Volume 31 — Number 2 — 2016 — ISSN 2190-6858

JLCL
Journal for Language Technology
and Computational Linguistics

Corpora and Resources for (Historical) Low Resource Languages

Herausgegeben von Edited by
Armin Hoenen, Alexander Mehler, Jost Gippert

GSCL Gesellschaft für Sprachtechnologie & Computerlinguistik

Contents

Impressum

Herausgeber	Gesellschaft für Sprachtechnologie und Computerlinguistik (GSCL)
Aktuelle Ausgabe	Band 31 – 2016 – Heft 2
Gastherausgeber	Armin Hoenen, Alexander Mehler, Jost Gippert
Anschrift der Redaktion	Lothar Lemnitzer Berlin-Brandenburgische Akademie der Wissenschaften Jägerstr. 22/23 10117 Berlin lemnitzer@bbaw.de
ISSN	2190-6858
Erscheinungsweise	2 Hefte im Jahr, Publikation nur elektronisch
Online-Präsenz	www.jlcl.org

Armin Hoenen, Alexander Mehler, Jost Gippert

Editorial

Im Februar 2016 fand an der Goethe Universität Frankfurt der Workshop "Corpora and Resources for Low Resource Languages with a Special Focus on Historical Languages" oder kurz CRiLL-HL statt.[1] Er wurde in Kooperation der GSCL-Arbeitskreise *Korpuslinguistik* und *Historisch-Vergleichende Sprachwissenschaft* mit dem *Centrum für Digitale Forschung in den Geistes-, Sozial- und Bildungswissenschaften* (CEDIFOR) der Goethe-Universität veranstaltet. Während für viele, vor allem für größere Sprachen mittlerweile eine gute bis sehr gute technologische Infrastruktur bereitsteht (d.m. in Bezug auf die Verfügbarkeit von Ressourcen einerseits und die Verfügbarkeit von grundlegenden Technologien andererseits), ist dies im Bereich so genannter *Low Resource Languages* (LRL), solcher Sprachen also, welche aus unterschiedlichsten Gründen wenig Zugang zu Ressourcen wie Korpora aller Art, Lexika, Grammatiken usw. aufweisen, noch nicht der Fall. Dies steht im Gegensatz zur großen Bedeutung dieser Sprachen, welche nicht nur in Europa selbst für einen Großteil der linguistischen Diversität verantwortlich sind. Die Situation verbessert sich in mancher Hinsicht stetig durch große Infrastrukturprojekte und Initiativen, sowie Organisationen, welche sich der Erschließung von LRL verschrieben haben. So sind beispielsweise CLARIN[2] mit seinem *Language Resource Inventory* oder die ELRA[3] zu nennen, welche für einen stetig besser werdenden Zugang zu Sprachressourcen sorgen. Unter anderem bedrohte Sprachen werden durch Projekte wie DOBES[4] noch einmal besonders ins Auge gefasst, da ihr unmittelbares Verschwinden droht.
WissenschaftlerInnen, die in einem dieser Kontexte zu LRL Sprachen forschen, sehen sich teilweise aber noch immer mit einer Reihe spezieller, schwer lösbarer Probleme konfrontiert, für deren Diskussion der Workshop ein Forum bieten und sich so in die Bestrebungen um eine bessere Verarbeitbarkeit der genannten Sprachen einreihen wollte. Besonders im Bereich der historischen Sprachen, welche innerhalb der LRL noch einmal eine besondere Stellung einnehmen, fand ein reger wissenschaftlicher Austausch statt. Dies betraf nicht nur die Präsentationen entsprechender Beiträge, sondern auch die Arbeit in themenorientierten Arbeitsgruppen, in welchen die TeilnehmerInnen spezielle Verfahrensweisen (wie z.B. die Lemmatisierung historischer Texte) intensiv diskutierten. In Bezug auf Annotationen korrespondieren einige der diskutierten Themen mit Fragestellungen, wie sie das kürzlich erschienene *Handbook of linguistic Annotation* thematisiert, was einmal mehr die Aktualität des Workshop-Themas unterstreicht.
Das vorliegende Heft des JLCL versammelt im Nachgang zu diesem Workshop nunmehr ausgewählte Beiträge, welche in diesem Kontext entstanden sind:

1. Der erste Beitrag von Florian Petran, Thomas Klein, Stefanie Dipper und Marcel Bollmann stellt mit ReM ein Referenzkorpus des Mittelhochdeutschen vor. Dabei

1 Informationen zum Workshop findet der interessierte Leser auch unter der Adresse http://gscl-ak-korpuslinguistik.hucompute.org

2 https://www.clarin.eu/content/language-resource-inventory

3 http://www.elra.info/en

4 http://dobes.mpi.nl

werden Korpusgenese, Quellen, Struktur und Annotationen genau beschrieben und mit Beispielen ausgeführt. ReM ist Teil eines bundesweiten Projektes zur Schaffung von Referenzkorpora für historische Sprachstufen des Deutschen. Es wurde semi-automatisch mit Annotationen angereichert, so u.a. im Hinblick auf Tokenisierung, Normalisierung, Parts of Speech, morphologische Analyse, Lemmata, wodurch eine Vielzahl weitergehender Analysen ermöglicht wird. Insgesamt umfasst ReM ca. 2,5 Millionen Token (in ca. 400 Texten).

2. Der zweite Beitrag von Roland Mittmann beschreibt eine Methode zur automatischen dialektalen Einordnung althochdeutscher Wortformen. Der Autor stellt das Konzept dieser Methode vor, welches auf aus Grammatiken extrahierten relativen Lautentsprechungen und deren grammatikalischen Funktionen beruht, und demonstriert erste Ergebnisse, welche auf die vielversprechende Anwendbarkeit seiner regel-basierten Methode zum Zwecke der automatischen Einordnung althochdeutscher Texte in Zeit-Dialekträume schließen lassen.
3. Der dritte Beitrag von Armin Hoenen und Lela Samushia stellt ein altgeorgisches Inschriftenkorpus vor, welches im Format der TEI (EpiDoc) codiert wurde, und erörtert die spezifischen Probleme, die dieser Texttyp an die technologische Verarbeitung stellt. Als *proof-of-concept* präsentieren Hoenen und Samushia Front- und Backend eines Tools, das aufzeigt, welche Eigenschaften wichtig sind, um bei der Entschlüsselung und Rekonstruktion der Botschaft von oft nur fragmentarisch überlieferten Inschriften zu helfen. Dabei kommen *language models*, *word embeddings* und frequenzbasierte Statistiken zum Einsatz.

Wir danken allen Gutachtern, der GSCL, den Herausgebern des JLCL sowie dem CEDIFOR für die gewährte Unterstützung und wünschen den LeserInnen ein angenehmes und hoffentlich erkenntnisreiches Leseerlebnis.

Armin Hoenen, Alexander Mehler und Jost Gippert
(Juli 2017, Frankfurt am Main)

Literatur

Ide, N., & Pustejovsky, J. (Eds.). (2017). *Handbook of Linguistic Annotation*. Springer.

Florian Petran, Marcel Bollmann, Stefanie Dipper, Thomas Klein

ReM: A reference corpus of Middle High German — corpus compilation, annotation, and access

1 Introduction

In recent times, there has been a growing interest in digitized and annotated corpora of historical language data, coming from both historical linguists as well as the emerging historico-cultural domain of digital humanities. For German, an initiative with the goal of creating a diachronic reference corpus was started in the 2000s, which has so far yielded four different research projects:[1]

- *Reference Corpus Old German* (ReA, 750–1050),
- *Reference Corpus Middle High German* (ReM, 1050–1350),
- *Reference Corpus Early New High German* (ReF, 1350–1650), and
- *Reference Corpus Middle Low German and Low Rhenish* (ReN, 1200–1650).

This paper describes ReM and the results of the ReM project and its predecessors. All projects closely collaborate in developing common annotation standards to allow for diachronic investigations. ReA has already been published and made available via the corpus search tool ANNIS[2] (Krause and Zeldes, 2016), while ReF and ReN are still in the annotation process.

The ReM project builds on several earlier annotation efforts, such as the corpus of the new Middle High German Grammar (MiGraKo, Klein et al. (2009)), expanding them and adding further texts, to produce a reference corpus for Middle High German, which we will also call "ReM" for short. The combined corpus, which consists of around two million tokens, provides a mostly complete collection of written records from Early Middle High German (1050–1200) as well as a selection of Middle High German texts from 1200 to 1350. Texts have been digitized and annotated with parts of speech and morphology (using the HiTS tagset, cf. Dipper et al. (2013)) as well as lemma information.

Release 1.0 of ReM has been published in December 2016 and is also accessible via the ANNIS tool. The project website at `https://www.linguistics.ruhr-uni-bochum.de/rem/` offers extensive documentation of the project and the corpus. The corpus

[1] ReA project: `http://www.deutschdiachrondigital.de/home/?lang=en`, ReM project: `https://www.linguistics.ruhr-uni-bochum.de/rem`, ReF project: `http://www.ruhr-uni-bochum.de/wegera/ref/`, ReN project: `https://vs1.corpora.uni-hamburg.de/ren/`

[2] `http://corpus-tools.org/annis/`

design as well as the transcription and annotation guidelines are described in Klein and Dipper (2016).

In the remainder of this paper, we briefly discuss the textual basis of the corpus (Sec. 2) and its annotation layers (Sec. 3). Sec. 4 explains the semi-automatic annotation process and the tools used for it, some of which date back to the mid to late 1980s. In Sec 5 we present the XML based document format that will be used to distribute the corpus. Sec. 6 deals with the presentation of the corpus in ANNIS.

2 Textual basis

The reference corpus of Middle High German (ReM) combines the work of several different research efforts:

1. the Cologne corpus of Hessian-Thuringian texts (created between 1986 and 1993; cf. Klein and Bumke (1997));
2. the Bonn corpus of Middle German texts (created from 1993 onwards);
3. the Bochum Middle High German corpus (BoMiKo) and its successor, the corpus of the Middle High German grammar (MiGraKo[3], Klein et al. (2009)); and
4. an extension/supplement of the aforementioned corpora, created during the ReM project.

MiGraKo is a balanced and structured corpus, composed of roughly equally-sized texts and text extracts from different dialect areas, time periods and text sorts (cf. Wegera, 2000). It already incorporates some of the texts annotated in the Cologne and Bonn corpora that preceded it. In total, MiGraKo consists of 102 texts and about 1,25 million tokens. The main goal of the ReM project was to create an even larger reference corpus of Middle High German, by combining data from all of the preceding projects, adding more texts, and also extending some of the existing annotations.

We distinguish two time periods within the corpus. The first half from ca. 1050 to ca. 1200, called Early Middle High German, is more important for the historical development of the German language, regarding the transition from Old High German, but also some of the beginnings of the development of New High German. At the same time, text sources from that period are scarce, so that it is hardly possible to obtain a structured and balanced selection. For that reason, the ReM corpus includes a mostly complete record of all available Early Middle High German texts, with the exception of a few heavily fragmented sources and those which are merely copies of an older text. Overall, the first part of the corpus includes about 700,000 tokens in 184 texts between 6 and 59,000 tokens in length.

For the second part of the corpus, the later Middle High German period, the availability of sources is much better. Here, the focus was on extending and supplementing

[3] `http://www.ruhr-uni-bochum.de/wegera/MiGraKo/`

the selection of texts in the MiGraKo corpus. In general, the selection is more diverse as the underlying MiGraKo part, e.g. including heterogeneous texts written by different authors in different dialects, texts whose manuscripts are considerably younger than the text's presumable time of origin, or larger text segments that are suitable for syntactic analyses. This part has 214 texts with between 20 and 55,000 tokens each, totalling about 1.8 million tokens.

The entire ReM corpus consists of around 2.5 million tokens.

3 Transcription and annotation

The earliest transcriptions and annotations, and with it the earliest version of the guidelines, date back to 1986. Therefore, they still reflect the computer technology of the 1980s in many ways.

The original transcriptions of the ReM texts served two goals. First, they encoded fine-grained properties of the historical word forms, resulting in a diplomatic transcription. The transcriptions used special characters and markup to encode historical graphemes, diacritics and abbreviations. For instance, '$' encoded historical 'f', 'o\v' stood for 'o͛', and 'o\-' for 'ō'.

