



Grundlagen der Annotation und Glossierung im Kali-Korpus

Grundsätzlich gilt: Annotiert werden sämtliche Verben. Komplett glossiert werden lediglich: Notker von St. Gallen, „De consolatione philosophiae“ (althochdeutsch; in Vorbereitung) und das Alexanderlied (mittelhochdeutsch).

Grundlagen der Annotation im Kali-Korpus:

§ 1 Annotation

- (1) Annotiert werden nur Verben.
- (2) In einer ersten Stufe werden Verben unterschieden nach den Kriterien der *Finitheit* als:
 - (a) *infinit*,
 - (b) *semifinit* (Imperative) oder
 - (c) *finit*.
- (3) Bei den *infiniten* Formen wird unterschieden zwischen:
 - (a) *Infinitiven*,
 - (b) *Partizip I* (Partizip Präsens) oder
 - (c) *Partizip II* (Partizip Perfekt).
- (4) Alle *infiniten* Formen werden zusätzlich klassifiziert als:
 - (a) *flektiert* oder
 - (b) *nicht flektiert*.Die Form der jeweiligen Flexion wird nicht weiter differenziert.
- (5) *Semifinite* Formen werden unterschieden nach:
 - (a) *Singular* oder
 - (b) *Plural*.Eine Unterscheidung der Person findet nicht statt, da in der überwiegenden Zahl der Fälle die 2. Person vorliegt und nur in Ausnahmefällen die 1. Person.
- (6) Bei den *finiten* Formen werden folgende Informationen annotiert:
 - (a) *Tempus* (Präsens oder Präteritum),¹
 - (b) *Modus* (Indikativ, Konjunktiv oder unklar),
 - (c) *Person* (1., 2. oder 3.) sowie
 - (d) *Numerus* (Singular oder Plural).Dabei steht *Präsens Konjunktiv* für den Konjunktiv I und *Präteritum Konjunktiv* für Konjunktiv II. Die Bezeichnung *unklar* signalisiert, dass eine eindeutige Zuordnung des Modus nicht möglich ist.

¹ Bei den Tempusformen Perfekt und Plusquamperfekt wird jedes Element der Periphrase als eigenständiges Verb annotiert.



```

æs | erdha | . | Dhuo | ir | himilo |
. iui | frumida | , | dhar | uuas |
dhanne | | mit | erena | euua |
. ndiu | uu | Verb (finit), Präteritum, Indikativ, 3. Person Singular, frumen, frummen
jida | , | dhuo | ir | erdha | ste-
uuac | , | mit | imu | uuas |

```

- (7) Auf die Markierung von *Negation* (auch beim Auftreten von Negationspartikeln am Verbstamm) wird verzichtet.

```

wi | hęc | non | modo | credenda | . | sed | saltim | audienda | uideantur | ? | Sĉ
sô | néahŧôn | ih | nĳeht | uuâreren | . | úbe | ih | án_dero | líuto | áhtâ | déncl
iuben | . | | e | sâr | gehôren | uuélle | ? |
| LUMEN | RATIONIS | NON | CLARESCERE | . |

```

§ 2 Lemmatisierung

- (1) Jeder annotierten Verbform werden Informationen über ihre *Grundform(en)* in der jeweiligen Sprachstufe zugewiesen. Als Grundlage dienen dabei grundsätzlich:
 - (a) Für das *Althochdeutsche*: R. Schützeichel, Althochdeutsches Wörterbuch, Tübingen: Niemeyer.
 - (b) Für das *Mittelhochdeutsche*: M. Lexer, Mittelhochdeutsches Taschenwörterbuch, Stuttgart: Hirzel.

