittelbare Übertragung der erkannten Textteile in eine für die Auszeichnung geeignete Form durchführbar, sodass die Textauszeichnung gemeinsam mit der Digitalisierung geschehen konnte.

Da die Digitalisierung in China erfolgte, war von einer Kenntnis der in den Glossaren verwendeten Sprachen, mit Ausnahme ggf. des Englischen, nicht auszugehen. Die Digitalisierer erhielten jedoch zu jedem Glossar eine bereits digitalisierte und ausgezeichnete Beispielseite als Vorbild, sodass es ihnen möglich war, analog zu verfahren. Im Fall von Ausdrücken, die sich nicht im Druckbild, wohl aber in der Sprache unterschieden, war eine Sprachauszeichnung daher jedoch nicht umsetzbar (vgl. hierzu auch Abschnitt 4). Zugleich verringerte dies aber die Gefahr von Fehlern durch unbewusste Korrektur von Wortformen in Anlehnung an andere Sprachstufen, vor allem an das Neuhochdeutsche.

Um sicherzustellen, dass mit der Digitalisierung zugleich eine möglichst präzise Textauszeichnung erfolgen konnte, war daher vonnöten, den Digitalisierern ein Auszeichnungsschema zur Verfügung zu stellen, das so gut wie möglich auf die Textsorte und die genannten Herausforderungen abgestimmt war und über kurze, aber dennoch in sich eindeutige und klar zuweisbare Tag-Namen verfügte. Denn anders als bei den meisten Wörterbuch-Digitalisierungsprojekten war das Ziel der Digitalisierung nicht die Publikation der Daten, sondern ausschließlich deren interne Weiterverarbeitung, also die Auslesung der Glossardaten, um das zugehörige Textkorpus damit annotieren zu können. Aus diesen Gründen wurde zwar als Auszeichnungssprache der De-facto-Standard XML gewählt, bei der Festlegung eines Schemas (Tag-Sets) allerdings von der Verwendung eines bestehenden Formats – z.B. TEI, wie bei Lemnitzer et al. (im Erscheinen) beschrieben und bei Christmann et al. (2001, S. 25 und passim) angewendet – abgesehen. Stattdessen wurde, ausschließlich für diesen spezifischen Zweck, ein idiosynkratisches Format entwickelt, bei dem insbesondere die Informationen, die später automatisiert ausgelesen werden sollten, zwischen den Glossaren so einheitlich wie möglich dargestellt wurden. Für die Tag-Namen, die im folgenden Kapitel näher beschrieben werden, wurden meist Längen von drei bis fünf Buchstaben, nur bei kombinierten Tag-Namen auch längere Bezeichnungen vergeben. Als Beispiele hierfür seien `<entry>`, `<lem>` (vgl. Abbildung 2) oder `<refLem>` (vgl. Abbildung 8) genannt.[4]

Über die XML-kodierten Beispielseiten hinaus wurde den Digitalisierern zu jedem Glossar je eine Liste der dort verwendeten Elemente, Attribute und Attributwerte sowie eine Liste der Sonderzeichen zur Verfügung gestellt – Letzteres vor allem, um sicherzustellen, dass die Sonderzeichen einheitlich und mithilfe des ihnen entsprechenden Unicode-Characters kodiert wurden. Bei diesem Vorgehen war ausgeschlossen, sämtliche im Glossar auftretenden Fälle im Vorfeld zu erkennen und zu beschreiben, sodass es insbesondere im Falle der ersten bearbeiteten Glossare einer intensiven Korrespondenz mit den Digitalisierern zur Klärung bislang unerfasster Fälle bedurfte.



gomman-barn *st. n. männliches Kind, masculinum: nom. sg.* 7, 2.
gomo *sw. m. im Compos.* brûtigomo.
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>gomman-barn</lem>
  <pos>st. n.</pos>
  <trlat>männliches Kind,
masculinum</trlat>
  <case>
    <form>nom. sg.</form>
    <inst>
      <rec>7, 2</rec>
    </inst>
  </case>
</entry>
```

Abbildung 2: Beispielhafter Eintrag eines einmal belegten Lemmas (vgl. Sievers, 1892, S. 343)

3 Die Wiedergabe der Datenstruktur

Der strikt hierarchische Aufbau von XML bedingt eine ebensolche Abbildung der Datenstruktur der Glossare. Die oberste Gliederungsebene nach dem Wurzelelement `<root>` bildet meist eine Gliederung nach dem Anfangsbuchstaben des Lemmas (`<let>`, vgl. Abbildung 3). An nächster Stelle in der Hierarchie folgen nun bei allen Glossaren die Lemma-Einträge (`<entry>`). Im Falle der Untergliederung nach Anfangsbuchstaben ist der erste Eintrag innerhalb eines `<let>`-Elements jedoch stets das entsprechende Graphem selbst (`<char>`). Einige Glossare enthalten darüber hinaus noch eine Eigennamenliste, die in der Hierarchie parallel zu den Anfangsbuchstaben in deren Anschluss gestellt wird (`<names>`).

