Referenzkorpus Altdeutsch:

Automatisierte Prozesse zur Konvertierung, Verknüpfung und Qualitätssicherung von Sprachdaten

20. Jahrestagung der ITUG

Mainz, 16.–18. September 2013

Roland Mittmann
Institut für Empirische Sprachwissenschaft
Goethe-Universität Frankfurt am Main
mittmann@em.uni-frankfurt.de

GOETHE UNIVERSITÄT FRANKFURT AM MAIN

Das Referenzkorpus Altdeutsch

- DFG-gefördertes Projekt (2008–2014)
- Teil der ‘Deutsch Diachron Digital‘-Initiative: Digitalisierung der wichtigsten Texte aller historischen Stufen des Deutschen
- Ziel: tiefenannotiertes Korpus aller althochdeutschen und altsächsischen Texte, ca. 750–1050
- Kooperation:
  - Humboldt-Universität (Berlin)
  - Goethe-Universität (Frankfurt am Main)
  - Schiller-Universität (Jena)

Das Korpus

- 650.000 Textwörter

- interlineare Übersetzungen lateinischer Texte
- freie Übersetzungen, Nachdichtungen und gemischte deutsch-lateinische Texte
- einige Texte unmittelbar auf Altdeutsch verfasst

- größte Teilkorpora:
  - Werke Notker Labeos (ahd.)
  - Werke Otfrids von Weißenburg (ahd.)
  - Übersetzung der Evangelienharmonie des Tatian (ahd.)
  - Heliand (as. Evangelienharmonie)

Automatisierte Vorannotation

Automatisierte Vorannotation?

- Erstellung eines tiefenannotierten historischen Textkorpus:
  - Wie die digitalisierten Texte annotieren?
    - → normalerweise: erheblicher Aufwand an manueller Arbeit
    - → Alternative: automatisierte Vorannotation mithilfe bestehender Glossare
- Vorgehensweise:
  - Digitalisierung der Glossare
  - Extraktion der benötigten Informationen
    - Ergänzung zusätzlicher Daten
    - Anpassung an die Standards des Korpus
  - Zuweisung der Datensätze zu den einzelnen Textwörtern
- Datenverarbeitung mithilfe von Perl-Skripten

Der Ausgangspunkt

- eine gedruckte Ausgabe pro Text, digitalisiert durch das TITUS-Projekt (Frankfurt am Main)

- **Old High German:**
  - **Hildebrandslied (diplomatic edition)**
    Data entry by R.J. Jacob (Neukirchen-Vluyn)
    [HTML version (UTF-8) with index available]
  - **Merseburger Zaubersprüche (diplomatic edition)**
    Data entry by R.J. Jacob (Neukirchen-Vluyn)
    [HTML version (UTF-8) with index available]
  - **Isidor**
    Data entry by Ma Pilar Fernández Alvarez and M. M. García-Bermejo Giner (Salamanca); Schuhmann (Gießen)
    [HTML version (UTF-8) with index available]
  - **Notker**
    Data entry under the guidance of R. Lühr by K. Lepper and S. Zeilfelder (Jena)
    [HTML version (UTF-8) with index available (TITUS members only)]
    [Boethius, Consolatio philosophiae (Latin text edited by J.J. O'Donnell, Virg
  - **Otfrid**
    Data entry under the guidance of R. Lühr (Jena) by R. Schuhmann and M. Bayer; further edi
    [HTML version (UTF-8) with index available]
  - **Regula Benedictorum**
    Data entry by Ma Pilar Fernández Alvarez and M. M. García-Bermejo Giner (Salamanca); corrections by A. Potthoff-Knoth and Roland Schuhmann (Gießen)
    [HTML version (UTF-8) with index available (TITUS members only)]
  - **Tatian**
    Data entry by Ma Pilar Fernández Alvarez and M. M. García-Bermejo Giner (Salamanca); Schuhmann (Gießen); corrections by S. Zeilfelder and A. Potthoff-Knoth (Gießen). Præfatio
    [HTML version (UTF-8) with index available]
    Extended Version: aligment with the original manuscript by Jost Gippert (Frankfurt)
    [HTML version (UTF-8) with index available]
  - **Monsee Fragments**
    Data entry by Emil Kroymann and Thorwald Poschenrieder (Berlin)
    [HTML version (UTF-8) with index available]
  - **Murbach Hymns**
    Data entry by Patrizia Noel (München)
    [HTML version (UTF-8) with index available]
  - **Physiologus (Old High German)**

ali] [Pra
Saka] [B
| [Pisidic
lovene]
on] [Mic
Portugue
tarily a
odern Ge
ttmann

TITUS

Tatian, Gospel Harmony

Part No. 2

Chapter: 1

Sentence: 1

In principio erat verbum et verbum erat apud deum et deus erat verbum.

In anaginne uuas uuort inti thaz uuort uuas mit gote inti got selbo uuas thaz uuort.

Sentence: 2

Hoc erat in principio apud deum. Omnia per ipsum facta sunt et sine ipso factum est nihil quod factum est.