Second, the original transcriptions encoded information about modern word boundaries, thus supporting further (semi-)automatic processing of the word forms. That is, markup was used to indicate modern word boundaries in cases where the historical word forms, as marked by whitespace, did not correspond to modern word forms. For instance, the historical form 'biftu' ('are you') would be transcribed as 'bi$|tu'. The vertical bar indicated a modern word boundary because the historical form corresponds to two word forms according to modern spelling rules: 'bif' + 'tu' ('are' + 'you').

In ReM corpus, this information has been projected to two different layers, called "diplomatic" (dipl) and "annotated" (anno). The diplomatic layer records historical graphemes, by converting special encodings for historical characters to appropriate UTF characters. The diplomatic layer also conserves original word boundaries and line breaks. The annotated layer uses ASCII characters only and adapts word boundaries to the rules of modern German. For an example, see (1).

(1) **dipl** fo biftu
anno so bis tu
'so you are'

Both the diplomatic and the modernized layers are annotated with further information. Each diplomatic token is assigned its exact location in the text (page number, line number, column, etc.).[4] All further annotations refer to the annotated token layer. These are:

[4] In some cases the original manuscript was lost or destroyed, in those cases the diplomatic tokens are assigned their location in the edition used for the transcription

Normalization (norm) This layer contains automatically-created word forms that closely correspond to word forms as used in traditional editions of historical manuscripts in German. For instance, a diplomatic form like 'chindelin' ('children') is mapped to the form 'kindelîn'.

Tokenization (tokenization) This layer annotates cases of diverging word boundaries, as in Ex. (1). The annotation follows the HiTS guidelines (Dipper et al., 2013). The tags encode two properties: first, whether the modernized form is a merger of several historical forms to one modern form (*Univerbierung*, label `U`), or a case of splitting one historical form to multiple modern ones (*Multiverbierung*, labels `M..1, M..2`, etc. for the different forms). Second, the tags also encode which character is used at the word boundary: a space (label `S`), a hyphen (`H`), or camel case, i.e. a word-internal capitalized letter (*Binnenmajuskel*, `B`). It is also encoded if the tokenization involves a line break (`L`). For some examples, see (2) (line breaks are marked by '⸗').

(2) a.

dipl	fo	biftu	
anno	so	bis	tu
tok		MS1	MS2

'so you are'

b.

dipl	Alfo	der	lichaname	er	ſtír	⸗bet
anno	Also	der	lichaname	erstirbet		
tok	–	–	–	US UL		

'as the body dies'

c.

dipl	be	durfeter	
anno	bedurfet		er
tok	US MS1		MS2

"you[pl] need"

Punctuation (punc) This layer encodes original punctuation marks and modern sentence and clause boundaries. Original punctuation marks correspond to modern sentence or clause boundaries in about 2/3 of the cases.

Modern boundaries are always annotated at the last (modernized) word in the sentence or clause. Labels used here are `DE, EE, IE, QE`, which stands for "end of a declarative / exclamative / imperative / interrogative clause". Other segment boundaries that are annotated include dependent and appositive clauses and enumerations (labels `S*, N*, NE`).

Original punctuation marks that correspond to some segment boundary are annotated with the tag `$E`, see (3).

(3)

dipl	fo	ne	mach	i̇nemen	gotegelichen		·
anno	so	ne	mach	niemen	gote	gelichen	.
punc						DE	$E

'so nobody can be like god'

Linguistic annotations: part of speech (pos), morphology (infl), lemma The original annotations have been created semi-automatically (Klein, 2001). In the ReM corpus, they have been mapped to tags that largely follow the HiTS guidelines (Dipper et al., 2013). This means, among other things, that words are annotated in two ways, once as a token (instance) and once as a type. The token annotation takes the actual context into account, type annotation encodes general properties of a word. Ex. (4) shows that the word 'gebornen' ('born') is basically a verb (past participle), which in this context is used like an adjective. Hence, the type is annotated with the part of speech "VVPP" (verb past participle), and the token is annotated with "ADJN" (postnominal adjective).

(4)

dipl	diu	chindelin	niu	gebornen
anno	diu	chindelin	niu	gebornen
norm	diu	kindelîn	niu	geborenen
pos (token)	DDART	NA	ADJD	ADJN
pos (type)	DD	NA	ADJ	VVPP
lemma	der	kindelîn	niuwe	ge-bor(e)n
lemmaID	29817000	89652000	121830000	48162000
infl	Neut.Nom.Pl	Nom.Pl	Pos.Neut.Nom.Pl.0	–
inflClass	–	st.Neut	–	–

'the newborn children'

In addition to the lemma, a lemma ID is also provided, which links to the corresponding lemma of the online lexicon 'Mittelhochdeutsches Wörterbuch'[5].

In Ex. (4), the layer inflClass refers to the token-specific inflection class. It is specified for nouns and verbs and represents the declension or conjugation class of the respective lemmas, in the given context. In the case of nouns, a preceding article and/or adjective can help in determining the gender of a noun (e.g. 'Neut'). For instance, like many other nouns in Middle High German, the lemma 'slange' ('snake') is underspecified for gender and frequently occurs in masculine or feminine gender. Ex. (5) shows examples where the context helps (a) or does not help (b) in disambiguating gender. The layer infl-class (type) shows the general, ambiguous properties of the noun, the layer infl-class (token) the context-specific features.

(5) a.

dipl	So	der	hirz	den	flangen	fihit
inflClass (token)	–	–	st.Masc	–	wk.Masc	–
inflClass (type)	–	–	st.Masc	–	wk.Masc,Fem	–

'as the deer sees the snake'

b.

dipl	Vō	flangē
inflClass (token)	–	wk.Masc,Fem
inflClass (type)	–	wk.Masc,Fem

'of snakes'

[5] http://www.mhdwb-online.de/lemmaliste/

Character alignments (char) Finally there is a layer that aligns characters from the annotated with the normalized forms. For instance, a word pair such as 'chindelin'–'kindelîn' ('children') gives rise to the mappings ch=k, i=i, n=n, d=d, e=e, l=l, i=î, n=n. The mappings can be used to investigate spelling variation between different dialect regions.

4 Semi-automatic annotation

Owing to the history of the corpus (cf. Sec. 2), the annotation process as a whole was quite eclectic. The pioneering work on the Cologne corpus used a suite of programs written in Macro SPITBOL for semi-automatic, rule-based part-of-speech and morphology annotation (Klein, 1991). At the core of this suite is an annotated index of normalized forms of Middle High German words based on the modernized tokenization.

The form to be annotated is analyzed with the known character alignments for Middle High German spelling and dialectal variations and inflectional affixes. Based on this analysis, a ranked list of approximate matches is returned from the normal form index. The list has lemma and part-of-speech (POS) annotations, as well as a pre-selection of possible morphology annotations for the recognized affixes. The index already has rankings according to the naive probability of each suggestion; an additional basic rule-based syntactic analysis re-ranks the suggestions appropriately for the token context. A human annotator then selects the correct annotation from the list, or adds the lemma to the index if the correct annotation was missing.

The opportunity for the annotator to add lemmas to the index ensured that the index coverage grew as it was associated with more projects of wider scope. After the annotation of the Cologne corpus, it was found to have a coverage of 90%, with the correct annotation presented as first choice in 60% of the cases. Since the beginning of the annotation efforts predates even standardized tagsets for modern German, customized tagsets were originally used for parts of speech and morphology. They were later mapped to HiTS tags (Dipper et al., 2013).

Annotating a sentence — example Table 1 shows part of the analysis for the beginning of a sentence from the manuscript "Rheinisches Marienlob", a poem in praise of the Virgin Mary: 'Wife Dine Burſte iⁿ dinen lif…' ('Show your breasts [that have suckled Jesus] and your body [that has born Jesus]…').

The first token has four suggestions: The adjective (ADJ) 'wîs(e)' ('wise'), the feminine noun (F) 'wîse' ('meadow'), the weak verb (SwV) 'wîsen' ('to know, to show'), and the adjective (ADJ) 'wîz' ('white'). The system ranked the choices purely according to their naive probabilities — no syntactic context has been encountered yet since this is the beginning of the sentence. This means that the correct analysis, the weak verb, is not ranked very highly in this case, and the annotation has to be corrected. The correct analysis comes with a number of suggestions for the morphology. To generate the suggestions, the inflectional paradigm of this verb was prefiltered according to the inflectional affixes the system recognized. Again, the human annotator has to select

Form	Lemma	POS	Morph
Wiſe	wîs(e)	ADJ	NP/-/0/NSmfnw/NASf/ASnw/NAP
	wîse	F	NS/AS/GFS/NAP
	wîsen	SwV	1SG/3SGK/1PG/2SGB/i
	wîz	ADJ	NSmfnW/NASf/ASnw/NAP
dine	dîn	PronPoss	NP/NSf/ASf/AP
burſte	brust	F(u)	NP/AP/GP/GS/DS
iñ	unde	Konj	–
dinen	dîn	PronPoss	ASm/DP/DSm/DSn
lif	lîb	M	AS/NS/DS
	loufen	stv7	3SVI/1SVI

Table 1: Lemma and annotation suggestions for the beginning of a sentence from "Rheinisches Marienlob". The leftmost column has the form as it was transcribed from the manuscript.

the correct analysis (2SGB, 2nd person imperative). The following tokens are largely unambiguous, only the correct morphological analysis has to be manually selected here. Table 2 shows the corrected annotation for this fragment.

Form	Lemma	POS	Morph
Wiſe	wîsen	SwV	2SGB
dine	dîn	PronPoss	AP
burſte	brust	F(u)	AP
iñ	unde	Konj	–
dinen	dîn	PronPoss	ASm
lif	lîb	M	AS

Table 2: The manually corrected annotation.

The annotator has selected a weak verb (SwV) in 2nd singular imperative form (2SGB) here, followed by a possessive pronoun (PronPoss) in accusative case and plural number (AP), and so on. However, the annotations need to be converted into HiTS-like tags, which have more categories (see Sec. 3) and more distinctions. This is not without its own challenges, as Table 3 below shows.

Mapping to HiTS In some cases, such as for the first token, the mapping from the internal tagset to HiTS is very straightforward. The internal tagset has the SwV POS tag indicating a weak verb, and the 2SGB morphology tag for a second person singular imperative form. This was re-distributed to a `pos (token)` tag for a full verb imperative

Token	Wise	dine	burste	in	dinen	lif
pos (token)	VVIMP	DPOSA	NA	KON	DPOSA	NA
pos (type)	VV	DPOS	NA	KO	DPOS	NA
infl	Sg.2	Fem.Akk.Pl.st	Akk.Pl	–	Masc.Akk.Sg.st	Akk.Sg
inflClass	wk	–	st(u).Fem	–	–	st.Masc

Table 3: Final annotations for this fragment. Lemma and other annotations are omitted here, but are visible in the final corpus. The tokens are shown in simplified spelling.

(VVIMP), a `pos (type)` tag for a full verb, `infl` showing only 2nd person singular form, and an `inflClass` tag showing the weak inflection class. The second token is annotated as a possessive determinative that precedes its noun phrase (DPOSA). This is not explicitly annotated in the internal tagset, but it can be easily inferred by precedence being the default case for determinatives.

Difficulties arise in cases where HiTS makes distinctions that are not made in the internal tagset. For example, the noun 'burſte' ('breast') is annotated as belonging to the strong inflection class in HiTS, but the internal tagset does not capture this information. This had to be solved by a combination of the analysis of the lemma form and list lookup: if the lemma ends in a consonant, the noun has a strong inflection class. Lemmas ending in '-e' have to be looked up for weak or strong inflection classes. Lemmas ending in other vowels are always weakly inflected. Similar lists had to be built for other parts of speech that lacked disctinctions, such as pronouns, articles and numerals, as well as verbs, auxiliar verbs, and modal verbs.

Some distinctions could not be reconstructed by looking at the token alone. One example for this is the annotation of pronominal adverbs that introduce a relative clause (as opposed to interrogative usage) as PAVREL. Reconstructing these distinctions would have required usage of the syntactic context which the tools are not capable of. In that sense, the tagset used here represents a subset of the entire HiTS tagset.

The final output of this annotation process is a flat XML file based on the modernized tokenization only; the historical tokenization has to be inferred using the transcription standards (see Sec. 3). It is converted into CorA-XML format (see Sec. 5) to re-gain flexibility with regards to the tokenization layers.