Die Lemma-Einträge sind im Druck stets klar voneinander abgegrenzt und enthalten zunächst das Lemma selbst (`<lem>`), das gegenüber dem übrigen Eintrag deutlich hervorgehoben ist, gelegentlich auch um eine oder mehrere Varianten (`<lemVar>`) davon ergänzt. Anschließend folgen stets Angaben zur Wortart sowie im Falle mancher Wortarten auch zur Flexionsweise (`<pos>`, vgl. Abbildung 2: "`st`[arkes] `n`[eutrum]"). Die selten erfolgende Angabe einer genauen Flexionsklasse kann aufgrund der markierten Darstellung in runden Klammern (z.B., wie in Abbildung 3 gezeigt, bei Hench, 1890) gesondert getaggt werden (`<flex>`).

Meist folgen nun eine oder mehrere mögliche Übersetzungen des Lemmas (`<trlat>`). Die einzelnen Belegfälle (`<inst>`, vgl. Abbildungen 4 und 5[5]) bestehen aus der belegten Form (`<expr>`) – entweder allein oder im Kontext, gelegentlich auch unter Angabe einer zusätzlichen Variante (`<var>`[6]) –, anschließend bei aus dem Lateinischen übersetzten oder übertragenen Texten oft der zugrundeliegenden Entsprechung (`<equi>`) sowie der Belegstelle (`<rec>`), meist bestehend aus Kapitel und Abschnitt, je nach Art des Textes aber etwa auch aus der fortlaufenden Versnummer oder Einzeltextnummer und Zeile in

A

abanst f. (i) invidia: acc. sg. ni uueiz abanst, nescit invidere

```
<root>
  <let>
    <char>A</char>
    <entry>
      <lem>abanst</lem>
      <pos>f.</pos>
      <flex>(i)</flex>
      <trlat l="lat">invidia</trlat>
```

Abbildung 3: Beginn eines Glossars (vgl. Hench, 1890, S. 145)

der Druckausgabe. Zu einer Belegform können auch mehrere Belegstellen genannt sein, an denen diese erscheint; insbesondere dann, wenn bei der Belegform auf eine Angabe des Kontextes verzichtet wird. Wird in einer solchen Aufzählung eine Belegstelle doch gesondert markiert – etwa durch einen erwähnenswerten Kontext oder ein besonderes Äquivalent –, so wird dieser Fall separiert ausgezeichnet (`<subinst>`). Bei flektierbaren Lemmata sind die Belegfälle nach den einzelnen morphologischen Kategorien und ihren Werten (`<form>`) auf bis zu zwei Hierarchieebenen (`<case>`, `<subcase>`) untergliedert. Die kursive Schreibung einzelner Buchstaben innerhalb einer Belegform wird ebenfalls in Form einer gesonderten Auszeichnung übernommen (`<i>`). Auch wenn zu einem im Kontext angegebenen Beleg weitere vergleichbare Belegstellen genannt sind, wird dies markiert (`<sim>`, vgl. Abbildung 7). Fußnoten zu einer Seite werden sowohl einzeln (`<fn>`) als auch im Block (`<foot>`) getaggt.

Einige Lemmata weisen zudem eine teilweise sehr differenzierte semantische Gliederung auf (`<usage>`, `<subusage>`, `<specusage>`, `<subspecusage>`, vgl. Abbildung 6). Oberhalb davon erscheinen in manchen Fällen noch bis zu zwei weitere Kategorien zu unterschiedlichen Wortarten oder etwa unterschiedlichen folgenden Kasus bei Präpositionen (`<qual>`, `<nat>`[7]). Die Zuordnung geschieht hier analog zur Auszeichnung im Glossar: Bei Sehrt (1966) etwa werden mit `<usage>` immer die arabischen Zahlen in der Gliederung ausgezeichnet, unabhängig davon, ob bei einzelnen Lemmata darüber hinaus auch `<qual>` (hier Großbuchstaben) oder `<nat>` (hier römische Zahlen) vorkommen. Wie in Abbildung 6 dargestellt, erscheinen zuweilen auch dem Lemma untergeordnete Komposita (`<comp>`) hier als semantische Unterkategorie eingruppiert.