Thaz uuas in anaginne mit gote. Alliu thuruh thaz vvurdun gitán inti ûzzan sín ni uuas uuiht gitanes thaz thar gitán uuas.

Sentence: 3

In ipso vita erat et vita erat lux hominum.

Thaz uuas in imo lîb inti thaz lib uuas lioht manno.

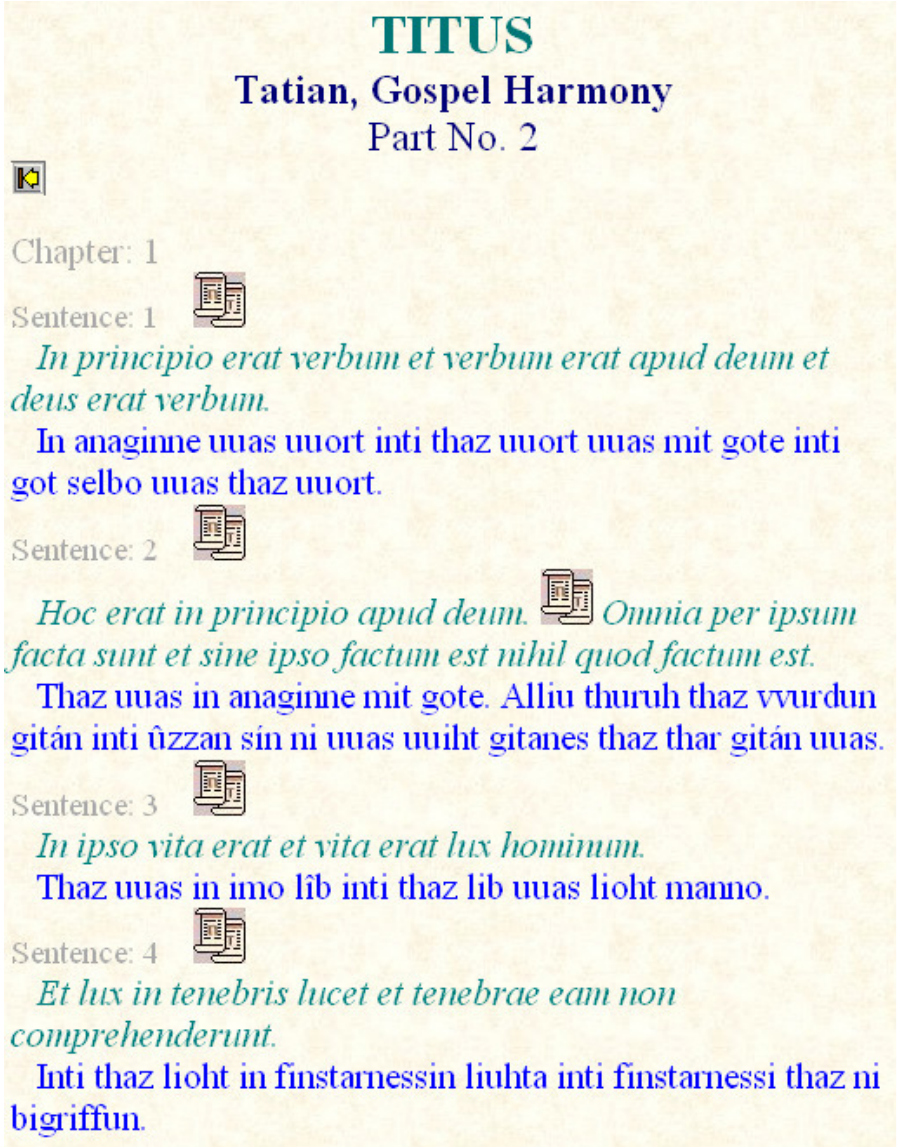
Sentence: 4

Et lux in tenebris lucet et tenebrae eam non comprehenderunt.

Inti thaz lioht in finstarnessin liuhta inti finstarnessi thaz ni bigriffun.

Der Ausgangspunkt

- eine gedruckte Ausgabe pro Text, digitalisiert durch das TITUS-Projekt (Frankfurt am Main)
  - HTML-/XML-Format
  - strukturelle Annotation, z.B. Kapitel und Zeilen für Manuskript und Edition
  - Annotation der Sprache



TITUS

Tatian, Gospel Harmony

Part No. 2

Chapter: 1

Sentence: 1

In principio erat verbum et verbum erat apud deum et deus erat verbum.

In anaginne uuas uuort inti thaz uuort uuas mit gote inti got selbo uuas thaz uuort.

Sentence: 2

Hoc erat in principio apud deum. Omnia per ipsum facta sunt et sine ipso factum est nihil quod factum est.

Thaz uuas in anaginne mit gote. Alliu thuruh thaz vvurdun gitán inti ûzzan sín ni uuas uuiht gitanes thaz thar gitán uuas.

Sentence: 3

In ipso vita erat et vita erat lux hominum.

Thaz uuas in imo lîb inti thaz lib uuas lioht manno.

Sentence: 4

Et lux in tenebris lucet et tenebrae eam non comprehenderunt.

Inti thaz lioht in finstarnessin liuhta inti finstarnessi thaz ni bigriffun.

