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 'o̬', 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 | n̄emen | 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

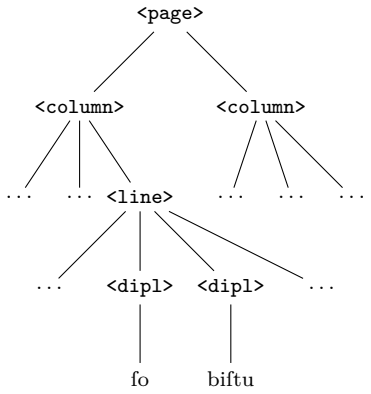


**Figure 2:** Simplified example of the layout hierarchy in CorA-XML

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/

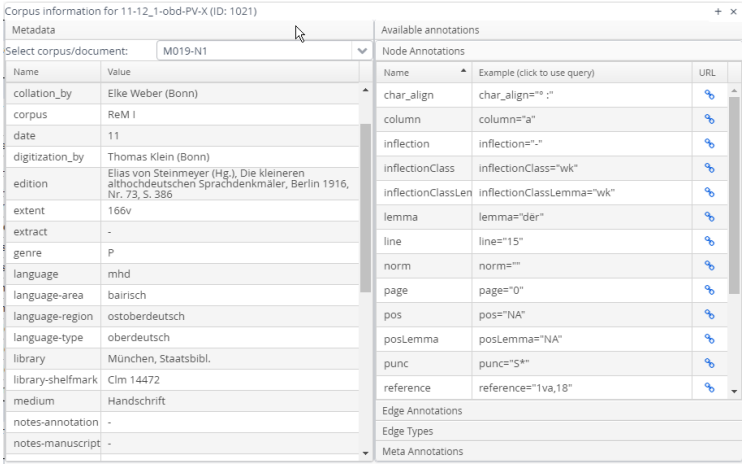


**Figure 3:** Part of the meta annotations for the text "Augensegen" ("blessing of the eyes"). Some of the meta annotations are important for diachronic searches, others (such as the annotators responsible for digitization) are merely informative.

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.

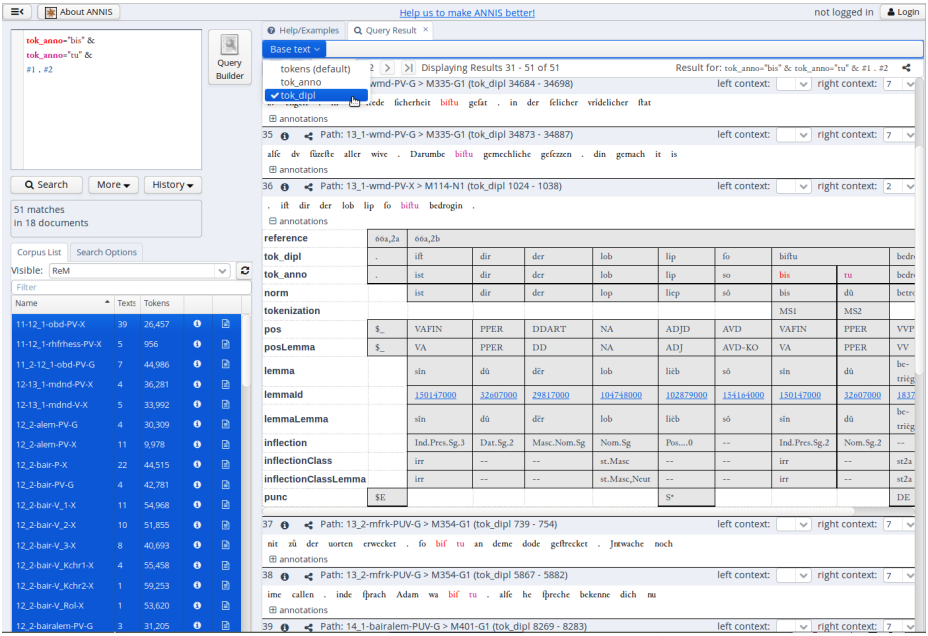


**Figure 4:** ANNIS window showing the results of a search for the sequence "bis tu" in modernized form. Part of the subcorpus list is shown on the lower left.

`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.

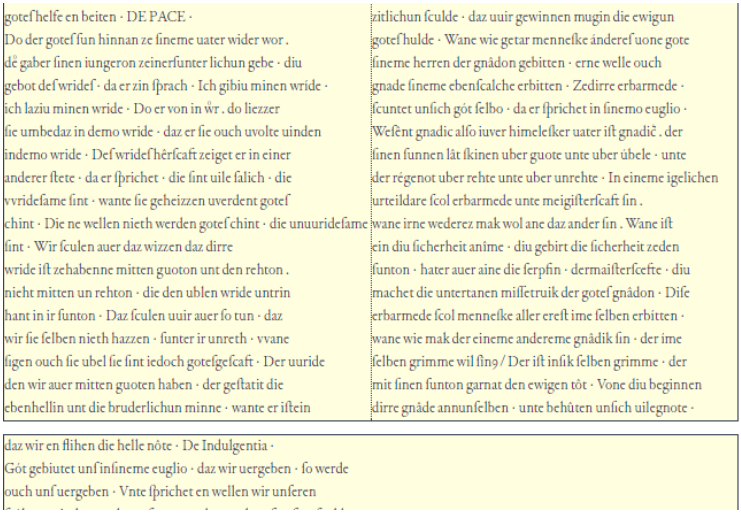


**Figure 5:** Diplomatic full text view of the Middle High German translation of Alkuin's "De virtutibus et vitiis"

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.