Über das bisher Beschriebene hinaus sind bei Sehrt (1966) innerhalb eines Lemmas semantische und morphologische Gliederung voneinander getrennt: Zunächst werden charakteristische Verwendungsweisen des Lemmas im Kontext dargestellt, anschließend folgen die bloßen belegten Formen, angeordnet nach Flexionskategorien (vgl. Abbildung 7). Da große Teile des *Heliand* eine parallele Überlieferung in mehreren Handschriften

hêr adj. exalted: comp. nom. sg. m. subst. *herro*, dominus 12, 21, — superl. hêrôsto, princeps: gen. sg. m. herostin 21, 21,

```
<usage>
  <case>
    <form>comp. nom. sg. m. subst.</form>
    <inst>
      <expr>h<i>erro</i></expr>
      <equi l="lat">dominus</equi>
      <rec>12, 21</rec>
    </inst>
  </case>
</usage>
<usage>
  <case>
    <form>superl.</form>
    <expr>hêrôsto</expr>
    <subcase>
      <form>gen. sg. m.</form>
      <inst>
        <expr>herostin</expr>
        <rec>21, 21</rec>
      </inst>
```

Abbildung 4: Morphologische Gliederung bei Hench (1890, S. 170)

mit diu 137, 2. 144, 1. 145, 1. 151, 4 (*dum*). 178, 4. 210, 5;

```
<inst>
  <expr>mit diu</expr>
  <rec>137, 2</rec>
  <rec>144, 1</rec>
  <rec>145, 1</rec>
  <subinst>
    <equi>dum</equi>
    <rec>151, 4</rec>
  </subinst>
  <rec>178, 4</rec>
  <rec>210, 5</rec>
</inst>
```

Abbildung 5: Darstellung gleichlautender Belege bei Sievers (1892, S. 464)

an (*got.* ana, *an.* á, *ahd.* ana an, *ags.* on, *afries.* ana an *FT 11*) *A adv.* (*Syn. § 10*) *an, auf, nach*; *hinan, hinauf*: *1) in Verbindung mit einem Verbum und dat. pers.* (*Syn. S. 211, 215, 217*): an uuas imu anst godes 784. that he themu uuîbe gedorsti stên an uuerpen 3877. dedun im eft ôđer (lakan) an 5498. than ûs liudi farad an 4141. *1a*) *c. adv. et verb.* (*vgl. Germ. XI, 214*) than hêr theobas an thingstedi halden 3745. *2*) *c. verb. et acc. pers.* (*Syn. S. 207, 214*): sah sie an langa 1291. thô bigan ina Crist sehan an mid is ôgun 3281. that sie ina than feteros an leggien môstin 3796. uueldun ina the andsacon stên an (ana) uuerpan 3941; 3871, 3946. that mugun uui ina gitellian an 5189. thes (*für* that) sie an iro môd spenit 1354. *3*) *mit einem Verbum ohne persönl. Objekt* (*Syn. S. 89, 207, 215*): huuat gi sculun an hebbean uueros te gewêdea 1664. bûtan sô gi than an hebbean te gareuuea 1856. bigunnun im (*reflex.*) tellien an 5072. *4*) *mit einem Verbum und adverb. Ausdrucke verbunden* (*Syn. §§ 168, 339*): that sie môstin an faren an thiu berhtun bû 3653.