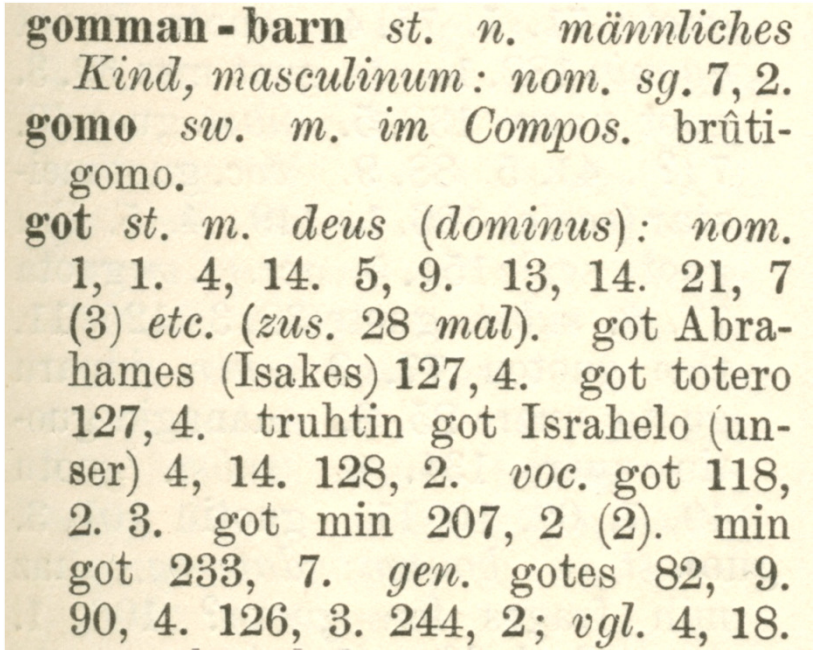
Der Ausgangspunkt

- eine gedruckte Ausgabe pro Text, digitalisiert durch das TITUS-Projekt (Frankfurt am Main)
  - HTML-/XML-Format
  - strukturelle Annotation, z.B. Kapitel und Zeilen für Manuskript und Edition
  - Annotation der Sprache

- gedruckte Glossare (v.a. spätes 19. / frühes 20. Jh.), eines pro Text oder Textsammlung

Strukture eines Glossareintrags

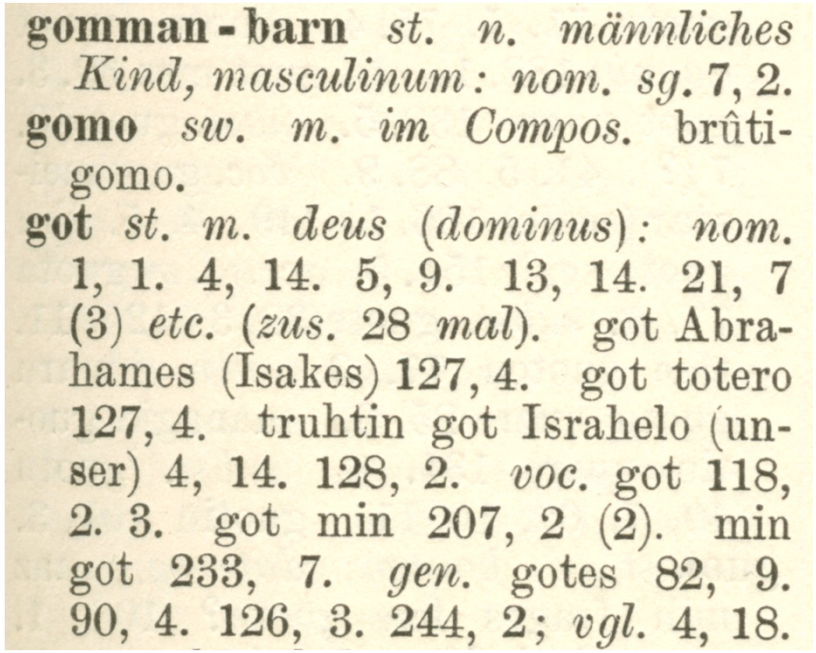
- Lemma
  - gefolgt von der zugehörigen
    - morphologischen Information (Wortart)
    - Übersetzung (teilweise)
- semantische Struktur (tw.)
- Einträge
  - sortiert nach
    - morphologischen Kategorien
    - Schreibung
    - Kontext
  - gefolgt von einem Verweis auf ihre Position im Text



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.