5 CorA-XML document format

For further processing of the annotated data, we choose to convert it into the CorA-XML document format. This XML-based format was originally developed for the web-based annotation tool CorA[6] (Bollmann et al., 2014), and is specifically designed for the needs of historical documents. CorA is actively used to annotate historical texts for the reference corpora of Early New High German (ReF) and Middle Low German/Low Rhenish (ReN), as well as the Anselm corpus of Early New High German (Dipper and Schultz-Balluff, 2013). Converting ReM to the same format therefore significantly

[6] `https://www.linguistics.rub.de/comphist/resources/cora/`

```
<token>
  <dipl utf="ſo" />
  <anno utf="ſo" ascii="so">
    <pos tag="AVD" />
    <lemma tag="sô" />
  </anno>
</token>
<token>
  <dipl utf="biſtu" />
  <anno utf="biſ" ascii="bis">
    <pos tag="VAFIN" />
    <lemma tag="sîn" />
  </anno>
  <anno utf="tu" ascii="tu">
    <pos tag="PPER" />
    <lemma tag="dû" />
  </anno>
</token>
```

Figure 1: Simplified CorA-XML representation of "ſo biſtu" with annotations

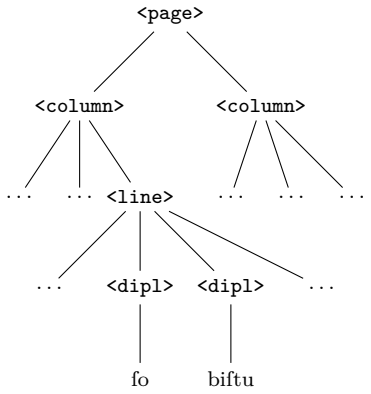


Figure 2: Simplified example of the layout hierarchy in CorA-XML

increases reusability of tools and facilitates further processing of the data. Furthermore, we are actively working on an automatic conversion from CorA-XML to a TEI-compatible format, which will open up the data for use with an even wider range of existing tools.

CorA-XML distinguishes between two different tokenization layers, whose elements are represented by `<dipl>` and `<anno>` tags respectively, corresponding to the distinction between diplomatic and annotated tokens in ReM (cf. Sec. 3). Since there can be a one-to-many (or even many-to-many) relationship between elements of these layers (as in the 'biſtu' example from Fig. 4 below), they are always wrapped by a virtual `<token>` element which establishes this correspondence. Within each layer, different representations of the wordforms can be included, e.g., a UTF-8 representation conserving special characters (such as 'ſ'), or a pure ASCII representation (mapping 'ſ' to 's'). On the annotated tokenization layer, arbitrary annotations can be added to each token, encoding the linguistic layer and punctuation layer described in Sec. 3. Figure 1 gives a simplified example of the CorA-XML representation for the sequence 'ſo biſtu' from Figure 4.

Layout information is encoded via a hierarchy of layout elements, namely 'pages', 'columns', and 'lines'. Each instance of an element contains a pointer to one or more elements of the next lower type in the hierarchy; i.e., pages refer to columns, which in turn refer to lines. Each 'line' element finally refers to one or more diplomatic tokens. Figure 2 provides an example visualization of this hierarchy. A valid layout specification in a CorA-XML document requires that each diplomatic token is contained in the span of exactly one 'line' element, thereby allowing to derive an exact page, column, and line specification for each diplomatic token.

6 Access via ANNIS

For the public release of the corpus, it was important that different user groups' needs can be satisified by a single visualization and search system. Users should be able to make (diachronically oriented) queries that disregard variation such as different use of diacritics, usage of long or normal 's', or tokenization peculiarities. At the same time, the transcription captures all such variation, so it was important to make them available as well for users that want to research those aspects of our texts. The corpus tool ANNIS[7] (Krause and Zeldes, 2016) addresses needs such as ours, by specifically targeting the visualization of complex, multi-layer corpora. It also offers Pepper[8], a modular conversion infrastructure that can be leveraged to convert a number of different formats into ANNIS native format for easy import. Since it did not originally recognize Cora-XML, we developed an import module for it which is now included in the Pepper distribution.

In spite of its flexibility, there are a number of technical and conceptual limitations. For technical reasons, there is a limit on the size of a corpus that can be imported into ANNIS. The exact limit depends on the number and nature of the annotations, in our case it amounts to around 60,000 tokens. We solved this by dividing the texts into smaller subcorpora. Since no single criterion provided a subdivision of appropriate size for all of their values, we used a combination of several criteria. The first subdivision is by the century, or half-century where the texts most likely originated, such as 11-1 for the first half of the 11th century (1000–1050). All centuries are further divided into more or less broad dialect areas, such as alem for Alemannic. Most dialects are attested well enough to warrant further subdivision into prose (P), verse (V), and charter (U – "Urkunde") texts. Finally, a suffix marks if the texts are from the original, balanced grammar corpus (G) or from the extension (X). In this way, the subcorpus list also allows for a quick access to some of the meta annotations. The texts are further annotated with more exact and specific meta annotations that are also searchable (Fig. 3).

Displaying annotations in ANNIS For the display of annotations, we chose the grid view, which is essentially a table with flexible column sizes. It fits the structure of our annotations, which are of two distinct categories. Linguistic annotations, such as parts of speech or lemma, relate to word tokens in modernized tokenization. Layout related information, such as page or line breaks, which is also treated as annotation by ANNIS on the other hand, relates to historical tokenization (see Sec. 3). Users have to be able to query for layout specific information in their searches, yet displaying all layout information in the grid would visually clutter the results. We therefore combined all layout information on the line level, while the specific higher levels are still searchable, but will not be displayed in the results. The names for the annotation categories were

[7] http://corpus-tools.org/annis/
[8] http://corpus-tools.org/pepper/

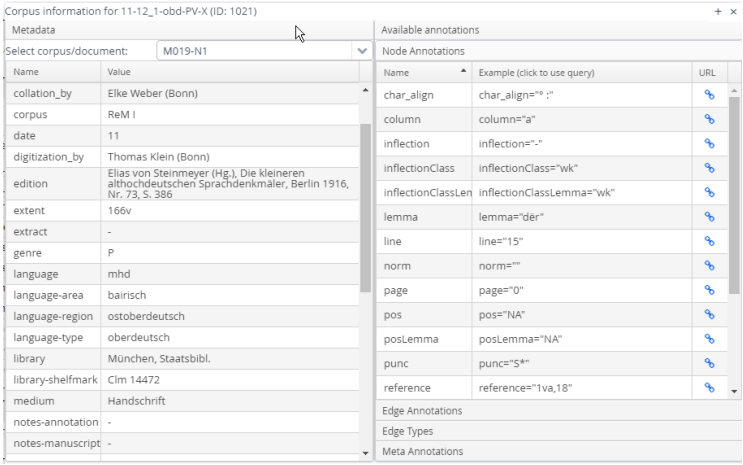


Figure 3: Part of the meta annotations for the text "Augensegen" ("blessing of the eyes"). Some of the meta annotations are important for diachronic searches, others (such as the annotators responsible for digitization) are merely informative.

chosen for consistency with other existent reference corpus projects where possible (see Sec. 1).

On the conceptual level, ANNIS default configurations assume a single, main token layer. However, in our case the simple surface token form already exists in two annotation dimensions: transcription (diplomatic or simplified), and tokenization (historical or modern). Displaying each possible combination would clutter the results more than it would help, so we chose only two token forms for the primary text: `tok_anno` and `tok_dipl`. `tok_anno` combines the modern tokenization with simplified spelling, while `tok_dipl` combines historical tokenization with diplomatic spelling. These two token variations make up the primary text and can be selected to be displayed in the KWIC view of the primary search results. Fig. 4 shows such a result for the search for the sequence "bis tu" in modernized form.

Each search result is shown in KWIC format with the currently selected tokenization layer, the main layer can be switched between `tok_dipl` and `tok_anno` via the menu on the top.[9] Below the main token is an expandable grid table displaying the annotations. It starts on top with layout information ("66a,2b"). Layers `tok_dipl` and `tok_anno` contain the two textual versions, followed by the layers with linguistic information. The layer `norm` contains the normalized form that closely corresponds to word forms as used in Middle High German dictionaries (see Sec. 3). Layer `tokenization` contains the information on the difference between modernized and historical tokenization. Layers

[9] The menu also shows the default token layer, which is empty, as it was only used to align the two tokenization layers.

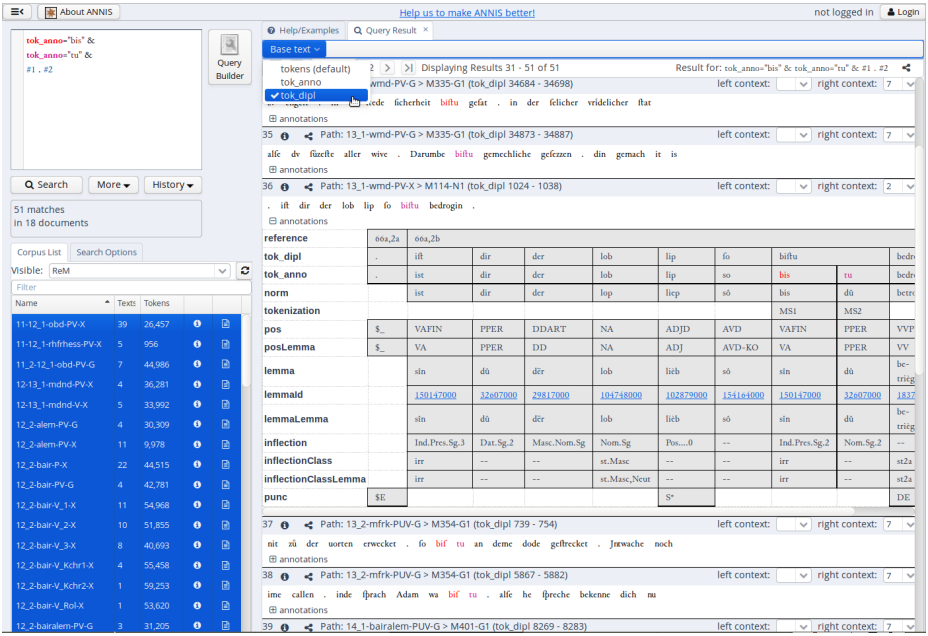


Figure 4: ANNIS window showing the results of a search for the sequence "bis tu" in modernized form. Part of the subcorpus list is shown on the lower left.

`pos` and `posLemma` correspond to the part of speech of the token and type respectively (see Sec. 3), as do the layers `inflectionClass` and `inflectionClassLemma`. Layer `punc` at the bottom encodes information on punctuation marks and segmentation.

Full text view The different user groups' needs are also taken into consideration for the full text view. While ANNIS has a default full text view, it does not work with our corpora, since it presumes a single main token layer. Instead, we used a functionality that allows a full text view to be generated as an HTML document by emitting any annotation as HTML elements, which can then be styled with CSS, thus making it adaptable for both diplomatic and modernized views.

A diplomatic view provides a version of the document that is as close to the original manuscript as possible. It displays all letter variation, diacritics, layout, and tokenization unchanged, and can be used as a more readable version of the original for many purposes. The layout levels are emitted as nested `div` elements, with the final line `divs` containing the `tok_dipl` as spans. Fig. 5 shows part of the diplomatic view for a text.

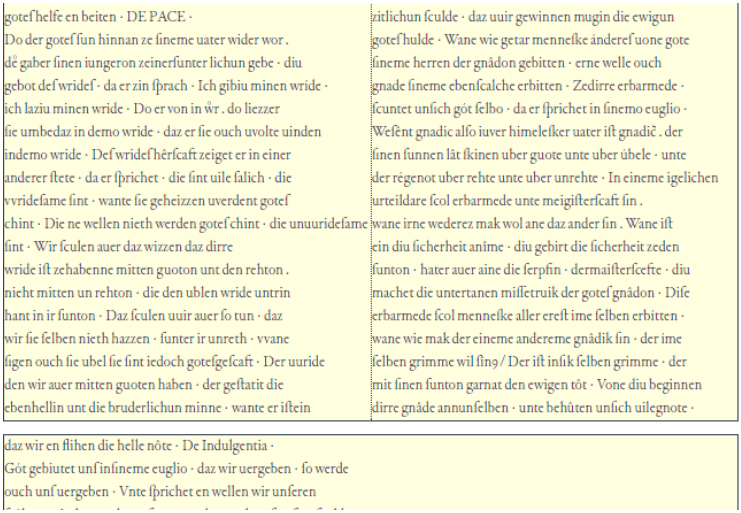


Figure 5: Diplomatic full text view of the Middle High German translation of Alkuin's "De virtutibus et vitiis"

The layout elements are then placed via CSS in a way that resembles the manuscript: the larger box represents a folio page, with the left and right side representing the back and front sides of the manuscript page. If the manuscript has multiple columns, they are placed next to each other. The text is rendered in a Unicode version that mirrors the original. The yellow tint provides a visual clue that the text presentation is oriented towards the original.