an *1*) **784 3877. 4141**, 5498*; *1a*) **3745**; *2*) **1291** (*V*). **1354** (*V*), **3281, 3796, 5189**; *3*) **1664, 1856. 5072**; *4*) **3653.** *an 1516 *M.* ana (an*) *2*) 3871 *M* (*C* auuerpe), **3941, 3946.** *B. präp.* (*Syn. §§ 163, 165, 237*) *I c. dat.*[1]) *1) rein räumlich, in* (*unter*), *an, auf, bei*: *a) zu einem Verbum tretend um den Ort zu bezeichnen, wo das durch das Verbum Ausgedrückte stattfindet*: *α) bei Verbis, welche an sich eine Ortsbeziehung voraussetzen oder eine Ergänzung durch eine Ortsbestimmung fordern, wie bei den intransitiven Verbis des sich Befindens, Verweilens, Vorhandenseins, Existierens*: thea liudi the hêr nu langa bidun an thesara middilgard 524. ne ik an them bendion mid thi bîdan uuillie 4682; 4947. bûan an them burugium *Gen.* 238. the habda an (*M* at) them uuîha sô filu uuintro endi sumaro gilibd an them liohta 465—66; *Gen.* 92. libbian an thesun lande *Gen.* 71, 76, 305, 333. *übertragen*: that ik scal an thînun heti libbian, forđ an thînun fiundscepi *Gen.* 60—61. ligid that kind an ênera cribbiun 407; 2140—41, 3364. huuô hîr uuegos tuêna liggead an thesumu liohte 1772; 1782. (ik) an feteron lag biklemmid an karkare 4399—4400; 5397. an is breostun lag 4602. liggian an ênam diapun dala *Gen.* 29. that corn, that thar an theru lêian gilag[2]) 2394. môste thar thô an thes mahtiges Kristes barme restien 4601; 2134—34—35. sâton ira heritogon an landa gihuem 59. sitit an is uuînseli 229; 549. thar he an is rîkea sat 716. thar sie an (*M* at) mahle sittiad 1312. thar he an is benki sat 2746; 5269. Lazarus sat bliđi an is barme 3362. sittian an them stênuuege 5462. he an themu uuîhe stôd 4240. Hiericho, thiu thar an Iudeon stâd 3625. he an middien stôd 3908. stêt thînes brôđor uurâca bitter an helli *Gen.* 79. thiu burg, thiu an berge stâd 1395. thie uurti, thea hîr an felde stâd 1673; 1680. the ubilo bôm, thar he an erđu stâd 1745. thuo

```
<entry>
  <lem>an</lem>
  [...]
  <qual n="B">
    <pos>präp.</pos>
    <ref>Syn. §§ 163, 165, 237</ref>
    <nat n="I">
      <pos r="13_1">c. dat.</pos>
      <usage n="1">
        <trlat l="deu">rein räumlich, in (unter), an, auf, bei</trlat>
        <subusage n="a">
          <trlat l="deu">zu einem Verbum tretend um den Ort zu
bezeichnen, wo das durch das Verbum Ausgedrückte stattfindet</trlat>
          <specusage n="α">
            <trlat l="deu">bei Verbis, welche an sich eine Ortsbeziehung
voraussetzen oder eine Ergänzung durch eine Ortsbestimmung fordern,
wie bei den intransitiven Verbis des sich Befindens, Verweilens,
Vorhandenseins, Existierens</trlat>
             [...]
             <subspecusage>
               <comp>an-innan</comp>
               <comp>an-uppan</comp>
```

Abbildung 6: Semantische Gliederung mit sechs Hierarchieebenen (gekürzt) bei Sehrt (1966, S. 13 f.)

aufweisen, wird hier auch dargestellt, welche Form an welcher Stelle in welcher Handschrift erscheint und welche Unterschiede zwischen den Schreibungen der einzelnen Manuskripte vorliegen. Hierbei werden die sonst als Belegfälle behandelten Beispiele innerhalb der semantischen Gliederung anderweitig ausgezeichnet (`<ex>`), während die Darstellung der Belegformen im Kontext selbst unverändert bleibt (`<expr>`), sodass die Auszeichnung durch `<inst>` den tatsächlichen Belegfällen innerhalb der morphologischen Gliederung vorbehalten wird. Hier werden nun allerdings die belegten Formen anders markiert (`<shape>`), da sie zwar ohne Kontext, aber unter Angabe des Manuskripts und Bezugnahme auf die semantische Gliederung aufgeführt sind.

Schließlich können innerhalb der Lemmata auch erläuternde Teilübersetzungen von Ausdrücken oder ihren lateinischen Entsprechungen (`<expl>`) sowie Anmerkungen (`<rem>`) und Bezüge bzw. Verweise (`<ref>`) auftreten (vgl. Abbildung 8). Vor allem Letztere enthalten Bezugnahmen (`<refLem>`) oder Verweise (`<refEntry>`) auf andere Lemmata sowie Bezüge auf Belegformen (`<refForm>`) oder Ausdrücke (`<refExpr>`), oft eingeleitet durch ein `<cf>cf.</cf>` oder `<cf>s.</cf>`. Auch die nicht anderweitig markierten Inhalte einer Anmerkung werden gesondert als Kommentare ausgezeichnet

firiu-uuord *stn.* (*Syn. S. 9*) *Frevelwort, Schmähung*: felgidun imu firinuuord [9]), bismersprâka 5116; 5299.

acc. pl. firinuuord 5116, 5299*.