Digitalisierung der Glossare

- alte Schrifttypen → manuelle Digitalisierung
- Listen von Elementen, Attributen und Werten
  - kurze Namen, da Berechnung per Zeichen
  - idiosynkratisches Format, da nur eigene Nutzung



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>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>
```

Verarbeitung der Glossardaten

- Extraktion aller Wörter zusammen mit zugehörigem/zugehöriger
  - Lemma
  - Wortart
  - Flexionsinformation
  - Position im Text

→ durch automatisiertes zeilenweises Scannen
  - Speicherung der genannten Werte
  - → Belege werden in Datei ausgegeben, zusammen mit zugehörigen Eigenschaften

- oft erscheinen Belege im Kontext
  - Identifizierung einfach für den Philologen, aber nicht für den Computer (falls nicht identisch mit dem Lemma)

```
- <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>
```

Identifizierung der Belege im Kontext

1. Beleg identisch mit Lemma?
2. nur ein Wort mit selbem Anfangsbuchstaben?
3. nur ein Wort mit selben zwei Anfangsbuchstaben? – usw.
4. wenn kein Ergebnis, Wiederholung des Vorgangs mithilfe eines Listenpaars mit Graphemen or Graphemclustern, die bei Lemmata und Belegen einander oft entsprechen
   - Ersetzung von Graphemen in einer Phrase, die in zweiter Liste enthalten sind, durch entsprechende Grapheme in erster Liste
5. wenn weiterhin kein Ergebnis, Wiederholung des letzten Vorgangs mit weiterem Listenpaar mit selteneren Entsprechungen, z.B. verbale Suppletion oder Flexionsformen von Pronomina
   - Trennung von Stufen 4 und 5 spart Rechenzeit und vermeidet fehlerhafte Anwendung seltenerer Entsprechungen

- testweise Entfernung des Präfixes vermuteter Perfektpartizipien
  - mögliche Formen: `#{g|k|ch|c}{i|e|a}°`

Identifizierung der Belege: Beispielfall

```
<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>
```

- **uuis** ist das einzige Wort mit initialem **u** → Beleg
- aber was wäre, wenn z.B. **uuela** in der Phrase erschiene?
  → übernächstes Graphem wird auch geprüft:
  uue passt besser zu **uuësan** als **uui**, aber
  uuis passt besser zu **uuësan** als **uuel**
- Programm ordnet auch z.B. **vuis** oder **ist** zu **uuësan** zu

Wortart- und Flexionsinformation

- Verwendung des *Deutsch-Diachron-Digital-Tagset* (DDDTS)
  - entwickelt durch Referenzkorpora Altdeutsch und Mittelhochdeutsch
  - basiert auf *Stuttgart-Tübingen-Tagset* (STTS) für Neuhochdeutsch
- Überführung der gesamten Wortart- und Flexionsinformation in diesen Standard (durch reguläre Ausdrücke)
  - Aufgabe erleichtert durch automatische Erzeugung von Listen zu allen Wortart- und Flexionsinformationen, die im Glossar vorkommen
  - beide Kategorien jedoch nicht immer klar getrennt in digitalisierten Glossaren
- Ergänzung durch manuell hinzugefügte Regeln auf Grundlage der Grammatiken
  - Verwendung etwa des Lemma-Ausgangs, um exakte Flexionsklassen von Verben und Nomina zu ermitteln
    (die meisten Wörterbücher geben nur “stark“/“schwach“ an)

STTS	DDDTS	LEMMA	
ADJA	ADJ	ADJ	Adjektiv, attributiv
ADJD	ADJD	ADJ	Adjektiv, prädikativ oder adverbial
ADJA	ADJE	ADJ	Adjektiv, attributiv, Teil eines Eigennamens
ADJA	ADJN	ADJ	Adjektiv, attributiv, nachgestellt
ADJN	ADJNE	ADJ	Adjektiv, attributiv, nachgestellt, Teil eines Eigennamens
ADJA	ADJO	ADJO	Adjektiv, ordinal, attributiv
ADJA	ADJON	ADJO	Adjektiv, ordinal, attributiv, nachgestellt
NN	ADJOS	ADJO	Adjektiv, ordinal, substantiviert
NN	ADJS	ADJ	Adjektiv, substantiviert
ADV	ADV	ADV	Adverb
ADV	ADVM	ADVM	Adverb, multiplikativ

Die Glossardaten-Datei

- Speicherung von Belegen und zugehörigen Informationen in Datei
- Titel und Exzerpt für genannten Beispielfall:

