RIDGES Herbology – Designing a Diachronic Multi-Layer Corpus

Authors

Carolin Odebrecht
Humboldt-Universität zu Berlin
Institut für deutsche Sprache und Linguistik
Korpuslinguistik und Morphologie
Dorotheenstraße 24, D-10117 Berlin
carolin.odebrecht@hu-berlin.de
+49 (0)30 2093 9618

Malte Belz
Humboldt-Universität zu Berlin
Institut für deutsche Sprache und Linguistik
Phonetik/Phonologie
Dorotheenstraße 24, D-10117 Berlin

Amir Zeldes
Department of Linguistics
Georgetown University
Poulton Hall
1421 37th St. NW, Washington, DC 20057

Anke Lüdeling
Humboldt-Universität zu Berlin
Institut für deutsche Sprache und Linguistik
Korpuslinguistik und Morphologie
Dorotheenstraße 24, D-10117 Berlin

Thomas Krause
Humboldt-Universität zu Berlin
Institut für deutsche Sprache und Linguistik
Korpuslinguistik und Morphologie
Dorotheenstraße 24, D-10117 Berlin

Acknowledgments
We would like to thank Vivian Voigt and Laura Perlitz, our two very capable student assistants, who helped with many aspects of corpus creation and consistency checking. We would also like to thank the many students who took part in the digitization and basic annotation, as well as Uwe Springmann, Florian Zipser and three anonymous reviewers for comments that greatly improved the manuscript. The project has been generously funded by two Google Digital Humanities Research Awards.

1. Introduction

This paper is concerned with the development of a diachronic corpus containing German excerpts from herbals, which has been constructed to study the emergence and change of scientific registers in a vernacular language of Europe (for more on the background of register development in a vernacular language see Klein 1999; Pahta and Taavitsainen 2010, among many others). Up to the 16th century almost all scientific writing in Europe was conducted in Latin. The different language communities changed to their respective vernacular languages at slightly different points in time; German being fairly late. The change was slow: It took about 300 years between a point in time when virtually all scientific communication was carried out in Latin to a point in time when almost all scientific publications were in a language other than Latin, and the process affected different text types, fields, and topics differently (Pörksen 2003; Vikør 2004). As such, it forms a prime example for the crystallization of a new register for a language, a topic of great interest for variation studies, linguistic theory, and cultural heritage studies, to name a few.

Since register changes affect all linguistic and extra-linguistic levels, register studies are always multifactorial (see Biber and Conrad 2009 for an overview), and register change can only be carried out using deeply and consistently annotated diachronic corpora. The construction and annotation of *historical* corpora is challenging in many ways (see Lüdeling et al. 2005; Claridge 2008; Rissanen 2008; Kytö 2011; Kytö and Pahta 2012, among many others). The construction of *diachronic* corpora (i.e. corpora covering a sequence of historical periods) has a number of additional issues. The lexicon changes with the formation of terminology, spelling regularities emerge, and word-formation, syntax, and text structure developments. All of this poses challenges to consistent annotation. At the same time we see changes in typesetting and printing methods which complicate automatic digitization. Furthermore, the emergence of scientific texts cannot be studied without taking into account the concurrent advancements in school systems, scientific fields and methods and university structure.

All these topics need to be covered and technically supported in a broad corpus design and architecture planned for a variety of studies on the development of the language of science, entailing special aspects of digitization, annotation, and natural language processing to produce a coherent and useful resource. In the planning of such a resource the following questions have to be addressed:

- What kind of transcription and which layers of normalization are essential for a diachronic corpus?
- How can we assign consistent categories to text types, words, utterances, etc. over time? How can we be sure that the same label refers to the same concept?
- What kind of corpus architecture is needed?
- How can we ensure comparability to other historical and modern corpora (of German and beyond)?
- How can we make the corpus reusable for other research questions in different scientific fields?

The project **R**egister **i**n **D**iachronic **Ge**rman **S**cience (RIDGES)[1] aims to address these questions for German by constructing a diachronic multi-layer corpus (Section 2.1). In this paper, our main focus will be on the challenges and solutions that we have found in the representation of diachronic data in German as the emerging language of science. We will address both the aspects of the technical infrastructure and the conceptual levels of analysis that together ensure an extensible, reusable and comparable corpus for the study of register development across time. Several case studies will illustrate how our corpus can be used to study the different levels of interpretation.