firio *siehe* firihos.

firi-wit (*ahd.* firi-wizzi, *ags.* fyr-wit *FT 231, 410*) *stn.* (*Syn. S. 10, 64,*

nm. zu 2428. 3) *ibid. S. 450, 32—33.*
S. 430, 13.
) *ibid. S. 465, 7; 398, 14.*
III, 203. 9) *Sievers, S. 430, 13.*

```
<entry>
  <lem>firin-uuord</lem>
  <pos>stn.</pos>
  <ref>Syn. S. 9</ref>
  <trlat l="deu">Frevelwort, Schmähung</trlat>
  <ex>
    <expr r="133_9">felgidun imu firinuuord, bismersprâka</expr>
    <rec>5116</rec>
    <sim>
      <rec>5299</rec>
    </sim>
  </ex>
  <case>
    <form>acc. pl.</form>
    <inst>
      <shape m="M, C">firinuuord</shape>
      <rec>5116</rec>
    </inst>
    <inst>
      <shape m="C">firinuuord</shape>
      <rec>5299</rec>
    </inst>
  </case>
  <foot>
    <fn n="133_9">Sievers, S. 430, 13</fn>
  </foot>
</entry>
```

Abbildung 7: Gliederung nach Handschriften bei Sehrt (1966, S. 133), vollständiger Lemma-Eintrag

oder an welchem sie stattfindet. Hinter **githuagan ist** *zu interpunktiren und* **reino** *als 3. Ps. sing. Conj. des Verbums* **reinôn** *aufzufassen, wird eben durch Joh. 13, 10, worauf man sich beruft, zurückgewiesen. Dort heisst es nämlich:*

```
<rem>
  <com>Hinter</com>
  <expr>githuagan ist</expr>
  <com>zu interpunktiren und</com>
  <refForm>reino</refForm>
  <com>als 3. Ps. sing. Conj. des
Verbums</com>
  <refLem>reinôn</refLem>
  <com>aufzufassen, wird eben
durch</com>
  <bib>Joh. 13, 10</bib>
  <com>, worauf man sich beruft,
zurückgewiesen.</com>
</rem>
```

Abbildung 8: Beispiel einer Anmerkung mit diversen Referenzen (vgl. Kelle, 1881, S. 156 f.)

(`<com>`). Daneben werden auch Verweise auf Bibelstellen markiert (`<bib>`).

Verschiedene weitere Informationen werden schließlich in Form von Attributen umgesetzt – etwa die Aussage darüber, ob über eine Angabe Zweifel bestehen (z.B. `d="?"`), die Angabe der Sprache (z.B. `l="lat"`), die einheitlich gemäß ISO 639-3[8] erfolgt, die Angabe des der angegebenen Form zugrundeliegenden Manuskripts (z.B. `m="C"`, vgl. Abbildung 7) und die Angabe einer Referenz (z.B. `r="133_9"`). Auch sonstige nummernartige Bezeichnungen, z.B. von Fußnoten oder semantischen Hierarchieebenen, werden markiert (z.B. `n="1"`).

4 Besondere Probleme der Digitalisierung

Bereits angesprochen wurde der Aspekt, dass nicht alle Problemfälle auf einer Beispielseite behandelt werden können (s. Abschnitt 2), was insbesondere bei den ersten digitalisierten Glossaren zu einer weiteren Überarbeitung und Erweiterung des Tag-Sets führte. Und auch wenn ein Großteil der Elemente sich in allen Glossaren einsetzen ließ, sind einige von ihnen nur für ein einziges Glossar konzipiert worden, wie in Abschnitt 3 exemplarisch dargestellt.

Der Umstand, dass XML eine strikte hierarchische Elementstruktur aufweist, erwies sich im Falle sich überschneidender Hierarchien als großer Nachteil. Ein eindrucksvolles Beispiel hierfür bieten die Fußnoten. Diese sind der Ebene der Seite untergeordnet, die für die Struktur des Glossars an sich aber keine Rolle spielt. Eine Lösung hierfür könnte sein, die Fußnoten umzunummerieren, indem man sie in eine vom übrigen Inhalt separierte Ebene am Ende des Dokuments überführt. Als weniger umständlich hat sich jedoch erwiesen, die Fußnoten parallel zu den (vollständigen) Lemma-Einträgen

anzuordnen und dazu an das Ende des Lemma-Eintrags, auf den sie sich beziehen, zu verschieben. In beiden Fällen ist es allerdings nötig, zur Fußnote die Seitenzahl dazuzumarkieren, um eine eindeutige Zuordnung zu gewährleisten.