```
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
```

- Daten aus dem Glossar, konvertiert und ergänzt
- **VA/VAIMP**: aus **v.** und **imp.**
  + manuell hinzugefügte Information „Auxiliar“
- **irr¦st5**: manuelle Ergänzung (Lemma kombiniert zwei Verben), **an.** ergäbe nur **irr**
- beide **irr¦st5** sollten **st5** lauten, manuell zu korrigieren
- **Imp_Pres_Sg_2** aus **imp. sg.**; **Pres** + **2** automatisch hinzugefügt

Vereinheitlichung der Lemmata

- unterschiedliche Lemmaschreibungen und -übersetzungen in jedem Glossar
- für Ahd. Jochen Spletts „Althochdeutsches Wörterbuch“ (1993) als Standard: gesamter ahd. Wortschatz, einheitliche Schreibung
- Anpassung der Glossarlemmata:
  - Erweiterung automatisch erzeugter Lemmalisten aus dem Glossar um Splett-Lemmata und Übersetzungen
  - wiederum Anwendung von 2 Paaren Graphem(cluster)ersetzungslisten:
    - eine mit Regeln, die (fast) immer gelten
    - eine mit Regeln, die nur versuchsweise angewendet werden – auch mit Ausnahmen von der ersten Liste
  - Formulierung der Regeln kontrolliert durch Prüfung der Änderung der Gesamtzahl an Übereinstimmungen bei Anwendung einer Regel

Vereinheitlichung der Lemmata

- gewichteter Gesamtdurchschnitt von 84 % aller Lemma-Konkordanzen für die 7 ahd. Glossare errechenbar
- verbleibende Lemmata manuell zuzuweisen
- wenn mehrere mögliche Ergebnisse, Ausgabe aller
- Listen genau zu prüfen
  - v.a. auf „falsche Freunde“: homographische Lemmata
- automatisierte Ergänzung der Lemmaübersetzungen

Lemmata: Markierung von ‹ë› und ‹ȝ›

- abweichend von z.B. Wilhelm Braunes „Althochdeutscher Grammatik" kennzeichnet Splett unumgelautetes ‹e› nicht als ‹ë› und frikatives ‹z› nicht als ‹ȝ›

→ Erstellung von Regeln für die Zuordnung der Grapheme, ~ ermittelbar aus der Geschichte des Althochdeutschen

  - liefern ein Ergebnis für gewichteten Gesamtdurchschnitt von 90 % aller 22.223 Fälle (94 % for ‹e›/‹ë›, 77 % for ‹z›/‹ȝ›)

  - manuelle Prüfung aller betroffenen Lemmata, v.a. der unentscheidbaren Fälle

Lemmavereinheitlichung: Beispielfall

- Beispielzeile aus finaler Lemmakonkordanz-Datei:

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

- Vorgehensweise:
  1. Ermittlung des Splett-Lemmas **wesan** aus dem Glossarlemma **uuësan** mithilfe der Ersetzungslisten