The modernized view is based on the simplified transcription and modern tokenization. It provides a quick way of accessing larger contexts, and, since it does not imitate the original layout, the opportunity to fit the text to varying screen sizes. Fig. 6 shows part of the modernized view of the same text.

Since the corpus in its current form only annotates boundary locations (see Sec. 3), and not the entire sentence spans, there is no structuring information that can be used by ANNIS' full text view. As the absence of any structuring would hinder readability, especially for longer texts, we used the pages and columns from the `dipl` structure to emit paragraph (`p`) elements which contain all `tok_anno` as spans. Unfortunately, this leads to paragraphs sometimes breaking up sentences, since they orient towards the layout. However, since the modernized view consists only of variable size elements, it can be easily adapted to different screen sizes and browser window sizes, as can be seen from the downscaled browser window.

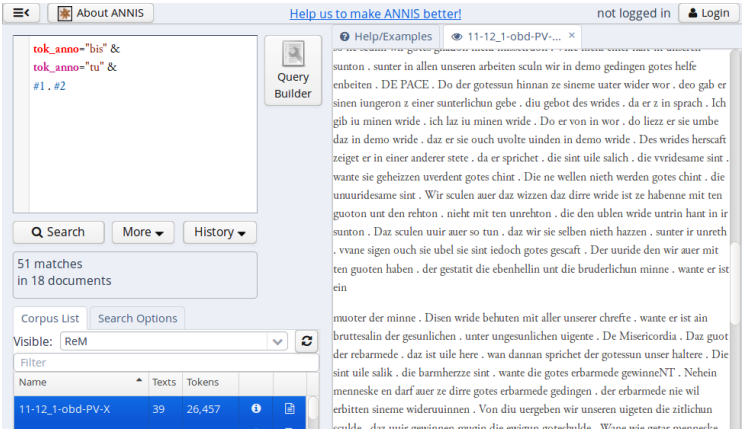


Figure 6: Modernized full text view of the document displayed in Fig. 5.

7 Conclusion

We presented the creation of the Reference Corpus Middle High German (ReM) with a focus on the compilation and annotation process and its implications for the preparation and release of the corpus.

The ReM corpus is a product of several annotation efforts stretching over the span of about 30 years, and starting as far back as 1986 (cf. Sec. 2). This explains the usage of annotation tools, formats, and tagsets that would be considered "out-dated" from a modern point of view. We discussed the types of annotation in the final corpus and how they were derived from the originally annotated data; e.g., creating two distinct tokenization layers ("diplomatic" and "annotated"/"modernized") from word boundary markings in the transcription, or mapping the custom part-of-speech tagset to the modern HiTS tagset (cf. Secs. 3 and 4).

By converting the corpus into an XML format (Sec. 5), we hope to make it more accessible for existing tools and computational analyses. Providing access to the corpus via the ANNIS tool (Sec. 6), on the other hand, provides an efficient way for querying and visualizing the corpus data.

Acknowledgments

We would like to thank the German Research Foundation (Deutsche Forschungsgemeinschaft) for financial support, Grants DI 1558/1, KL 472/6, WE 1318/14, WI 3664/2.

References

Bollmann, M., Petran, F., Dipper, S., and Krasselt, J. (2014). CorA: a web-based annotation tool for historical and other non-standard language data. In *Proceedings of the 8th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH)*, pages 86–90, Gothenburg, Sweden.

Dipper, S., Donhauser, K., Klein, T., Linde, S., Müller, S., and Wegera, K.-P. (2013). HiTS: ein Tagset für historische Sprachstufen des Deutschen. *Journal for Language Technology and Computational Linguistics, Special Issue*, 28(1):85–137.

Dipper, S. and Schultz-Balluff, S. (2013). The Anselm corpus: Methods and perspectives of a parallel aligned corpus. In *Proceedings of the NODALIDA Workshop on Computational Historical Linguistics*.

Klein, T. (1991). Zur Frage der Korpusbildung und zur computergestützten grammatischen Auswertung mittelhochdeutscher Quellen. In Wegera, K.-P., editor, *Mittelhochdeutsche Grammatik als Aufgabe*, pages 3–23. E. Schmidt, Berlin.

Klein, T. (2001). Vom lemmatisierten Index zur Grammatik. In Moser, S., Stahl, P., Wegstein, W., and Wolf, N. R., editors, *Maschinelle Verarbeitung altdeutscher Texte V. Beiträge zum Fünften Internationalen Symposion, Würzburg 4.-6. März 1997*, pages 83–103. de Gruyter, Berlin.

Klein, T. and Bumke, J. (1997). *Wortindex zu hessisch-thüringischen Epen um 1200*. Niemeyer, Tübingen. Unter Mitarbeit von B. Kronsfoth und A. Mielke-Vandenhouten.

Klein, T. and Dipper, S. (2016). Handbuch zum Referenzkorpus Mittelhochdeutsch. *Bochumer Linguistische Arbeitsberichte*, 19.

Klein, T., Solms, H.-J., and Wegera, K.-P., editors (2009). *Mittelhochdeutsche Grammatik. Teil III: Wortbildung*. Niemeyer, Tübingen.

Krause, T. and Zeldes, A. (2016). ANNIS3: A new architecture for generic corpus query and visualization. *Digital Scholarship in the Humanities*, 31:118–139. `http://dsh.oxfordjournals.org/content/31/1/118`.

Wegera, K.-P. (2000). Grundlagenprobleme einer neuen mittelhochdeutschen Grammatik. In Besch, W., Betten, A., Reichmann, O., and Sonderegger, S., editors, *Sprachgeschichte. Ein Handbuch zur Geschichte der deutschen Sprache und ihrer Erforschung*, volume 2, pages 1304–1320. de Gruyter, Berlin, New York, 2nd edition.

Roland Mittmann

Automatisierter Abgleich des Lautstandes althochdeutscher Wortformen

Abstract

Um Texte einer Sprache automatisiert auf ihren möglichen Entstehungszeitraum und ihre dialektale Zugehörigkeit hin zu untersuchen, werden für jedes erwartete Graphem und jede Flexionsendung zunächst Entsprechungsregeln zwischen einer idealisierten Sprachform und den Sprachformen in Grammatiken beschriebener Zeit-Dialekt-Räume erfasst. Anschließend werden mithilfe eines Computerprogramms unter Anwendung dieser Regeln die belegten Wortformen mit ihren Entsprechungen in der idealisierten Sprachform abgeglichen und für jeden Text die Übereinstimmungsgrade mit den einzelnen Zeit-Dialekt-Räumen angegeben. Exemplarisch wird dieser Abgleich für eine althochdeutsche Wortform beschrieben und das Ergebnis der Analyse des zugehörigen Gesamttextes dargestellt.

1 Untersuchungsthema

Seit jeher verändern sich Sprachen im Laufe der zeitlichen Entwicklung. Sobald ihre Sprechergemeinschaften in verschiedene Gruppen zerfallen, die nicht mehr dauerhaft miteinander in Kontakt stehen, entwickeln sie zudem verschiedene Varietäten. Solange die Normierung einer Sprache nicht erfolgt ist, bleibt die textliche Überlieferung daher sprachlich uneinheitlich.

Auch innerhalb eines Textes können Schwankungen auftreten, etwa wenn Sprecher verschiedener Dialekte am selben Text arbeiten oder einen bestehenden Text korrigieren (vgl. etwa Braune/Reiffenstein 2004, § 3 und Anm. 1). Ein einzelner Autor kann ebenfalls verschiedenen dialektalen Einflüssen unterworfen sein oder die im Laufe seines Lebens erfolgte sprachliche Veränderung in seinen Niederschriften wiedergeben. Da vor der Erfindung des Buchdrucks Texte allein durch Abschrift vervielfältigt wurden, kam es schließlich auch seitens der Kopisten – bewusst oder unbewusst – zu sprachlichen Anpassungen bei dialektalen Formen bzw. infolge der zeitlichen Entwicklung.

Sind zu einem Teil der textlichen Überlieferung einer Sprache keine genaueren zeitlichen und örtlichen Angaben bekannt, erscheint es denkbar, diese automatisiert auf ihre Übereinstimmung mit verschiedenen Zeit-Dialekt-Räumen – also Zeitabschnitten mit Bezug auf die verschiedenen örtlichen Varietäten – zu untersuchen. Diese Untersuchung wird im Folgenden beschrieben. Voraussetzung dafür ist, dass Angaben zu den üblichen Entsprechungen der verschiedenen Phonem-Graph-Entsprechungen (Lautverschriftungen, vgl. Mittmann 2015b, 248) und der Flexionsendungen in den einzelnen Zeit-Dialekt-Räumen vorliegen.

2 Untersuchungsobjekt und Datengrundlage

Als Untersuchungssprache wird das Althochdeutsche gewählt, das ab etwa 750 überliefert ist und um 1050 ins Mittelhochdeutsche übergeht (vgl. PAUL ET AL. 2007, § E 7). Das althochdeutsche Textkorpus umfasst etwa 560.000 Wortformen und erscheint damit für die Untersuchung hinreichend groß. Ein Teil des Korpus sind umfangreiche Texte mit einheitlicher Sprachform und bekannter Überlieferungsgeschichte (etwa die Werke Notkers des Deutschen mit ca. 320.000 Wortformen)[1] Daneben umfasst es aber auch zahlreiche mittelgroße und kleinere Texte, deren Herkunft oft nicht genau bekannt ist, sodass sich die Angaben in den Grammatiken zu zeitlichen und dialektalen Unterschieden beim Laut- und Formenstand nicht auf sie stützen können. Ein Zirkelschluss ist für diese Texte also ausgeschlossen, und die erstgenannten können zur Überprüfung der Untersuchungsmethode dienen.

Das Althochdeutsche weist eine deutlich erkennbare zeitliche Entwicklung auf und lässt sich in fünf gut bezeugte Dialektgebiete (Alemannisch, Bairisch, Ostfränkisch, Rheinfränkisch und Mittelfränkisch, vgl. BRAUNE/REIFFENSTEIN 2004, §§ 4-6) gliedern. Zur Abgrenzung werden zudem drei westgermanische Nachbarsprachen (Altsächsisch, Altniederfränkisch und Langobardisch, vgl. BRAUNE/REIFFENSTEIN 2004, § 2 u. Anm. 1) bei der Untersuchung hinzugenommen.[2]

Die für die Untersuchung verwendeten Daten sind auf Grundlage des DFG-geförderten Projektes *Referenzkorpus Altdeutsch* erstellt worden. Das entstandene Korpus umfasst alle althochdeutschen und altsächsischen Texte und weist eine umfangreiche morphologische Annotation auf (vgl. *www.deutschdiachrondigital.de*, LINDE/MITTMANN 2013 sowie MITTMANN 2013).

3 Vorbereitung der Texte

Beleg	Kestírnis
Lemma	Gistirni
Übersetzung	Gestirn
Wortart Lemma	NA
Wortart Beleg	NA
Flexion Lemma	ja_Neut
Flexion Beleg 1	ja_Neut
Flexion Beleg 2	Sg_Gen

Abbildung 1: Korpusauszug aus der *St. Galler Schularbeit* (vereinfacht)

Ausgehend von den im Korpus angegebenen Lemmata und morphologischen Angaben werden mithilfe eines Computerprogramms zunächst idealisierte Wortformen gebildet, die den für die Untersuchung der zeitlichen und dialektalen Zuordnung benötigten Abgleich mit den belegten Wortformen ermöglichen (vgl. MITTMANN 2015a, 68-74). Dazu werden je nach morphologischen Werten Endungen angefügt oder ersetzt; bei Umlaut oder starken Verben auch Vokalersetzungen im Inlaut durchgeführt. Sowohl das für die Lemmatisierung der althochdeutschen Texte verwendete Wörterbuch (SPLETT 1993) als auch die zur Ermittlung der Flexionsformen verwendete Referenzgrammatik (BRAUNE/REIFFENSTEIN 2004) orientieren sich am ostfränkischen Dialekt um 830, sodass auch die idealisierten Wortformen den größten Teil der im ältesten Althochdeutschen noch bewahrten lautlichen Unterschiede wiedergeben (vgl. MITTMANN 2015b, 249).