Als Grundformen eines Lemmas gelten dabei sämtliche Formen, die in den bezeichneten Wörterbüchern fettgedruckt sind. Durch Klammern gekennzeichnete Schreibvarianten werden in zwei separate Einträge aufgelöst.
 Bsp.: heiz(z)an: heizan, heizzan
 Dies gilt auch für Verben, bei denen an einer Grundform gleichzeitig Kürze und Länge des Vokals gekennzeichnet sind.
 Davon abgesehen orientiert sich die Schreibung der Grundformen an jener der jeweiligen Wörterbücher.
- (2) Neben den Grundformen eines Lemmas werden auch seine neuhochdeutschen *Übersetzungen* angeführt. Hierbei dienen wiederum die Wörterbücher von Schützeichel und Lexer als Grundlage. Sonderbedeutungen, idiomatische Verwendungsweisen sowie Übersetzungsvarianten bei Kombination mit speziellen Kasus werden nicht aufgeführt.
- (3) Lassen sich Verben eindeutig als *Präfixverben* klassifizieren, werden sie mit ihrer präfigierten Grundform annotiert, selbst wenn das Präfix getrennt vom Verbstamm steht. Die Verknüpfung des trennbaren Präfixes mit dem Stamm wird farblich markiert. Die Annotationsinformationen des Präfixverbs können auch am Präfix abgelesen werden.
 Das „Mittelhochdeutsche Taschenwörterbuch“ von M. Lexer führt Präfixverben häufig nicht als eigene Lemmata, so dass zur Bestimmung der Grundformen das



„Mittelhochdeutsche Handwörterbuch“ herangezogen wird. Dieses gibt jedoch nicht immer neuhochdeutsche Übersetzungen an. In diesen Fällen wird auf Basis von Präfix und Verbstamm eine passende Übersetzung erstellt. Derart interpretatives Vorgehen wird durch einen kurzen Kommentar dokumentiert.

```

-----
e | unde | ze | mieten | , |
r | in | abe | huobe |
n | sô | s | he | begrubo | ! |
t | gewa | Verb (finit), Präteritum, Indikativ, 2. Person Singular, abe heben
rbe | er | her | in | werft | chom | , |
z | er | sînen | lîchnamen |

```

- (4) Bei den ahd. Verben *gân/gangan* und *stân/stantan*, sind beide Formen als Grundform verzeichnet. Dies gilt, obwohl in den Wörterbüchern separate Einträge aufgeführt sind. Die angegebenen Übersetzungen basieren auf jenen der kürzeren Form.

```

tinode | , | endi | sines | mundes | gheistu |
standit | al | iro | meghin' | . | In | dhemu |
druhtî | s | nemin | archennemes |
chiuui | Verb (finit), Präsens, Indikativ, 3. Person Singular, stantan, stân
chilaubemes | sunu | . | In | sines | mun-
des | gheistu | lîntandemes | lîhi

```

- (5) Einen Sonderfall bei der Zuweisung der Grundformen stellt ahd. *sîn/wesan* bzw. mhd. *sîn/wësen* dar. Während im Althochdeutschen zwischen den Grundformen noch klar differenziert wird, ist diese Trennung im Mittelhochdeutschen aufgehoben. Lediglich der mhd. Infinitiv *wësen* wird auch der Grundform *wësen* zugewiesen. Alle anderen Formen sind unter der Grundform *sîn* verzeichnet.
- (6) Verben mit dem Präfix *gi-/ge-* werden mit der entsprechenden präfigierten Grundform aufgeführt, wenn sie sich eindeutig zuordnen lassen. Formen im Partizip II sind im Allgemeinen nicht eindeutig.

§ 3 Material

- (1) In erster Linie relevant für die Lemmatisierung der Verben sind:
- Für das Althochdeutsche: Rudolf Schützeichel, Althochdeutsches Wörterbuch, Tübingen: Niemeyer ⁵1995.
 - Für das Mittelhochdeutsche: Matthias Lexer, Mittelhochdeutsches Taschenwörterbuch, Stuttgart: Hirzel ³1885 [ND 1999].
- (2) Für die Annotation einzelner Texte stehen detaillierte Glossare zur Verfügung. Dies sind:
- Für Tatians Evangelienharmonie: Eduard Sievers (Hg.), Tatian. Lateinisch und altdeutsch mit ausführlichem Glossar, Paderborn: Schöningh 1892 (Bibliothek der älteren Literaturdenkmäler 5).
 - Für Notkers De Consolatione Philosophiae: Edward H. Sehr, William K. Legner



(Hgg.), Notker-Wortschatz, Halle/Saale: Niemeyer 1955.