  2. Anpassung zu **wësan**, da ‹e› vor ‹a› in nächster Silbe steht
  3. manuelle Ergänzung von **sīn** (Glossarlemma umfasst beide Verben)
  4. nach Ergänzung der Übersetzungen manuelle Löschung von „**ermattet, kraftlos**“ und „**Sein, Grundlage**“, da adjektivische und substantivische homographische Lemmata **wësan** nicht im Glossar erscheinen

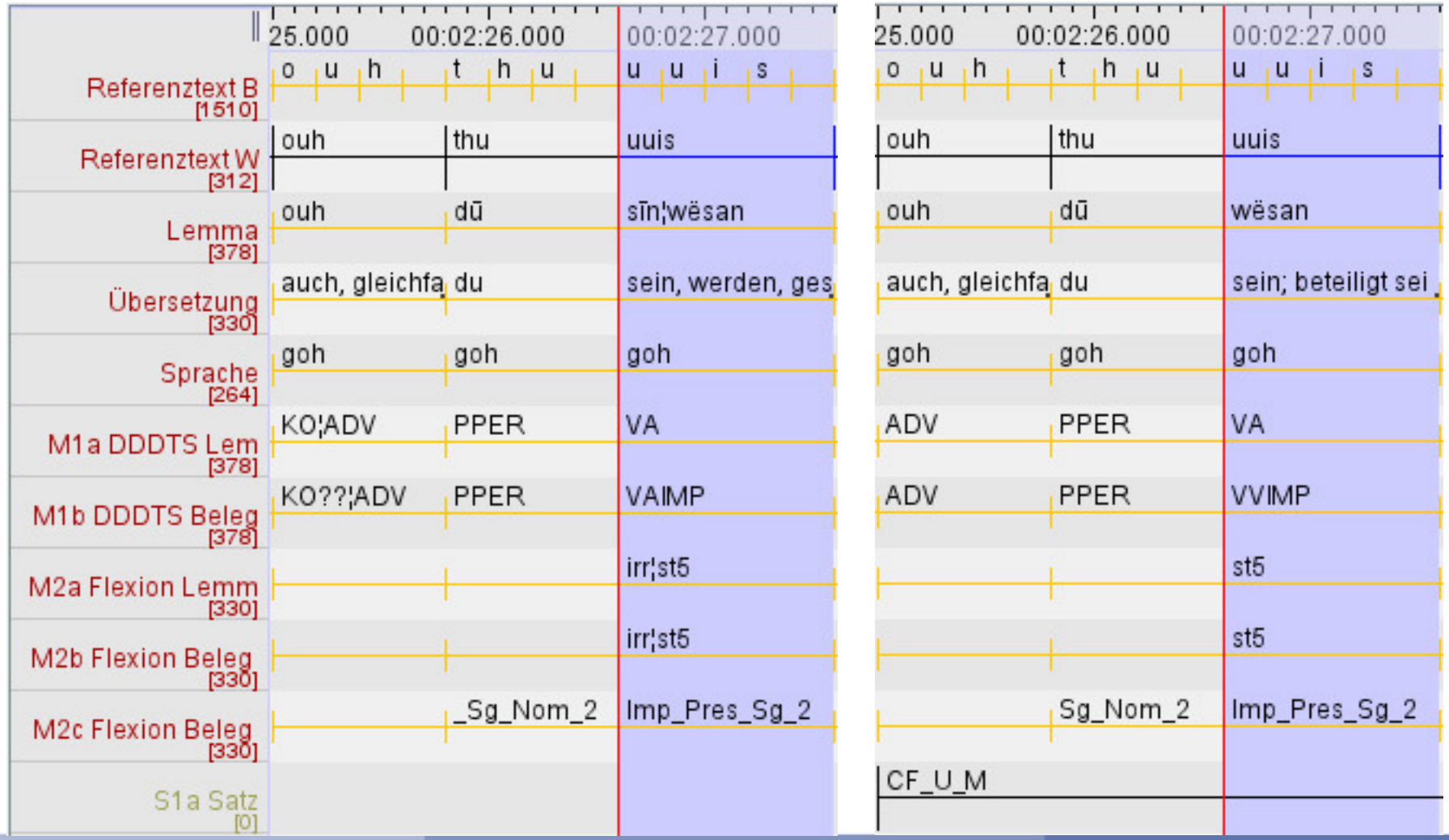
Verbindung der 2 Dateien mit dem Text

- Programm gleicht jedes Textwort mit Belegen in der Glossardaten-Datei ab
- wenn Nummerierung der Belegpositionen in TITUS und dem Glossar identisch ist
  - → 1:1-Zuordnung
  - sonst: Zuordnung aller entsprechenden Datensätze, alle außer einem manuell zu entfernen

→ Wort **uuis** in Beispielphrase
Themo quad her: ouh thu uuis obar fimf burgi. (Tatian 151, 6)
wird korrekt vorannotiert
  - nicht auch als unflektierte Form des Adjektivs *wīs* ‘weise‘

- Ergänzung der Lemmakonkordanzdatei: Ersetzung Lemmata durch Splett-Lemmata, Hinzufügung Übersetzungen
- Überführung der Daten in das Format der Annotationssoftware ELAN (XML-basiert, MPI Nimwegen)

ELAN-Daten vor/nach manueller Annotation



ANNIS-Datenbank (SFB 632 Uni Potsdam)

www.deutschdiachrondigital.de:

`edition="uuis"` (oder über Dropdown-Menüs) → 4 Ergebnisse

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)					

Prüfung der manuellen Annotation

Prüfung der Lemmata

- ggf. Veränderung der Lemmata durch manuelle Annotation
- Prüfung auf ‹ë›/‹ȝ› und Vokallängen
- Vergleich der Lemmata mit Lemmakonkordanz-Datei zu jeweiligem Glossar und Standardwörterbuch
- falls nicht enthalten:
  - Neuerzeugung von Formen mit (bzw. ohne) ‹ë›/‹ȝ›
    - Ausgabe in Log-Datei bei unklarer Lautgestalt
  - erneuter Abgleich, ggf. Korrektur
    - Ausgabe in Log-Datei, ob Korrektur möglich oder nicht
  - Prüfung anderer Vokallängen (lang, kurz, ambig), beginnend am Wortende und in allen möglichen Kombinationen
    - Ausgabe in Log-Datei:
      - bei einem Ergebnis: Ersetzung
      - bei mehreren Ergebnissen: Hinweis auf Ersetzung durch 1. Ergebnis
      - ohne Ergebnis: Hinweis, keine Korrektur
  - erneute Prüfung auf ‹ë›/‹ȝ›
- bei Mehrwortlexemen im Text: Einzellemmaprüfung

Prüfung der Lemmata

Log-Datei mit Beispielfällen aus der Praxis