Abbildung 1 zeigt eine Wortform aus dem *Referenzkorpus* mitsamt einigen Annotationszeilen. Um die idealisierte Wortform zu erzeugen, wird die Ersetzungsregel für die Angaben *„N[3], ja,*

{Masc|Neut}, Sg, Gen" (Genitiv Singular eines maskulinen oder neutralen *ja*-stämmigen Substantivs) – also die grau hinterlegten Informationen – auf das Lemma angewendet. Das auslautende *-i* wird somit durch *-ies* ersetzt, sodass eine idealisierte Wortform *gistirnies* entsteht, die dann mit dem belegten *kestírnis* verglichen werden kann.[4]

4 Vorbereitung der Untersuchung

Vor der Durchführung der Untersuchung müssen noch Entsprechungsregeln zwischen den idealisierten und den belegten Wortformen mit Bezug auf die jeweiligen Zeit-Dialekt-Räume erstellt werden. Dafür wird zunächst in Anlehnung an die Angaben bei BRAUNE/REIFFENSTEIN (2004) die Zeit von 750 bis 1150 in acht 50-Jahres-Abschnitte unterteilt, damit auch sprachliche Entwicklungen zum Mittelhochdeutschen (ca. 1050–1350) hin erfasst sind. Auf diese Weise ergeben sich in Kombination mit den zusammen acht Dialekten und Nachbarsprachen 64 Zeit-Dialekt-Räume, etwa *„Alemannisch, 900–950".*

Anschließend werden auf Grundlage von Referenzgrammatiken für die althochdeutschen (BRAUNE/REIFFENSTEIN 2004)[5] und frühmittelhochdeutschen Dialekte (PAUL ET AL. 2007) sowie die Nachbarsprachen – GALLÉE/TIEFENBACH (1993) fürs Altsächsische, BREMMER/QUAK (1992) fürs Altniederfränkische/Altniederländische und BRUCKNER (1895) fürs Langobardische – Entsprechungsregeln erstellt. Die einzelnen Phonem-Graph-Entsprechungen werden dabei nach ihrer Stellung im Wort unterschieden, um der jeweils unterschiedlichen Entwicklung Rechnung zu tragen: Konsonanten nach Anlaut, Inlaut und Auslaut; Vokale (und Diphthonge) nach Präfix, Tonsilbe, Mittelsilbe und Endsilbe.[6] Darüber hinaus ist in zahlreichen Fällen auch die Umgebung der jeweiligen Phoneme entscheidend, wie Abbildung 2 zeigt:

g	**> {g|k|c} / #_i**	für "Alemannisch oder Bairisch, 750–1050"
g	**> {g|gh|j|Ø} / #_i**	für *"Altsächsisch"* (ohne zeitliche Eingrenzung)
i	**> e / _rC** (Tonsilbe)	für "Altniederfränkisch oder Altsächsisch"

Abbildung 2: Beispiele für Entsprechungsregeln für spezifische Zeit-Dialekt-Räume

Neben den Phonem-Graph-Entsprechungen werden für die flektierenden Wortarten auch Flexionsendungen mit gleicher Funktion, aber über den unterschiedlichen Lautstand hinaus auch dialektal unterschiedlicher morphologischer Bildeweise, betrachtet. So lässt sich etwa beim Nominativ Plural der *a*-stämmigen maskulinen Substantive die altalemannische Endung *-ā* auf gemeingermanisches rekonstruiertes **-ōz* und schließlich auf urindogermanisches **-oes* zurückführen, die altsächsische Endung *-os* dagegen auf gemeingermanisches **-ōsiz* und schließlich auf urindogermanisches **-óeses* (vgl. KROGH 1996, 295 i. V. m. BAMMESBERGER 1990, 43 f.).[7] Bei der althochdeutschen Endungsvariante *-a* wiederum liegt wohl eine Übernahme der gleichlautenden Akkusativendung vor (vgl. KROGH ebd.).

Auf diese Weise ergeben sich 707 Entsprechungsregeln: 203 für Flexionsendungen (plus 21 Regeln für Stammallomorphie) sowie 483 für Phonem-Graph-Entsprechungen (166 für Konsonanten und 317 für Vokale).

5 Durchführung der Untersuchung

Für jeden Text wird eine Wort-für-Wort-Prüfung durchgeführt, indem die Entsprechungsregeln auf dessen einzelne Phonem-Graph-Entsprechungen und Flexionsendungen angewendet werden. Dabei

wird berücksichtigt, dass das Korpus nur wortweise aligniert ist und oft keine 1:1-Zuordnungen der Phonem-Graph-Entsprechungen zueinander vorliegen. Im Althochdeutschen gilt, von Ausnahmen abgesehen, Anfangsbetonung, und aufgrund der schon in ältester Zeit (vgl. BRAUNE/REIFFENSTEIN 2004, § 54) beginnenden Vokalreduktion in den unbetonten Silben – bis hin zum Schwund – sowie des Auftretens epenthetischer Vokale und Konsonanten (vgl. ebd., § 69) ist vor allem in den Mittelsilben mit mangelnden Entsprechungen zu rechnen. Um möglichst viele automatisierte Lautzuordnungen zu ermöglichen, erfolgt die Prüfung daher von beiden Wortenden aus zur Mitte hin. Da Komposita und Präfixe im Korpus nicht markiert sind, wird jedoch auch in Mittel- und Endsilben geprüft, ob ein betonter Vokal vorliegt. Zudem wird die Möglichkeit von Doppel- und Einfachschreibung stets mitberücksichtigt.

Sofern das Wort flektiert, wird zunächst die Endung ermittelt und mit den in den einzelnen Zeit-Dialekt-Räumen bezeugten Formen abgeglichen. Unterschiede zwischen verschiedenen Endungsformen mit dem gleichen zugrundeliegenden Lautstand werden an dieser Stelle nicht berücksichtigt. Unabhängig davon erfolgt dann der Abgleich eines eventuellen auslautenden (sowie eventueller diesem vorangehenden inlautender) Konsonanten, danach in mehrsilbigen Wörtern des Endsilbenvokals und des diesem vorangehenden inlautenden Konsonanten. Ist die Endung hiervon noch nicht vollständig abgedeckt, werden noch weitere Mittelsilbenvokale und inlautende Konsonanten abgeglichen. Anschließend wird der Abgleich am Wortanfang mit einem eventuellen anlautenden (sowie eventuellen diesem folgenden inlautenden) Konsonanten und dem Tonsilbenvokal fortgesetzt, gefolgt von Mittelsilbenvokalen und inlautenden Konsonanten. Falls der Lautstand den Beginn mit einem unbetonten Präfix zulassen könnte, wird diese Möglichkeit für den ersten Vokal mit einbezogen.

Findet sich keine passende Entsprechungsregel, bricht das Programm die Untersuchung der Wortform vom Ende zur Mitte hin bzw. vom Anfang zur Mitte hin ab. Erfolgt der Abbruch unmittelbar im Auslaut oder Anlaut, wird die Prüfung also ausschließlich in die jeweils andere Richtung durchgeführt. Passen sowohl die Endung als auch Auslaut und Anlaut nicht, wird die Wortform vollständig übersprungen, da in diesem Fall von einer Fehlzuweisung auszugehen ist, die bei einem Korpus dieser Größe nicht ausgeschlossen werden kann.

Um den Übereinstimmungsgrad eines Textes mit den einzelnen Zeit-Dialekt-Räumen zu ermitteln, wird dieser mit jeder Regelanwendung für alle 64 Zeit-Dialekt-Räume – jeweils beginnend bei 0 – erhöht oder gesenkt. Da eine nur in wenigen Zeit-Dialekt-Räumen verbreitete Lautform für die Zuordnung eines Textes als deutlich signifikanter gelten kann als eine, die in fast allen Zeit-Dialekt-Räumen vorkommt, wird jedes Mal die Gesamtzahl der Zeit-Dialekt-Räume (64) durch die Zahl der (nicht) zutreffenden Zeit-Dialekt-Räume geteilt und 1 davon abgezo gen: So ergibt sich etwa eine Veränderung von ± 0, wenn alle Zeit-Dialekt-Räume zutreffen (64/64 - 1), und eine Veränderung von + $4{,}\overline{3}$ für die in Abbildung 2 zuoberst genannte Regel, die auf zwölf Zeit-Dialekt-Räume zutrifft (64/12 - 1). Der ermittelte absolute Übereinstimmungsgrad jedes einzelnen Zeit-Dialekt-Raums wird schließlich prozentual auf den Übereinstimmungsgrad eines fiktiven („idealen“) Zeit-Dialekt-Raums bezogen, der dem jeweiligen Text exakt entspricht, sodass sich ein relativer Übereinstimmungsgrad ergibt.

Graphem(e)	Idealisierte Wortform	Belegte Wortform	Angewandte Regel	für „Alemannisch, 900–950“ Zutreffend?	Gleiche Fälle	Übereinstimmungsgrad Veränderung	Übereinstimmungsgrad Kumulierte Summe
Endung	gistirnies	kestirnis	**es → is** {Sg, Gen}	-	41	- 0,56	- 0,56
C final	gistirnies	kestirnis	s = s	+	64	± 0,00	- 0,56
V Endsilbe	gistirnies	kestirnis	e > i	-	49	- 0,31	- 1,78
prävokalisches *j*	gistirnies	kestirnØis	**j > Ø** / C_V	+	64	± 0,00	- 0,87
C vor V	gistirnies	kestirnis	**n = n** / C_	+	64	± 0,00	- 0,87
C initial	gistirnies	kestirnis	**g > k** / _i	+	12	+ 4,33	+ 3,46
Präfix-V	gistirnies	kestirnis	**i > e** / g_-	-	33	- 0,94	+ 2,52
C medial	gistirnies	kestirnis	s = s	+	64	± 0,00	+ 2,52
C medial	gistirnies	kestirnis	**t = t** / s_	+	64	± 0,00	+ 2,52
V Tonsilbe	gistirnies	kestirnis	**i = i** / _rC	+	48	+ 0,33	+ 2,85
C medial	gistirnies	kestirnis	r = r / _C	+	64	± 0,00	+ 2,85

Abbildung 3: Ermittlung des absoluten Übereinstimmungsgrades einer Wortform mit einem Zeit-Dialekt-Raum

6 Beispielfall (Einzelwort)

Abbildung 3 zeigt das Vorgehen des Computerprogramms für die bereits in Abbildung 1 dargestellte Wortform *kestírnis* aus der *St. Galler Schularbeit*. Die rechten vier Tabellenspalten enthalten dabei die Veränderungen der absoluten Übereinstimmungsgrade für den Zeit-Dialekt-Raum *„Alemannisch, 900–950"*. Die Spalte „Zutreffend?" gibt an, ob nach den Angaben bei BRAUNE/REIFFENSTEIN (2004) die Entsprechungsregel, die das Programm durch Abgleich von idealisierter und belegter Wortform als anzuwenden erkennt, auf diesen Zeit-Dialekt-Raum zutrifft oder nicht. Die Spalte „Gleiche Fälle" nennt die Anzahl an Zeit-Dialekt-Räumen, die sich ebenso verhalten wie *„Alemannisch, 900–950"*, sodass anhand dieses Wertes die Veränderung des Übereinstimmungsgrades ermittelt werden kann.

Bestünde der Text nur aus diesem einen Wort, würde abschließend das Ergebnis von + 2,63 in Bezug zu dem Ergebnis für den „idealen" Zeit-Dialekt-Raum gesetzt, das für dieses Wort + 10,77 (= 100 %) beträgt. Der relative Übereinstimmungsgrad des Wortes mit dem Zeit-Dialekt-Raum *„Alemannisch, 900–950"* beläuft sich nach dieser Methode also auf etwa 26,5 %.