(c) Für die Kaiserchronik: W.F. Tulasiewicz, Index Verborum zur deutschen Kaiserchronik, Berlin: Akad.-Verl. 1972 (Deutsche Texte des Mittelalters 68).

Grundlagen der Glossierung im Kali-Korpus:

Glossiert wird auf Basis der [Leipzig Glossing Rules](#). Abweichungen vom Standard und spezielle Eigenarten der Glossierung im Kali-Korpus werden im Folgenden dokumentiert.

§ 1 Glossierung

- (1) Die Glossierung einzelner Tokens erfolgt vertikal, jede Zeile beinhaltet feste Angaben:
 - (a) Zeile 1: *Token*
 - (b) Zeile 2: *Glosse*
 - (c) Zeile 3: *Grundform*
 - (d) Zeile 4: *Übersetzung*

Diz	lit	, daz
D.AKK.SG.N	Lied[AKK.SG]	D.AKK.SG.N
dër, diu, daz/dez D	liet stn.	dër, diu, daz/dez D
der, die, das	Lied	der, die, das

Jede Zeile trägt Information. Wenn einzelne Tokens nicht glossiert werden, enthält die Zeile „Glosse“ eine Übersetzung des Tokens mit Angaben zur Wortart.

- (2) Bei der Glossierung werden folgende Wortarten unterschieden:
 - (a) *Determinator* [D]: Artikel, Relativpronomen und Demonstrativpronomen
 - (b) *Possessivpronomen* [POSS]
 - (c) *Personalpronomen* [PRON]
 - (d) *Indefinitpronomen* [INDF]
 - (e) *Interrogativpronomen* [ITR]
 - (f) *Numeral* [NUM]: unbest. Artikel/Indefinitpronomen/Zahlwort *ein*
 - (g) *Konjunktion* [KON]
 - (h) *Verb* [stv., swv., an.v.]
 - (i) *Adjektiv* [ADJ]
 - (j) *Adverb* [ADV]
 - (k) *Nomen* [stm., stf., stn., swm., swf., swn.]
 - (l) *Interjektion* [INT]
 - (m) *Präposition* [PRÄP]
 - (n) *Partikel* [PK]
 - (o) *Reflexivpronomen* [REFL]
- (3) Die Bestimmung der Wortart und der Grundform erfolgt mithilfe von Wörterbüchern. Als Referenzwerke zugrundegelegt werden:



AHD: R. Schützeichel, Althochdeutsches Wörterbuch, Tübingen: Niemeyer.

MHD: M. Lexer, Mittelhochdeutsches Taschenwörterbuch, Stuttgart: Hirzel.

Lassen die Wörterbücher für einzelne Lemmata mehrere Interpretationen zu, wird aus dem Kontext entschieden, wie ein Token einzuordnen ist.

- (4) Segmentiert werden Präfixe und Flexionsendungen. Dies gilt für die Zeilen *Token* und *Glosse* gleichermaßen. Innere Flexion, die nicht vom Stamm separiert werden kann, wird ohne Bindestrich mit einem einfachen Leerschritt an den Stamm angefügt. Grammatische Information, die nicht an der Oberfläche realisiert wird, ist in eckigen Klammern ohne Leerschritt direkt an den Stamm angeschlossen.

<p>märe Märe[DAT.SG] mære str. Märe, Geschichte, Kunde, Bericht</p>

- (5) Zwischen einzelnen grammatischen Informationen werden Punkte gesetzt (eine Ausnahme bilden lediglich die Angaben zu Person und Numerus, die nicht durch einen Punkt voneinander getrennt werden). Dabei sind für die einzelnen Wortarten folgende Informationen festgehalten:
- (a) D: *Kasus.Numerus.Genus*(ggf.)
(b) POSS: *Person.Numerus.Genus* (Stamm) – *Kasus.Numerus.Genus* (ggf.) (Flex.)