```
unklare Schreibung (e): leben
Quantitäten-Ersetzung - nicht gefunden: leben -
     stattdessen: lebēn
Korrektur nach Konkordanzliste (e): lëbēn
ERSETZUNG: lëben -> lëbēn !

Korrektur nach Konkordanzliste (z): fizzusheit
ERSETZUNG: fiʒʒusheit -> fizzusheit !

Quantitäten-Ersetzung - nicht gefunden: so - stattdessen: sō
Quantitäten-Ersetzung - nicht gefunden: so - stattdessen: sō
ERSETZUNG: so wio so -> sō wio sō
```

- anschließend manuelle Korrektur nach Log-Datei

Prüfung der morphologischen Annotation

- Erzeugung idealisierter flektierter Wortformen aus Lemmata, Wortart- und morphologischer Annotation
- Ziel:
  - Abgleich dieser Idealwortformen mit tatsächlichen Belegen, um Fehler bei morphologischer Annotation zu erkennen
- Nebeneffekte:
  - idealisierte Wortform kann mit tatsächlicher Wortform verglichen werden, weitere Forschung auf dieser Grundlage möglich
  - idealisierte Wortform kann in Datenbank gesucht werden (anstelle der Angabe morphologischer Eigenschaften)

Prüfung der morphologischen Annotation

- Überführung der Flexionsinformation aus Referenzgrammatiken (ahd./as.) in Computerprogramm
  - dabei zunächst Prüfung nötig, ob in Datei gleiche Zahl an Sprachcodes, Lemmata und Belegen vorhanden
- Übereinstimmung zwischen Standardwörterbuch und Referenzgrammatik nötig:
  - ‚Althochdeutsches Wörterbuch' (J. Splett, 1993):
    - Ahd.: „Sprachstufe, die keine allgemein gültige Leitvarietät besitzt"
    - Entscheidung: „die Idealform des Ostfränkischen, das der Tatian überliefert"
  - ‚Althochdeutsche Grammatik' (W. Braune, 15 Aufl., 1886–2004):
    - „In diesem Buch wird die ostfränk. Sprache des ahd. Tatian zugrundegelegt."

Variation innerhalb der Standardvarietät

- in einigen Fällen gibt Braunes Grammatik mehrere mögliche Flexionsendungen

- z.B. Dativ Plural mehrerer Flexionsklassen:
  -um, -om; -un, -on
- im Tatian: 119x *-un*, 113x *-on*, 5x *-m* ➔ *-un* or *-on*

→ adverbialisierte Dativ-Plurale in Spletts Wörterbuch:
gestaron ‘gestern‘, *zwiron* ‘zweimal‘, *simbalum* ‘immer‘

- Tatian:
  - 1x *gestaron*
  - 1x *zuiron*
  - *simbulun*: 9x *-un*, 2x *-um*, 2x *-on*

→ willkürliche Entscheidung für *-un* als die ältere Form (nach Braune)

Vorliegende Daten: Beispielfall (gekürzt)

Beleg	bigunnon
Lemma	biginnan
Wortart Lemma	VV
Wortart Beleg	VVFIN
Flexion Lemma	st3a,wk1a
Flexion Beleg 1	st3a
Flexion Beleg 2	Ind_Past_Pl_3

Vorliegende Daten: Beispielfall (gekürzt)

Beleg	bigunnon
Lemma	biginnan
Wortart Lemma	VV
Wortart Beleg	VVFIN
Flexion Lemma	st3a,wk1a
Flexion Beleg 1	st3a
Flexion Beleg 2	Ind_Past_Pl_3

Erzeugung der flektierten Lemmata

- Anwendung morphologischer Regeln (reguläre Ausdrücke) auf die Lemmata, z.B.:

Lemma	biginnan
Wortart Beleg	VVFIN
Flexion Beleg 1	st3a
Flexion Beleg 2	Ind_Past_Pl_3

- Lemma: *biginnan* ‘beginnen‘
- VVFIN, Past, Ind, Pl – i?{a|e}n# → Ø *(biginn)*
- st{2|3} – {ë|i|io|ū} → u / _C+# *(bigunn)*
- 3 – a?# → un *(bigunnun* ‘(sie) begannen‘*)*

→ resultierende Wortform: ***bigunnun*** (Beleg: *bigunnon*)

Erzeugung der flektierten Lemmata

- schwach flektierende Form desselben Lemmas *(biginnan)*:

Lemma	biginnan
Wortart Beleg	VVFIN
Flexion Beleg 1	**wk1a**
Flexion Beleg 2	Ind_Past_Pl_3

- VVFIN, Past, wk1a – C_1C_1i?{a|e|ē}n# → C_1ta *(biginta)*
  lemmaspezifische Regel – int → ond / big_a# *(bigonda)*
- Ind, Pl, 3 – a?# → un *(bigondun)*

→ Ergebnis: ***bigondun*** '(sie) begannen'
(Belege: e.g. *bigondun, begonton*)

Sonderfälle

- in einigen Fällen können aufgrund von Lemma und Wortart- und morphologischen Informationen keine ausreichenden Regeln aufgestellt werden, zwei mögliche Formen verbleiben

Lösung 1: Lemma-Liste

→ Lemma-Liste für einen der beiden Fälle erstellen

- in jedem Fall korrekte Ergebnisse, aber großer Aufwand

Lösung 2: belegte Wortformen

→ tatsächlich belegte Wortformen einbeziehen

- einfacherer Ansatz, aber durch Schreibungsvariation könnten sich falsche Resultate ergeben

Sonderfälle

Konsonantenvereinfachung (Substantive und Adjektive)

- im Standard: $C_1C_1 > C_1 / \_\{C,\#\}$

- wenn Lemma auf Konsonant endet, keine Regel für Verdopplung in Flexionsformen möglich: *bal, balles* 'Ball' vs. *wal, wales* 'Wal'

→ automatisierte Erzeugung einer Wortliste (VC#), manuelle Prüfung
→ Reduzierung auf Lemmaliste mit Konsonantenverdopplung

- Homographe mit Unterscheidung bei der Verdopplung:
  - *far, farres* m. 'Stier' vs. *far, fares* n. 'Hafen; Leuchtturm'; adj. 'gehend'
    → Unterscheidung nach Kategorien
  - *ram, rammes* m. 'Widder' vs. *ram, rames* m. 'Rabe; Rahmen'
    → Unterscheidung nach Schreibung des Belegs

Sonderfälle

Vergleichbare Probleme

- Adjektivsteigerung: Komp. *-ōr-/-ir-*, Superl. *-ōst-/-ist-*
  - -i-Steigerung: *ja,jo*-Stämme (i#), irreguläre Stammbildungen
  - -ō-Steigerung: andere mehrsilbige Adjektive (VC+V)
  - → sonst Unterscheidung nach Belegschreibung (Typ variiert oft innerhalb desselben Lemmas!)

- Konsonantenvereinfachung (Verben)
  - *stellen, stellis* ‘stellen‘ vs. *zellen, zelis* ‘(er)zählen‘

Log-Datei mit Beispielfällen aus der Praxis (manuell zu prüfen)