Die Verarbeitung von Varianten wird oft dadurch erschwert, dass diese nur abgekürzt dargestellt werden – so handelt es sich bei dem in Fußnote 6 dargestellten Beispiel `<expr>ahtu</expr><var>ahto</var>` um eine Digitalisierung von gedrucktem „ahtu (*var.* -o)“ (vgl. Sievers, 1892, S. 302). Entscheidend ist jedoch, diese in ihrer vollständigen Form zu digitalisieren, um sie so später automatisch auslesen zu können. Dies stellt oft ein besonderes Problem für die Digitalisierer dar. Bei der Darstellung von Alternativen – etwa Lemmaformen, Flexionskategorien oder belegten Formen – kann zudem keine Art der Wiedergabe in XML völlig verhindern, dass bei der Vorbereitung des maschinellen Auslesens der Daten auf ihre Berücksichtigung besonders geachtet werden muss, um sie nicht zu übergehen. Während innerhalb eines Lemma-Eintrags angegebene Komposita, die an anderer Stelle separat erscheinen, bloß als solche markiert zu werden brauchen, stellt sich die Frage, inwieweit Sublemmata eigene Einträge konstituieren sollten. Mehr-Wort-Lemmata bereiten bei der XML-Digitalisierung zwar keine Schwierigkeiten, werden allerdings aufgrund ihrer Seltenheit beim Auslesen so eventuell nicht erwartet und könnten nicht erfasst werden – wenn in die Suchabfrage etwa keine Leerzeichen einbezogen werden.

Trotz der Bereitstellung von Listen der vorkommenden Sonderzeichen können auch bei der Kodierung von Buchstaben Fehler auftreten, am häufigsten durch Fehlinterpretation aufgrund von Veränderungen der Glyphe beim oder seit dem Druck oder Darstellung in mangelhafter Qualität nach dem Einscannen der Bücher für die Digitalisierer (z.B. `c` für ‹e›, `r` für ‹n› oder auch `n` für ‹u›). Selten erscheinende Sonderzeichen, die nicht in der Sonderzeichenliste enthalten sind, können zudem fehlinterpretiert und falsch kodiert werden. Dies kann etwa bei Angabe einer Lemmaentsprechung in einer außergermanischen Sprache oder bei ungewöhnlichen Beleg-Schreibungen (z.B. Wiedergabe von belegtem ‹ŏ› als `ô`) vorkommen. Hinzu kommt das Problem von aus der Printausgabe übernommenen Druckfehlern. Wenngleich einzelne Fehler systematisch auftreten und daher mithilfe von Ersetzungsregeln automatisiert korrigiert werden können, hat sich die Datenqualität der digitalisierten Glossare insgesamt als so gut erweisen, dass sich für den Zweck des Projekts eine systematische Fehlersuche erübrigte und Korrekturen meist nur einzelfallbezogen erfolgen, wenn bei der Weiterverarbeitung der Daten ein Fehler offensichtlich wird.

Wie die Abweichungen im Druck sind auch semantische Unterschiede, die nicht explizit markiert sind, für Digitalisierer, die die im Glossar verwendeten Sprachen nicht beherrschen, nicht zu erkennen. Dies gilt insbesondere etwa für Lemmaübersetzungen, deren Sprache aufgrund der für einen Philologen des 19. bzw. 20. Jahrhunderts bestehenden Offensichtlichkeit oft nicht angegeben ist. Im Falle der Glossare zum Althochdeutschen und Altsächsischen handelte es sich dabei um Deutsch, Englisch und Latein, auf deren Angabe daher bei der Digitalisierung verzichtet werden musste, sofern die Position oder Darstellung von Ausdrücken in den einzelnen Sprachen die Sprache nicht eindeutig erkennen ließ (vgl. Abbildung 2).

Auch die Abbildung der semantischen Strukturierung – einschließlich der Gliederung eines Lemmas in mehrere Wortarten – mit bis zu sechs Hierarchieebenen innerhalb eines Lemmas (vgl. Abbildung 6) verlangt gerade bei sehr häufigen Lemmata genaue Abstufungen; hier kommen nicht selten auch Fehler im gedruckten Glossar vor.