7 Auswertung eines vollständigen Textes

Nach Abgleich aller 104 Wörter der *St. Galler Schularbeit* (Text bei STEINMEYER 1916, 121) mit den einzelnen Zeit-Dialekt-Räumen zeigt Abbildung 4 die Auswertung für den Gesamttext:[8]

	vor 800	800 – 850	850 – 900	900 – 950	950 – 1000	1000 – 1050	1050 – 1100	nach 1100
Langobardisch	-65	-65	-65	-65	-65[9]	-65	-65	-65
Alemannisch	-26	-12	+17	+22	+53	+78	+47	+47
Bairisch	-25	-21	+14	+14	+16	+41	+9	+9
Ostfränkisch	-51	-24	+10	+10	+12	+39	+10	+10
Rheinfränkisch	-52	-28	+6	+6	+8	+35	+6	+8
Mittelfränkisch	-60	-55	-19	+3	+3	+29	±0	+2
Niederfränkisch	-58	-47	-45	-46	-46	-45	-35	-13
Altsächsisch	-35	-35	-35	-35	-35	-35	-35	-35

Abbildung 4: Relative Übereinstimmungsgrade der *St. Galler Schularbeit* mit den Zeit-Dialekt-Räumen (gerundet, %)

Die mit Abstand höchste Übereinstimmung, ca. 78 %, wird also für *„Alemannisch, 1000–1050"* ermittelt, gestützt durch die nächstbesten Werte in den benachbarten Zeitabschnitten sowie Dialekten des Althochdeutschen. Die Annahme SONDEREGGERS (1980, Sp. 1049), die *St. Galler Schularbeit* stamme aus der ersten Hälfte des 11. Jahrhunderts, wird somit bestätigt.

8 Ausblick

Erst nach Analyse sämtlicher althochdeutscher Texte wird eine Aussage dazu möglich sein, inwiefern die beschriebene Untersuchungsmethode tatsächlich dazu geeignet ist, neue Erkenntnisse zu Entstehungszeiten und Schreibdialekten zu liefern. Dafür muss zunächst aber die Analyse der umfangreicheren Texte zeigen, ob die Angaben aus den Grammatiken in Summe einen plausiblen

Regelapparat bilden: Das ist der Fall, wenn der Forschungsstand zur zeitlich-dialektalen Zuordnung dieser Texte einerseits den Zeit-Dialekt-Räumen mit den bei Analyse jeweils höchsten Übereinstimmungsgraden andererseits entspricht. Dann ließe sich die Untersuchung auf die kleineren Texte mit weniger bekannter Überlieferungsgeschichte übertragen und sich so neue Erkenntnisse zu ihrem zeitlich-dialektalen Sprachstand gewinnen. Dies habe ich im Rahmen meiner Dissertation unternommen.

[1] Vgl. etwa BRAUNE/REIFFENSTEIN (2004, XII) zu den für die Erstellung der Grammatik verwendeten „hauptquellen" und den „nur soweit […] als nötig" zugezogenen Quellen.

[2] Zwar sind aufgrund der Überlieferungsform des Langobardischen die meisten Flexionsendungen dort nicht überliefert (vgl. BRUCKNER 1895, § 98), diese Fälle werden in der Untersuchung jedoch übersprungen und nicht als unzutreffende Übereinstimmung gewertet.

[3] Der zweite Teil der Angabe „NA" steht für „Appellativum" und ist für die Flexion irrelevant.

[4] Bei den vokalischen Substantivklassen wird stets die in BRAUNE/REIFFENSTEIN (2004) aufgeführte Form mit erhaltenem /j/ (dort als ‹i› geschrieben) verwendet, um den Prozess des Schwindens von /j/ nach Konsonant und den damit verbundenen Zusammenfall von Flexionsklassen (vgl. etwa ebd., § 210) deutlicher verfolgen zu können.

[5] Da für Altalemannische bis heute eine eigene Dialektgrammatik fehlt (vgl. BRAUNE/REIFFENSTEIN 2004, XII f.), kann als Referenzgrammatik für das Althochdeutsche nur eine Grammatik dienen, die alle Varietäten der Sprachstufe beschreibt.

[6] Ein Ausschnitt aus der Zuordnungstabelle für die konsonantischen Anlaute findet sich bei MITTMANN (2015b, 253).

[7] Das Doppel-Makron zeigt jeweils Überlänge des Vokals an.

[8] Nicht negative Werte sind von 0 ausgehend mit sich in 10-Prozentpunkt-Intervallen verdunkelnden Graustufen hinterlegt.

[9] Die letzte Änderung des langobardischen Lautstandes ist bei BRUCKNER (1895) für „[i]m Laufe des 9. Jhs." (§ 48) angegeben und darf somit für um 900 abgeschlossen angenommen werden. Die Angaben zum Langobardischen nach 950 erscheinen daher nur aus systematischen Gründen.

Literatur

BAMMESBERGER, A. (1990). Die Morphologie des urgermanischen Nomens (= Untersuchungen zur vergleichenden Grammatik der germanischen Sprachen, Band 2). Heidelberg: Winter.

BRAUNE, W. / REIFFENSTEIN, I.-(2004). Althochdeutsche Grammatik. Band I: Laut- und Formenlehre (= Sammlung kurzer Grammatiken germanischer Dialekte, A. Hauptreihe Nr. 5). 15. Auflage, bearbeitet von I. Reiffenstein. Tübingen: Niemeyer.

BREMMER, R. H. JR. / QUAK, A. (Hg.) (1992). Zur Phonologie und Morphologie des Altniederländischen (= North-Western European Language Evolution (NOWELE), Supplement Volume 6). Odense: Odense University Press.

BRUCKNER, W. (1895). Die Sprache der Langobarden (= Quellen und Forschungen zur Sprach- und Culturgeschichte der germanischen Völker, Bd. LXXV). Straßburg: Trübner.

GALLÉE, J. H. / TIEFENBACH, H. (1993). Altsächsische Grammatik (= Sammlung kurzer Grammatiken germanischer Dialekte, A. Hauptreihe Nr. 6). 3. Auflage mit Berichtigungen und Literaturnachträgen von H. Tiefenbach. Tübingen: Niemeyer.

KROGH, S. (1996). Die Stellung des Altsächsischen im Rahmen der germanischen Sprachen (= Studien zum Althochdeutschen, Band 29). Göttingen: Vandenhoeck & Ruprecht.

LINDE, S. / MITTMANN, R. (2013). „Old German Reference Corpus. Digitizing the knowledge of the 19th century. Automated pre-annotation using digitized historical glossaries." In: Bennett, P. et al. (Hg.) (2013). New Methods in Historical Corpora (= Korpuslinguistik und interdisziplinäre Perspektiven auf Sprache / Corpus Linguistics and Interdisciplinary Perspectives on Language – CLIP, Band 3). Tübingen: Narr, 235-246.

MITTMANN, R. (2013). „Old German and Old Lithuanian: the Creation of Two Deeply-Annotated Historical Text Corpora." In: Коллектив авторов (Ответственные редакторы: Захаров, В. и др.) [Autorenkollektiv (Verantwortliche Redakteure: Zakharov, V. et al.)] (2013). Труды международной научной конференции «Корпусная лингвистика – 2013» / Proceedings of the international conference «Corpus linguistics – 2013». Санкт-Петербург: Санкт-Петербургский государственный университет, Филологический факультет / St. Petersburg: St. Petersburg State University, Philological Faculty, 103-111.

MITTMANN, R. (2015a). „Automated quality control for the morphological annotation of the Old High German text corpus. Checking the manually adapted data using standardized inflectional forms." In: Gippert, J. / Gehrke, R. (Hg.) (2015). Historical Corpora. Challenges and Perspectives. Proceedings of the conference Historical Corpora 2012 (= Korpuslinguistik und interdisziplinäre Perspektiven auf Sprache / Corpus Linguistics and Interdisciplinary Perspectives on Language – CLIP, Band 5). Tübingen: Narr, 65-76.

MITTMANN, R. (2015b). „Automatisierte Zeit- und Dialektzuordnung althochdeutscher Texte." In: Oettinger, N. et al. (Hg.) (2015). Münchener Studien zur Sprachwissenschaft – MSS, Heft 69/2 – 2015. Dettelbach: Röll, 245-256.

PAUL, H. ET AL. (2007). Mittelhochdeutsche Grammatik (= Sammlung kurzer Grammatiken germanischer Dialekte, A. Hauptreihe Nr. 2). 25. Auflage, neu bearbeitet von Th. Klein et al. Mit einer Syntax von I. Schöbler, neubearbeitet und erweitert von H.-P. Prell. Tübingen: Niemeyer.

SONDEREGGER, S. (1980). Art. „'St. Galler Schularbeit'." In: Ruh, K. et al. (1978–2008). Die deutsche Literatur des Mittelalters: Verfasserlexikon. Begründet von Wolfgang Stammler, fortgeführt von Karl Langosch. 2., völlig neu bearbeitete Auflage unter Mitarbeit zahlreicher Fachgelehrter. Band 11. Berlin; New York: de Gruyter, Sp. 1049-1051.

SPLETT, J. (1993). Althochdeutsches Wörterbuch. Analyse der Wortfamilienstrukturen des Althochdeutschen, zugleich Grundlegung einer zukünftigen Strukturgeschichte des deutschen Wortschatzes. 3 Bände. Berlin: de Gruyter.

STEINMEYER, E. V. (1916). Die kleineren althochdeutschen Sprachdenkmäler. Berlin: Weidmann.

Armin Hoenen, Lela Samushia

Gepi: An Epigraphic Corpus for Old Georgian and a Tool Sketch for Aiding Reconstruction

In the current paper, an annotated corpus of Old Georgian inscriptions is introduced. The corpus contains 91 inscriptions which have been annotated in the standard epigraphic XML format EpiDoc, part of the TEI. Secondly, a prototype tool for helping epigraphic reconstruction is designed based on the inherent needs of epigraphy. The prototype backend uses word embeddings and frequencies generated from a corpus of Old Georgian to determine possible gap fillers. The method is applied to the gaps in the corpus and generates promising results. A sketch of a front end is being designed.

1 The Old Georgian Corpus

Basis for the corpus are the transcriptions present on the TITUS web thesaurus, Gippert (1995).[1] 91 inscriptions have been transcribed into digital form and annotated. The corpus comprises Old Georgian inscriptions with the oldest dated to the 5^{th} century A.D. written in Old Georgian Majuscle (Asomtavruli). However, some of the inscriptions stem from the new Georgian period and are written in the modern version of the alphabet (Mxedruli). The majority of inscriptions are building inscriptions (churches), yet there are some gravestone inscriptions and inscribed crosses and other objects. Of special importance for regional and national history are people mentioned mostly on gravestones and correlated data from the inscriptions. As Georgian has been written in three alphabets throughout its history, all inscriptions have been transcribed into the modern version of the alphabet in previous projects.

1.1 Corpus generation

Whilst the corpus is online and accessible via the Titus archive, a translation and annotations have been added. Additionally, the corpus has been transformed into the TEI-format in a way conforming to EpiDoc guidelines. EpiDoc, according to their website[2] is "an international, collaborative effort that provides guidelines and tools for encoding scholarly and educational

[1] http://titus.fkidg1.uni-frankfurt.de/texte/etcg/cauc/ageo/inscr/carcera/carce.htm
[2] https://sourceforge.net/p/epidoc/wiki/About/:last accessed on 07.02.2017

editions of ancient documents" which originated from an effort for publication of ancient inscriptions. In technical terms it uses a subset of the TEI. EpiDoc provides guidelines for the encoding of ancient documents, which the Old Georgian Corpus follows.

Each inscription is encoded in its own tei-xml file in order to ensure complete informativity on metadata and textual levels. The header contains meta information such as language, alphabet, place and time of the inscription as well as a link to its images if available on the TITUS web thesaurus which hosts the inscriptions electronically prepared at the National Museum of Georgia for the Georgian National Corpus (GNC), which they are part of.[3] The body of the document contains the four text divisions typical for EpiDoc: edition, translation, commentary and bibliography.

Annotations are applied to the text in the modern transcription. This transcription forming the TITUS base text previously already included expansions of abbreviations, fillers of gaps, most probable readings of unclear letters, letters the scribe had omitted and so forth (the canon of epigraphic annotation). The modern trancription thus displays one reconstructed text version for the inscription (where reconstruction was possible) and is consequently stored in the text division *edition*. Besides, each file provides the original characters (similar to the text in majuscules in Latin) preserving original linebreaks. Alongside, in a separate text division a full English translation is provided, which has been newly compiled and added to the corpus. People, titles, places and dates have been annotated in order to enable semantic technologies at later stages. Named entity annotation is encoded through the tag named *term* specified by its attributes *type* and *subtype*.