```
807 liupostun -> liobōst  (Sup)
662 caplasan -> giblāsan  (gi- 1)
1037 braenni -> brenni  (Imp)
889 unsero -> uns  (DPOS ohne -er-)
27 ana 28 analegi  ana-_lęgi
```

Sonderfälle

Partizipien der Vergangenheit

- An- und Abwesenheit des Präfixes *gi-* kann nicht errechnet werden

→ Verfahren entsprechend der Belegschreibung

- Setzung nur, wenn Beleg Präfix enthält und Lemma nicht
- auch nach abtrennbaren Präfixen zu setzen:

Beleg	nidargiuualzten
Lemma	nidarwelzen (‘niederbeugen‘)
Wortart Beleg	VVPPA
Flexion Beleg 1	wk1a
Flexion Beleg 2	P_Pos__Pl_Dat_st

- Ergebnis: ***nidargiwalztēn*** ‘(den) Niedergebeugten‘

Sonderfälle

Präverbien

- wenn ein Präverb (‘PTKVZ‘) in einem Text erscheint, wird in beide Richtungen nach dem zugehörigen Verb gesucht
- wenn ein Verballemma das Präverb enthält, der Beleg jedoch nicht, wird das Präverb der erzeugten Idealform getilgt

Beleg	ges	…	úz
Idealform	gās	…	ūʒ
Lemma	ūʒgān	…	ūʒ
Wortart Beleg	VVFIN	…	PTKVZ
Flexion Beleg 1	irr	…	
Flexion Beleg 2	Subj_Pres_Sg_2	…	

Beispielphrase

Beleg	Hich	gio	cote	almactigen
Idealform	ih	jihu	gote	alamahtīgin
Lemma	ih	jëhan	got	alamahtīg
Übersetzung	ich	bekennen	Gott	allmächtig
Sprache	goh	goh	goh	goh
Wortart Lemma	PPER	VV	NA	ADJ
Wortart Record	PPER	VVFIN	NE	ADJN
Flexion Lemma		st5	a_Masc	a,o
Flexion Beleg 1		st5	a_Masc	n
Flexion Beleg 2	Sg_Nom_1	Ind_Pres_Sg_1	Sg_Dat	Pos_Masc_Sg_Dat_wk

Beginn St. Galler Beichte I

Prüfung der morphologischen Annotation

Eigentliches Programm zur Fehlerprüfung

- Reduktion von Beleg und Idealform auf für altdeutsche morphologische Unterschiede relevante Aspekte, u.a.:
  - Tilgung sämtlicher Nichtbuchstaben, Vokallängen und Diakritika
  - Standardisierung von Diphthongen (z.B. *ua*, *ue* > *uo*)
  - Standardisierung von Konsonantenclustern (z.B. *th* > *d; ph* > *pf*)
  - Standardisierung von Ausgängen (z.B. *m#* > *n*; *st#* > *s*)
  - Tilgung von Stimmtonunterschieden
  - Ersetzung von Vokalen in nichtletzter Silbe durch *e*

- Berechnung der absoluten und der relativen Levenshtein-Distanz zwischen beiden Wortformen
  - (Quotient aus durchschnittlicher Wortlänge und) minimale(r) Anzahl von Einfüge-, Lösch- und Ersetzungsoperationen, um die eine Zeichenkette in die andere zu überführen
  - Ergänzung von **3**, wenn unterschiedlicher Auslaut vor Regelanwendung
  - Wenn ein Wert **> 0**, Ausgabe, sortiert nach relativer Distanz

Prüfung der morphologischen Annotation

Log-Datei mit Beispielfällen aus der Praxis (manuell zu prüfen)

```
4.00, 2, gi, ir | ke, ir       | DH_De_Heinrico
3.80, 2, dir, du | ter, tu     | JB_Juengere_Bairische_Beichte
3.75, 2, uns, unsih | uns, ensek      | FP2_Federprobe_II
3.67, 1, des, daz | tes, taz   | BR1_Basler_Rezepte_I
3.36, 2, sculu, sculun | skelu, skelen | FP2_Federprobe_II
3.25, 2, slaphanto, slafenti | slepento, slefente    | JB
0.75, 3, hab, haben | kan, kepen      | GGB3_Sang_Gl_B_III
0.67, 3, rib, anarib | rep, enerep     | BR2_Basler_Rezepte_II
0.36, 2, unielih, wiolih | enelek, felek      | CH_Chr_und_Sam
0.33, 2, giscufe, scuofi | keskefe, skefe     | KB_Klostern_G
0.18, 1, zesuun, zesun | zesfen, zesen        | GC_StGallCredo
```

www.deutschdiachrondigital.de

Vielen Dank für Ihre Aufmerksamkeit!
Dankōn iuwih furi iuwera anadāht!
Thankon iu for iuwa waru!