In Section (2) we will introduce the corpus design (2.1) and the general corpus architecture (2.2). Building on these we will discuss different layers of corpus annotation in Section (3), starting with transcription (3.1) and different normalization layers (3.2), before talking about graphical and structural annotations such as line breaks and rendering

[1] http://korpling.german.hu-berlin.de/ridges/index_en.html. The corpus is freely available under a CC-BY license at the LAUDATIO Repository http://hdl.handle.net/11022/0000-0000-2D85-8. Accessed 1 March 2016.

(3.3), and different layers of linguistic annotation (3.4). We will discuss our decisions vis à vis other historical and diachronic corpora and their architectures (3.5). In Section (4) we will exemplify the need for an open, multi-layer architecture by a number of case studies that involve some of the different annotations.

2. The RIDGES Herbology Corpus

2.1 Corpus Design

In order to study the development of the scientific language throughout the period of interest, we require a subject domain that is sufficiently well represented in all subperiods. To this end, we have chosen to focus on excerpts from herbals, which are available throughout much of the written transmission of German, first as manuscripts but from an early point in time as prints (see e.g. Riecke 2004; Gloning 2007; for an overview of the transmission, for more specific issues regarding herbal and medicinal texts in German see e.g. Habermann 2001; Riecke 2007; Squires 2010). Other disciplines, by contrast, did not exist for the entire period of time covered by the corpus, or meant much more disparate things across periods (e.g. the transition from astrological to astronomical texts). The RIDGES corpus, in version 4.1 used in this paper, contains 29 excerpts from 24 publications of herbals, ranging from 1478 to 1870, with approximately 30 years between the texts. New texts are added to the corpus at irregular intervals.[2] The corpus contains excerpts of about 3000–4000 words each such as herbal treatises, lectures, and scientific texts (currently 154,267 tokens in total).[3] Each document is stored with comprehensive bibliographic metadata such as title, author, editor, publication place, publisher and year as well as other metadata concerning the preparation of the text. The topics of the early texts in the corpus are medicinal (describing a medical problem and its herbal remedy), and later texts also contain botanical and chemical information. The early texts are often (liberal) translations or collections of earlier Latin and Greek texts (famous treatises by Galenus, Paracelsus, Dioscorides, etc.), while later authors add their own observations and, even later, scientific experiments and methods are described. The texts were published in different parts of Germany, Switzerland and Austria and therefore vary with respect to dialect. As the basis for digitization, freely available, good quality scans of the historical books provided by research libraries[4] were chosen. If a historical book is not captured by such services we use scans from Google Books[5]. The texts are digitized diplomatically (Section 3.1), normalized (Section 3.2), and deeply annotated (Sections 3.3 and 3.4).

The corpus is annotated in MS Excel format and converted with the converter framework Pepper[6] (Zipser and Romary 2010) into several formats. The corpus is stored in the stand-off format PAULA XML (Dipper 2005), and its annotations are accessible via ANNIS[7], a browser-based search and visualization platform (Chiarcos et al. 2008; Krause and Zeldes 2016). The corpus with all formats is archived long-term and extensively documented for reuse scenarios in the LAUDATIO-Repository (Odebrecht et al. 2015).[8]

---

[2] The corpus texts were collected and initially prepared in several graduate and undergraduate seminars at Humboldt-Universität zu Berlin. The texts were extensively corrected and checked for consistency before publication. The corpus is growing; Version 5 (containing 36 excerpts, 183.724 tokens) was published in June 2016.

[3] The size of the text excerpts is chosen depending on the teaching context, i.e. whether the data is collected in a graduate or undergraduate seminar.