Figure 1 illustrates some of the mentioned encodings. The Georgian abbreviation tradition is especially complex and features many models, see Boeder (1987). Contraction, the mode of abbreviating by first and last letter which gained prominence in the Christian era, compare for instance Driscoll (2009) was very prominent in Old Georgian (abbreviation 1). According to (Danelia and Sarzhveladze, 2012, p.312), the following types of abbreviation are available in Old Georgian: the abbreviation of a word to its initial letter, suspension, contraction and elision of vowels. Suspension is very rare and only found on epigraphic monuments (it is not evidenced in manuscripts). Unlike manuscripts, in epigraphy often uncommon, unfamiliar abbreviations are present, which are difficult to decipher. When it came to suffixes, in Old Georgian affix chains are quite common. In order not to lose the meaning, the suffixes had to be encoded in the abbreviation and scribes may have had different opinions (apart from different spatial considerations) on how to extend the contraction principle consistently in this case (abbreviation

[3] `http://titus.uni-frankfurt.de/indexe.htm`: last accessed on 10.02.2017, `http://gnc.gov.ge/gnc/static/portal/gnc.html`,`http://museum.ge`

Abbreviation 1: ***k(rist')e***

```
<expan>
        <abbr>ქ</abbr>
        <ex>რისტ</ex>
        <abbr>ე</abbr>
</expan>
```

Abbreviation 2: ***k(rist')h(s)i***

```
<expan>
        <abbr>ქ</abbr>
        <ex>რისტ</ex>
        <abbr>ჩ</abbr>
        <ex>ს</ex>
        <abbr>ი</abbr>
</expan>
```

Abbreviation 3: ***k(rist')hsi***

```
<expan>
        <abbr>ქ</abbr>
        <ex>რისტ</ex>
        <abbr>ჩსი</abbr>
</expan>
```

Abbreviation 4: ***a(gh)m(a)***

```
<expan>
        <abbr>ა</abbr>
        <ex>ღ</ex>
        <abbr>მ</abbr>
        <ex>ა</ex>
</expan>
```

Named Entities: ***mepeta mepe davit***

```
<term type="namedEntity" subtype="title">მეფეთა მეფე</term>
<term type="namedEntity" subtype="anthroponym">დავით</term>
```

Filled Gap and Line Break: ***[va]r***

მე ვარ, რომელიც, თუ მე ვარ, ასრე გახდება, როგორც მე`<lb n="8"/>`
`<supplied reason="lost">`*ვა*`</supplied>`*რ.*

Figure 1: Examples of xml encodings in the corpus.

2 and 3). While contraction was especially important for named entities and in particular biblical individuals and places (nomina sacra), for other word classes other ways of abbreviation are found (abbreviation 4). We annotated titles such as king of kings (named entities) in order to relate inscription type to state organization and to better distinguish individuals of the same name.

Throughout the corpus one sees that inscriptions are fragmentary, some to the extent not to allow a full reconstruction of their texts. On average, an inscription had roughly 33 words, 4 gaps and 11 abbreviations.[4] Experts on language and inscriptions have been able to provide hypotheses about the full text of many inscriptions. However, many gaps remain. Not only for the already encoded inscriptions, but also for a planned extension to the corpus some computer-aided assistance in the reconstruction could be welcome. Since largely transcription and other epigraphic work is done in digital environments already, this paper asks: Can there be a tool assisting in reconstructing the complete texts of inscriptions? What will distinguish

[4] Not all abbreviations are counted here since some cannot be read or are concealed in undeciphered gaps.

such a tool from the traditional methods and resources such as lexica of abbreviations, lists of historical named entities and so forth.

2 Towards a Tool: Necessities

For reconstruction, a tool in the digital medium could be designed which assists in two important excerises of the epigrapher: expanding abbreviations and filling gaps. For this purpose, the text of the inscription could be represented digitally, where abbreviations and gaps could be marked and filled with precomputed guesses. However, machine leraning and related techniques have seemingly not yet been applied much to epigraphy, compare Bodel (2012). Some studies on abbreviations and word prediction in *psycholinguistics* may provide interesting and relevant insights eventhough they are not replicating the epigraphic context, see for instance Yang et al. (2009); McWilliam et al. (2009); Slattery et al. (2011); Taylor (1953).

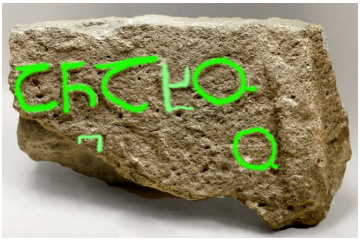
In computation, for tasks similar to epigraphic reconstruction such as *abbreviation generation*, *sequence prediction* and *spelling correction* feasible solutions have been found. But, those often rely on pretrained statistical models which need large amounts of input data. An example is the application of ngram language models for sequence prediction, where Manning and Schütze (1999) note that ngram models to be effective usually need large amounts of training data.[5] Even the full amount of Old Georgian data digitally available [6] is still not large enough to perform and thoroughly evaluate the majority of such approaches. Methodologically, there is an additional factor complicating assessment: Any gap can hold any number of abbreviations making gap filler generation (GFG) a more complex task than simple sequence prediction or abbreviation generation.

Additionally, the epigraphic record is very heterogeneous with the easier cases often already manually solved. In order to exemplify the heterogeneity of the epigraphic record and thus the range a tool aiding in reconstruction has to be able to address, we give some examples from the Georgian inscriptions, images come from the Corpus of Old Georgian Inscriptions.[7] On the one end there are inscriptions with so fragmentary evidence that no super computer can probably ever help to decipher the message, on the other there are reconstructions as trivial as to be performed without much effort even by laymen correctly.

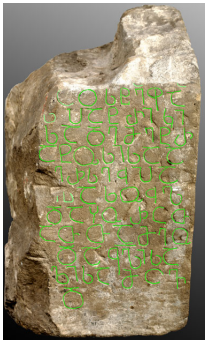
[5] See chapter 6 for discussion, (Manning and Schütze, 1999, p.201): "In general, four gram models do not become usable until one is training on several tens of millions of words of data."

[6] A subcorpus on Old Georgian not containing different redactions of the same texts from the TITUS server comprises roughly 4 million words.

[7] Collection under: `http://titus.fkidg1.uni-frankfurt.de/texte/etcg/cauc/ageo/inscr/carcera/carce.htm`



Although some letters survived, the extent and the placement of the gaps make a complete reconstruction almost impossible.



Here, the broken off part to the left can be reconstructed with a good level of confidence, since each line has more surviving than missing letters and since the amount of missing letters is of a minor magnitude. Also, there are few abbreviations.

Finally, in this example, only abbreviations of moderate difficulty have to be expanded, which could be done by a beginner to epigraphy knowing Old Georgian.

Facing such variety and difficulties, rather than provide a completed feasible solution for GFG, which given the scope of this article and the current landscape of computational epigraphic reconstruction would seem unrealistic, this paper primarily aims at making computational scholars aware of the inseparable interplay of abbreviation and gap which so characterizes the epigraphic record in many epochs, regions and languages and which may represent a new computational challenge. Towards a proof of concept however, a basic method for GFG is being formulated and tested.

In order to demonstrate the utility of such a tool, we concentrate on examples promising to yield some useful results. This is why we restrict ourselves for the time being to single words if possible not on broken off edges, the extent of which is unclear. We argue that if we are able to provide useful guesses for these, then larger units might be in reach for future research.

3 Method

We are looking for lexical matches of the gap context comparing two methods, pure frequency based cues and word embedding based cues. In the face of formulas and a very standardized language of inscriptions pure frequencies and conditional frequencies (of a word given a predecessor or follower) may be a sufficiently strong cue and could be feasible as a baseline. Word embeddings Mikolov et al. (2013a) on the other hand can be used for sequence prediciton since their training includes an optimization of immediate contextual similarity. To this end, semantic and syntactic similarity are captured by word embeddings which makes them a possible cue for a gap filler. Furthermore, Mikolov et al. (2013b) state that "neural network based language models significantly outperform N-gram models", compare Bengio et al. (2003); Mikolov et al. (2011); Schwenk (2007). In fact, in a pre-experiment, we found an ordinary n-gram language model to perform well only for the preditcion of the content of very short gaps. In order to generate word embeddings, we can later use for GFG, we compiled yet another corpus of Old Georgian texts fromt the TITUS archive.

3.1 Corpus

The TITUS website provides texts for many ancient languages and is (one of) the most comprehensive archive(s) (in close collaboration with the GNC) for Old Georgian text. Among the texts present are the Bible, lectionaries, hagiographical, theological and apocryphal texts, psalms and odes, song, historical texts, homiletic and exegetic texts, liturgical texts, canonical law texts, philosophical texts and for instance an astrological and a grammatical

text. Texts are often translated (from ancient Greek, Syriac and Armenian). For details, see the website.

These texts are thus of different genres than inscriptions, but the language stage is essentially the same. We extracted a subcorpus of roughly 4 million words, where for instance critical aparatuses or differing redactions have been omitted and only the critical text been taken. Punctuation as present has been separated from the tokens and tokens arranged so that sentences were approximately arranged in lines, as is usual for word embedding training. We do not entirely exclude the presence of noise. From the corpus, word2vec generated roughly 230,000 vectors for the wordforms in the corpus.

3.2 Approach

Each inscription was processed. An inscription internal gap was detected and the following mechanism tried to generate a filler.[8] First, the context of the gap was extracted. Here, ignoring any space within the gap(s), the continuous context to the left and right of the current gap until the next/previous space character has been extracted. If a subsequent gap was directly adjacent, there would be more than one gaps in such a "word". For instance same[bis]a[j] was so captured. Square brackets mark gaps, letters within are reconstructed, samebisaj means 'from Trinity'. This was then converted to a regex by simply substituting the letters of the gap by a placeholder: (same...a.).[9]

The regex was then used to match all candidates conforming to this pattern in the database of words of the Old Georgian corpus from which the word embeddings have also been generated.[10] [11] The outcome was a list of candidate fillers. However, depending on the extent and position of the gap, the number of fillers could easily become large. When one thinks of an aid for reconstruction, confronting the reconstructor with a large number of tokens, half of which is probably quite unlikely, will not be statisfactory. Therefore, we tried to use different cues for ranking candidates. Each candidate receives three values, firstly the cosine vector similarity to the word vector of the previous word if this word is in the lexicon (in the

[8] For the time being, gaps at the beginning or end of lines were left aside since their extent may be hard to estimate and validate, while the mechanism elaborated is under more based on gap breadth information.

[9] A more sophisticated approach would be to use the true breadth of the gap if annotated in absolute numbers. One could then assign a typical breadth to each letter and check if fillers are suitable for the gap at hand. A possible filler in its most condensed form should not be longer than the gap and its fully spelled out form not shorter. The way in which to generate the most condensed or gap matching form would pertain to abbreviation generation. One could for instance take the first letter of each word.

[10] For training, we used the default settings apart from the minCount feature which we set to 1 since the corpus is not huge and in this way, we capture hapax legomena and significantly enlarge embedding vocabulary.

[11] Neo4j was our data base system accessed via java.

Old Georgian corpus) and not gappy. Secondly, the same for the following word and thirdly, the filler is given its frequency from the Old Georgian corpus (in which it must occur since it has been extracted from there). From these values, we generate a weight for the candidates. This enables us then to sort the fillers and so limit the number of candidates to be offered to the reconstructor to a number he/she may deem useful. Such a number could be the top 10 for instance. However, since the weight may be the same for several candidates, we allow the limit to be exceeded and to include all candidates with a weight larger or equal to that of the tenth candidate.

3.3 Results

In the Old Georgian corpus, in overall 65 gappy "words" no fillers were retrieved for 25, whereas 26 of the 40 filler sets contained the correct filler. Results are encouraging, the correct filler was generated at a ratio of 0.65 decreasing to 0.6 if limiting the output to the top weights as described above using frequency as weighting cue. Recall was 0.62.[12] The average number of top fillers generated was roughly 7 which is not too confusable in terms of overview. Limiting to the top ranks had another effect, namely the Damerau Levenshtein distance, Damerau (1964) of the fillers to the correct solution decreased for more than half to be 4.21 which shows that even if the correct filler has not been included it is not unlikely to have a moderately similar or similar word in the top fillers. Using the word embedding cues, and only in case the previous and next word would both not be present in the word embedding lexicon frequency, deteriorated results. Taking the similarity to the last word if present (otherwise frequency) resulted in precision of 0.525, taking similarity to the next word if present (otherwise frequency) resulted in a precision of 0.5 and combinations such as the average of the similarities of last and next words if both were present, if only one of them was present that value and only in case none was present frequency, was still worse at 0.475. The correctly captured fillers from the embeddings however were largely coinciding but no subset of the ones captured by frequency.