Innerhalb von Erläuterungen und Kommentaren erscheinen bisweilen interne sowie externe Querverweise auf andere Lemmata, Sublemmata oder Wortformen und Mehrwortausdrücke samt Übersetzung und Textstellenangabe. Diese sind ebenso wie Kommentare an nahezu jeder denkbaren Stelle innerhalb des Lemma-Eintrags – etwa mitten in Mehrwortausdrücken, die einen Beleg enthalten – unter Wahrung der hierarchischen Struktur sowie der Zusammengehörigkeit ihres Kontextes entsprechend auszuzeichnen.

5 Zur Verwendung der Daten

Der Einsatz der digitalisierten Glossare im Textkorpus wird bei Linde und Mittmann (im Erscheinen) ausführlich dargestellt. Hier wird auch auf die Grenzen der Glossardaten und ihrer Verknüpfbarkeit mit dem Text eingegangen und erläutert, wie Lemma-Formen und Übersetzungen aus den Glossaren an einen Standard angepasst werden, der für alle Texte der jeweiligen Sprachstufe Verwendung finden kann. Diese Standardisierung wäre zudem auch die Voraussetzung für die Aufnahme der Glossare in ein digitales Wörterbuchnetz, wie es Burch und Rapp (2007) beschreiben.

6 Anhang: Ein Anwendungsbeispiel

Abbildung 9 zeigt eine mit Hilfe von Glossar-Daten automatisiert vorgenommene Vorannotation der Wortform *gommanbarn* im althochdeutschen *Tatian* (vgl. Abbildung 2 sowie Sievers, 1892, S. 25 und 343; leicht vereinfachte Darstellung). Die obersten beiden Zeilen enthalten das annotierte Wort, die letzten beiden Zeilen geben dessen Position im Text an. In den dazwischen befindlichen Zeilen erscheint die Lemmatisierung auf Grundlage des Glossars sowie der dort angegebenen weiteren Informationen, bereits umgewandelt in ein standardisiertes Format.[9] In einem nächsten Schritt müssen die Angaben nun manuell geprüft werden, bevor die Annotation in die Datenbank überführt werden kann. Weitergehende Erläuterungen zum genauen Vorgehen und zu dabei auftretenden Problemfällen finden sich bei Linde und Mittmann (im Erscheinen) sowie Linde (in diesem Band).

Anmerkungen

[1] Für das Althochdeutsche sind Heffner (1961), Hench (1890, 1893), Kelle (1881), Sehrt (1955) sowie Sievers (1874, 1892) verwendet worden, für das Altsächsische Sehrt (1966) und Wadstein (1899).

[2] Die Digitalisierung erfolgte im Rahmen des TITUS-Projekts (Thesaurus Indogermanischer Texte und Sprachen) – `http://titus.uni-frankfurt.de` –, über diese Seite sind auch die Texte abrufbar.

[3] Für ein vergleichbares Vorgehen vgl. Christmann et al. (2001, S. 23).

[4] Da die Glossare vollständig erfasst wurden, wäre eine spätere Überführung des idiosynkratischen XML-Formats in ein standardisiertes XML-Format (etwa TEI), um eine Publikation der Glossare zu ermöglichen, prinzipiell denkbar. Hierzu müssten die hierarchischen Strukturen angepasst und

Referenztext Wort	gommanbarn									
Referenztext Buchstaben	g	o	m	m	a	n	b	a	r	n
Lemma	gommanbarn									
Übersetzung	männlicher Nachkomme									
Wortart Lemma	NA									
Wortart Beleg	NA									
Flexion Lemma	a,z_Neut									
Flexion Beleg 1	a,z_Neut									
Flexion Beleg 2	Sg_Nom									
Kapitel	7									
Unterkapitel	2									

Abbildung 9: Automatisierte Vorannotation eines Wortes auf Grundlage von Glossardaten (NA: Nomen, Appellativum; a,z: germanische Nominalflexionsklassen)

die Namen der Elemente, Attribute sowie Attributwerte in den entsprechenden Standard überführt werden (vgl. hierzu und zum Folgenden die Abschnitte 3 und 4). Dabei müssten insbesondere die ungewöhnlicheren Merkmale der Glossare – die Angabe von Belegstellen, die Nennung von Belegformen im Kontext ohne Markierung der Belegform selbst, die Einfügung von Kommentaren, internen und externen Referenzen an nahezu jeder denkbaren Stelle sowie das Vorkommen von Druckfehlern – berücksichtigt werden. Zudem ist die digitalisierte Form der Glossare trotz der hohen Datenqualität nicht systematisch auf verbleibende Druck- oder neu hinzugekommene Lesefehler geprüft worden. Auch dass bei Übersetzungen oftmals die Angabe der Sprache fehlt, würde eine Nutzung dieser Angaben erschweren. Eine standardisierte digitale Form der Glossare herzustellen, wäre unter Berücksichtigung dieser kleineren Erschwernisse (von denen einige jedoch auch bei standardkonformer Digitalisierung bestanden hätten) möglich, erwies sich jedoch, wie angeführt, für das Projektvorhaben als nicht optimal.