<p>sin-iu POSS.3SG.M-NOM.PL sîn POSS sein</p>

- (c) PRON: *PersonNumerus.Genus*(ggf.).*Kasus*
(d) INDF: nur Übersetzung
(e) ITR: ahd.: *Kasus.Numerus.Genus*, mhd.: nur Übersetzung (ggf. mit *Kasus*)
(f) NUM: *Kasus.Genus*
(g) KON: nur Übersetzung
(h) Verb: *(Präfix-)stamm-Tempus.Modus.PersonNumerus*
(i) ADJ: *Kasus.Numerus*
(j) ADV: nur Übersetzung
(k) Nomen: *Kasus.Numerus*
(l) INT: nur Übersetzung
(m) PRÄP: nur Übersetzung
(n) PK: nur Übersetzung
(o) REFL: nur Übersetzung

Für alle Pronomen und Determinatoren gilt, dass das Genus nur markiert wird, wenn die Zuweisung eindeutig erfolgen kann. Das Genus von Nomen und Adjektiven wird zusammen mit der Grundform angegeben.



- (6) Besonderheiten bei der Glossierung von Verben:
- (a) Steht ein Verb im Indikativ, wird das nicht extra dokumentiert. Lediglich Formen im Konjunktiv werden als solche gekennzeichnet.
 - (b) Für Formen des PART.PRÄS bzw. PART.PERF werden die Bezeichnungen PARTI bzw. PARTII verwendet.
 - (c) Flektierte Infinitive und Partizipien werden mit *fl* markiert. Liegt ein Flexionsmorphem vor, wird dieses durch Bindestrich abgetrennt, ansonsten wird der Flexionsmarker direkt angeschlossen.
 - (d) Kurze Verben wie *sein* oder *tun* werden in der Glosse mit der vollen Infinitivform angegeben, während bei den übrigen Verben neben der Flexionsinformation nur der Stamm angegeben ist.
 - (e) Untrennbare Präfixe – besonders von Partizipien – werden durch einen Bindestrich vom Stamm getrennt, aber ausgeschrieben. Auf eine mögliche Einordnung als Zirkumfix wird verzichtet.

<p>ir-ging-en er-geh-PRÄT.3PL ergân, ergên stv.red.an zu Ende gehen, geschehen, sich ereignen, sich ergehen</p>	.
---	---

- (f) Kontraktionen werden durch den Vermerk KONTR als solche gekennzeichnet. Kontrahierte Elemente werden durch ein Gleichheitszeichen voneinander getrennt.

<p>het=er hab PRÄT.3SG = PRON.3SG.M.NOM haben, hân swv. = er PRON KONTR haben = er</p>
--

§ 2 Sigelverzeichnis

(lat.)	lateinisch
1	1. Person
2	2. Person
3	3. Person
ABL	Ablativ
ADJ	Adjektiv
ADV	Adverb
AKK	Akkusativ
an	anormal (unregelmäßig)
CARD	Kardinalzahl
D	Determinator (Artikel, Demonstrativpronomen, Relativpronomen)
DAT	Dativ
f	Femininum



fl	flektiert
GEN	Genitiv
IND	Indikativ
INDF	Indefinitpronomen
INF	Infinitiv
INT	Interjektion
ITR	Interrogativpronomen
KOMP	Komparativ
KON	Konjunktion
KONJ	Konjunktiv
KONTR	Kontraktion
m	Maskulinum
n	Neutrum
NEG	Negation
NOM	Nominativ
NUM	Numeral
PART	Partizip
PARTI	Partizip I bzw. Partizip Präsens
PARTII	Partizip II bzw. Partizip Perfekt
PK	Partikel
PL	Plural
POS	Positiv
POSS	Possessivpronomen
PRÄF	Präfix
PRÄP	Präposition
PRON	Pronomen
PRON-ADJ	Pronominaladjektiv
PRON-SUB	Pronominalsubstantiv
PRPR	Präteritopräsens
red	reduplizierend
SG	Singular
SMF	Semifinit
st	stark
subst	substantiviert
SUBST	Substantiv
SUP	Superlativ
sw	schwach
unfl	unflektiert
v	Verb

Kali_{Korpus} <http://www.kali.uni-hannover.de>



Versionshinweise

Version 1.0, erstellt am 25.05.2008 von

Dr. phil. Elena Smirnova – elena.smirnova@germanistik.uni-hannover.de

Anne Jäger – anne.jaeger@germanistik.uni-hannover.de