3.4 Discussion and Post Experiment

Frequency is plainly connected with probability through bare counts, while word embeddings capture syntagmatic and paradigmatic similarity. Similarities to previous and following words performed at an almost equal level. One reason for the reduced performance in respect to frequency using only the immediately adjacent neighbours (the larger the context, the more probable the occurrence of a gap or abbreviation within the context) could be the

[12] Using fewer dimensions (10) only improved the result marginally in lowering the average rank at which the correct filler was to be found.

nature of language, namely the dichotomy between high frequency function words and content words. For the former, naturally many more neighbours exist in a training text which may make their vectors less specific and in turn less reliable ranking cues.

However, the amount of data tested on is not sufficient to conclude anything. Consequently, we tested the same method on 1,000 inscriptions of a Latin data base for inscriptions, the *Epigraphic Database Heidelberg.*[13] The text database, we used for computing word embeddings and extracting the frequency lexicon were the Latin Wikipedia[14] and the classical texts of the Packard Humanities Institute.[15] We found the same pattern as in Old Georgian, meanwhile with lower recall and precision. Frequency alone was the best cue. More research may shed light on the true reasons behind this pattern.

For the Latin dataset, another approach is feasible. A preliminary attempt is described and first results given in what follows as an outlook to future elaboration. Since there are more than 70,000 inscriptions, it makes sense to produce for instance 10 chunks of equal size (in terms of numbers of inscriptions). Then for gaps in any 1 chunk symbolizing the unreconstructed inscriptions, one can extract context and use pattern search in the 9 training chunks symbolizing the until then reconstructed inscriptions. Since inscriptions are highly stereotypical this may lead to good results. To test this assumption, in a small follow up on Latin, we extracted the context, this time regardless of spaces until the next/previous gap and then matched the resulting pattern *left_context.+right_context* from the inscriptions in the 9 held out chunks. The matches were checked for suitable length given the gap breadth. As described above, the most condensed form (each word abreviated by its first letter) should not be significantly longer and the fully spelled out form not significantly shorter than the space the gap offers. For each gap, we decreased context size by one character on each side and repeated matching until the context consisted in one character only. The matches (or fillers) were weighted for the length of the context at which they had been matched and for frequency of the match ($\sum_{i=1}^{n} |left\_context| + |right\_context|$ for n matches).

Here, we found a recall of 0.33 with the correct filler being present at a rate of 0.46 in the filler sets, whilst at a rate of 0.2 the correct filler was in the top 10 fillers. The average DL of the top fillers was 3.96 for those filler sets, where the correct match was not present. The highest ratios of correct matches per context lengths were achieved with longest contexts and balanced contexts, but length was a better cue than balance. To exemplify, a context of 5 characters to the left and 5 characters to the right is in total

[13] `http://edh-www.adw.uni-heidelberg.de/`

[14] `https://la.wikipedia.org`: last accessed on 16.12.2015

[15] `http://latin.packhum.org`: last accessed on 09.12.2015

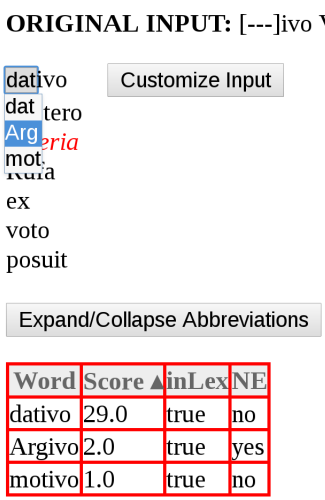


Figure 2: Simple front end example: The slightly transformed original transcription is visible in the first line. For each word, either the user is provided with a dropdown list restricted to the most probable automatically generated fillers or can choose to edit the gap filler manually. Abbreviations can be collapsed or expanded to support imagination of an original in the reconstructive process. A sortable table at the bottom informs him/her of all possibilities, which can be considerably more than in the thresheld dropdown menu and which contains additional information.

a 10 character context, but these contexts captured relatively less correct fillers than contexts of 0 characters to the left but 9 to the right. It seems that the longer a match in a continuous context, the better the cue.

4 User Interface

For the development of an "EpigraphyHelper" a user front end would have to be set-up. A sketch of this has been done using a platform independent HTML/Javascript solution which provides the most probable fillers in a drop-down container, see Figures 2 and 3. Future design and usability of this rendering should be made subject of an online survey for domain experts. The front end once finalized is completely independent from the technical backend, which is to say that the current method of generating gap fillers can be exchanged as soon as more effective methods are available.

The front end has several features. Firstly, the original transcription is presented on top, giving the epigrapher the context, he/she habitually encounters. Then per line each word is rendered either as non changeable text if readable as such on the inscription ('ex votu posuit' in the example) or

ORIGINAL INPUT: [P/ublio/] [M]ummio [P/ubli/] [f/ilio/] [Gal/eria/] [S]isenna[e] [Rutiliano] Xv[ir/o/] [stlitibus] [iudicandis] [---------]

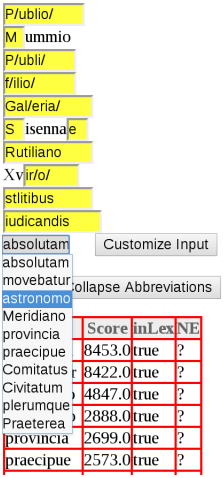


Figure 3: More complex example: Per word a separate line is assumed. Gaps filled by previous scientists as most probable reconstructions are editable. Visible and reconstructed abbreviations can be collapsed and expanded. They are marked differently.

with an expanded abbreviation, where the expansion is rendered in red and italics (Val*eria* in the example) or for each word which was reconstructed within a gap an editable textfield appears with yellow background, where abbreviations are marked by slashes (P/ublio/ in the example). Abbreviations can be collapsed and expanded per button. Finally, for gaps which have not been reconstructed, the algorithm computes candidates as described above and displays them in a drop-down list (Argivo in the example). Following Shneiderman's principle Shneiderman (1996), only in case of demand can the user obtain a sortable table with many more possibilities and additional annotations for the words. If none of the proposed fillers is deemed correct, the user can activate a 'Customize Input' button and transform the drop-down into an editable textfield.

5 Future Work and Experimentation

Of course, epigraphers have tried hard and succeeded well in reconstructions of inscriptions both internalizing abbreviation and text completion, connecting this with typical functional epigraphic formula and historical events and individuals. The frustration of not being able to decipher the message of certain inscriptions is probably a well known feeling for epigraphers and each one may have found his/her own way to deal with this issue. An application of AI to epigraphy should therefore not pretend to be a remedy for this frustration since it is clear that a too fragmentary inscription cannot be reasonably reconstructed. Yet, since the capacity of the human brain to keep in mind all relevant words, names and orthographic variants (and in consequence all possible reconstructions) is limited in comparison with a computer, a reconstruction aid may, in the best case find reasonable fillers for some of the not yet reconstructed gaps which had slipped the conscience of previous reconstructors. Especially in the case of Named Entities, a vast array of possibilities exists.

Furthermore, unreasonable candidates which such a system produces can be discarded by a human expert in a matter of seconds, leaving the technologically open user with a positive net outcome. One crucial question for an application of AI to epigraphy will be at which rate good guesses can be produced. Assessing such a question, databases such as the epigraphic database Heidelberg or the database Clauss/Slaby[16] may be seen as a benchmark dataset which will enable computer scientists to evaluate their approaches against the recontructions already conducted.

[16] `http://www.manfredclauss.de/`

6 Conclusion

A corpus of Old Georgian inscriptions has been compiled. Additionally, a tool for epigraphic reconstruction has been sketched in order to raise awareness in the Computer Scientific community that such a task exists, that data sets for its evaluation exist and that the task is an interesting computational challenge involving both abbreviation resolution or generation and sequence prediction. To this end, we have only been able to show that in the case of Old Georgian, thanks to a large resource of Old Georgian texts from the internet, a reconstruction aid can produce on average 7 fillers for roughly 60% of gaps with 60% of filler sets containing the correct solution. We hope for more general results and solutions in the future.

References

Bengio, Y., Ducharme, R., Vincent, P., and Jauvin, C. (2003). A neural probabilistic language model. *Journal of machine learning research*, 3(Feb):1137–1155.

Bodel, J. (2012). Latin Epigraphy and the IT Revolution. *Proceedings of the British Academy*, 177:275 – 296.

Boeder, W. (1987). Versuch einer sprachwissenschaftlichen Interpretation der altgeorgischen Abkürzungen. *Revue des études géorgiennes et caucasiennes*, 3:33 – 81.

Damerau, F. J. (1964). A technique for computer detection and correction of spelling errors. *Communications of the ACM*, 7:171–176.

Danelia, K. and Sarzhveladze, Z. (2012). *Kartuli p'aleograpia [Georgian Paleography]*. Nekeri.

Driscoll, M. (2009). Marking up abbreviations in old norse-icelandic manuscripts. In *Medieval Texts–Contemporary Media*. Ibis.

Gippert, J. (1995). Titus. das projekt eines indogermanistischen thesaurus ("titus. the project of an indo-european thesaurus"). *LDV-Forum*, 12(2):35–47.

Manning, C. D. and Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA, USA.

McWilliam, L., Schepman, A., and Rodway, P. (2009). The linguistic status of text message abbreviations: An exploration using a stroop task. *Computers in Human Behavior*, 25(4):970–974.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013a). Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013b). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Mikolov, T., Deoras, A., Kombrink, S., Burget, L., and Černockỳ, J. (2011). Empirical evaluation and combination of advanced language modeling techniques. In *Twelfth Annual Conference of the International Speech Communication Association*.

Schwenk, H. (2007). Continuous space language models. *Computer Speech & Language*, 21(3):492–518.

Shneiderman, B. (1996). The eyes have it: A task by data type taxonomy for information visualizations. In *Proceedings of the 1996 IEEE Symposium on Visual Languages*, VL '96, pages 336–, Washington, DC, USA. IEEE Computer Society.

Slattery, T. J., Schotter, E. R., Berry, R. W., and Rayner, K. (2011). Parafoveal and foveal processing of abbreviations during eye fixations in reading: making a case for case. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 37(4):1022.

Taylor, W. (1953). Cloze procedure: A new tool for measuring readability. *Journalism Quarterly*, 30:415–433.

Yang, D., Pan, Y.-c., and Furui, S. (2009). Automatic chinese abbreviation generation using conditional random field. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*, pages 273–276. Association for Computational Linguistics.

Author Index

Marcel Bollmann
Sprachwissenschaftliches Institut
Ruhr-Universität Bochum
Universitätsstr. 150, 44801 Bochum, Germany
https://marcel.bollmann.me/
bollmann@linguistics.rub.de

Stefanie Dipper
Sprachwissenschaftliches Institut
Ruhr-Universität Bochum
Universitätsstr. 150, 44801 Bochum, Germany
https://www.linguistics.rub.de/~dipper
dipper@linguistics.rub.de

Armin Hoenen
CEDIFOR
Institut für Empirische Sprachwissenschaft
Johann Wolfgang Goethe-Universität Frankfurt
Senckenberganlage 31 (Juridicum), 60325 Frankfurt am Main, Germany
https://hucompute.org/team/armin-hoenen/
hoenen@em.uni-frankfurt.de

Thomas Klein
Institut für Germanistik und Vergleichende Literaturwissenschaft
Universität Bonn
Am Hofgarten 22, 53113 Bonn, Germany

https://www.germanistik.uni-bonn.de/institut/abteilungen/germanistische-linguistik/abteilung/personal/klein_thomas
thomas.klein@uni-bonn.de

Roland Mittmann
Institut für Empirische Sprachwissenschaft
Johann Wolfgang Goethe-Universität Frankfurt
Senckenberganlage 31 (Juridicum), 60325 Frankfurt am Main, Germany
http://titus.uni-frankfurt.de/personal/mittmann.htm
mittmann@em.uni-frankfurt.de

Florian Petran
Sprachwissenschaftliches Institut
Ruhr-Universität Bochum
Universitätsstr. 150, 44801 Bochum, Germany
https://www.linguistics.rub.de/~petran/
florian.petran@gmail.com

Lela Samushia
Institut für Empirische Sprachwissenschaft
Johann Wolfgang Goethe-Universität Frankfurt
Senckenberganlage 31 (Juridicum), 60325 Frankfurt am Main, Germany
samushia@em.uni-frankfurt.de