[5]Verweise auf Abbildungen gelten in diesem Artikel i. d. R. auch für die im Folgenden genannten Tags, solange kein anderweitiger Verweis erfolgt.

[6]Z.B. `<expr>ahtu</expr><var>ahto</var>` zum Lemma *ahtu* 'acht' bei Sievers (1892, S. 302).

[7]Die Benennung erfolgte durch Abkürzung der englischen Termini *quality* und *nature*, jeweils im Sinne von 'Beschaffenheit, Eigenschaft'.

[8]Nähere Informationen zu der Norm finden sich bei der zuständigen Registrierungsstelle SIL International unter `http://www.sil.org/iso639-3/`.

[9]Die Übersetzung ist ebenfalls bereits standardisiert. Da eine Belegform im Glossar nicht explizit angegeben ist, wird die Lemmaform hier als Belegform angenommen. Der in Abbildung 2 dargestellte Bindestrich ist hier regelmäßig getilgt, zumal er nicht in den Belegformen erscheint und diese so besser mit den Lemmata verglichen werden können.

Literatur

Burch, T. & Rapp, A. (2007). Das Wörterbuch-Netz: Verfahren – Methoden – Perspektiven. In: D. Burckhardt, R. Hohls & C. Prinz (Hrsg.), *Beiträge der Tagung .hist 2006 = Historisches Forum* (Bd. 10, S. 607–627). Berlin: Clio-online.

Christmann, R., Hildenbrandt, V. & Schares, T. (2001). Ein "heiligthum der sprache" digitalisiert: Das Deutsche Wörterbuch von Jacob und Wilhelm Grimm auf CD-ROM und im Internet. In: N. Castrillo Benito & P. Stahl (Hrsg.), *TUSTEP educa. Actas de Congreso del Interna-*

tional TUSTEP User Group. Peñaranda de Duero (Burgos) Octubre 1999 (S. 13–37). Burgos. (http://kompetenzzentrum.uni-trier.de/files/8513/1349/4217/Grimmbu.pdf)

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

Hench, G. A. (1890). *The Monsee Fragments.* Straßburg: Trübner.

Hench, G. A. (1893). *Der althochdeutsche Isidor.* Straßburg: Trübner.

Kelle, J. (1881). *Glossar der Sprache Otfrids.* Regensburg: Manz.

Lemnitzer, L., Romary, L. & Witt, A. (im Erscheinen). Representing human and machine dictionaries in Markup languages (SGML, XML). In: R. H. Gouws, U. Heid, W. Schweickhard & H. Erns (Hrsg.), *Dictionaries. An international encyclopedia of lexicography. Supplementary volume: Recent developments with special focus on computational lexicography. Handbooks of Linguistics and Communication Science (HSK).* Berlin/New York: de Gruyter. (http://www.dwds.de/media/publications/text/Lemnitzer_Romary_Witt-HSK-Article-v2009-12-15_-_deformatted.pdf)

Linde, S. & Mittmann, R. (im Erscheinen). Old German Reference Corpus. Digitizing the knowledge of the 19th century. In: P. Bennett, M. Durrell, S. Scheible & R. J. Whitt (Hrsg.), *New Methods in Historical Corpus Linguistics = Korpuslinguistik und interdiziplinäre Perspektiven auf Sprache – Corpus linguistics and Interdisciplinary perspectives on language (CLIP)* (Bd. 3). Tübingen: Narr.

Sehrt, E. (1955). *Notker-Wortschatz.* Halle: Niemeyer.

Sehrt, E. (1966). *Vollständiges Wörterbuch zum Heliand und zur altsächsischen Genesis* (2. Aufl.). Göttingen: Vandenhoeck & Ruprecht.

Sievers, E. (1874). *Die Murbacher Hymnen.* Halle: Buchhandlung des Waisenhauses.

Sievers, E. (1892). *Tatian. Lateinisch und althochdeutsch mit ausführlichem Glossar* (2. Aufl.). Paderborn: Schöningh.

Wadstein, E. (1899). *Kleinere altsächsische Sprachdenkmäler.* Norden/Leipzig: Soltau.