[4] Bayerische Staatsbibliothek https://www.bsb-muenchen.de/, Münchener Digitalisierungszentrum http://www.digitale-sammlungen.de/, Universitätsbibliothek Heidelberg http://www.ub.uni-heidelberg.de/helios/digi/digilit.html. Accessed 1 March 2016. The corpus is currently based on printed texts only. We used the original version wherever possible (that is, wherever we were able to find a high-quality scan) and the earliest available version otherwise. The complete bibliographical information for each text is given in the metadata. We plan to add some manuscripts at a later stage, and also envision adding some of the Latin sources.

[5] https://books.google.de/. Accessed 1 March 2016.

[6] http://corpus-tools.org/pepper. Accessed 8 June 2016.

[7] ANNIS, which stands for ANNotation of Information Structure, was originally designed to provide access to the data of the SFB 632 - Information Structure, see http://corpus-tools.org/annis/. Accessed 1 March 2016.

[8] LAUDATIO, which stands for Long-term Access and Usage of Deeply Annotated Information, is an open access repository for historical corpora. http://www.laudatio-repository.org. Accessed 1 March 2016.

2.2 Multi-layer Architecture

Some of the early approaches to historical corpora have relied on inline text and annotations to encode both the primary text and linguistic analyses such as morpho-syntactic information (see Section 3.5 for more discussion). However, many of the questions that we will discuss in in this article, including the study of orthographic, grammatical and conventional changes, require more complex architectures. This applies perhaps most strongly to representations of tokens in older texts, which are much less standardized and call for different approaches to normalization. In this section we therefore want to motivate the need for a multi-layer corpus architecture with the possibility for multiple tokenizations. By tokenization we mean the segmentation of primary data[9] into units (Schmid 2008) and more precisely segmentation into the smallest annotatable units. By annotation we mean the explicit assignment of a category, or tag, to a token or sequence of tokens. We will start by explaining the need for multiple tokenizations.

For modern European languages tokens often correspond to graphemic words (or sequences of characters between white spaces). Technically, however, a token can be any segment that is the base for annotation. In historical texts the decision of what constitutes a word may be difficult because white spaces are distributed in different ways from modern usage (the extreme case being *scriptio continua*, writing without any spaces). A segmentation is an interpretation of the primary data, and – depending on the research question and the assignment criteria – there can be different interpretations (cf. Lüdeling 2011; more on this in Section 3.1). The segmentation directly influences the annotation. As a trivial example, consider cliticized negations such as *don't* or *can't*. If they are segmented as one element, only one part-of-speech tag (pos tag) can be assigned (the pos tag may itself be complex). If they are segmented into several tokens one has to decide where and how to split, cf. Figure (1). While each of the decisions in Figure (1) can be challenged, it must be clear that it is impossible *not* to decide and each decision has consequences: The number of tokens may differ (which is relevant for statistical analysis), and pos tag assignment can vary.[10]

Fig. 1 Different tokenizations for *we can't do that*

<table>
<tr><td>tok_a</td><td>4 units</td><td>we</td><td colspan="3">can't</td><td>do</td><td>that</td></tr>
<tr><td>tok_b</td><td>5 units</td><td>we</td><td>can</td><td colspan="2">t</td><td>do</td><td>that</td></tr>
<tr><td>tok_c</td><td>5 units</td><td>we</td><td>can</td><td colspan="2">'t</td><td>do</td><td>that</td></tr>
<tr><td>tok_d</td><td>6 units</td><td>we</td><td>can</td><td>'</td><td>t</td><td>do</td><td>that</td></tr>
<tr><td>tok_e</td><td>5 units</td><td>we</td><td>can</td><td colspan="2">n't</td><td>do</td><td>that</td></tr>
</table>

Especially in 'non-standard' texts such as historical texts it may be desirable to have different tokenizations, in order to deal with different research questions.

While these tokenizations are the basis for other annotations and are in principle independent of each other, there are research questions for which it is necessary to align the tokens of different segmentations in some way. For example in Figure (1), it is implicitly suggested that the *can't* token of the *tok_a* layer is aligned with both the *can* and *t* token of the *tok_b* layer. Our goal is not to enforce a single minimal tokenization to which the other tokenizations refer, but to allow conflicting segmentations. Three different data models are used to represent RIDGES from a technical standpoint: PAULA XML to serialize the data, ANNIS to allow searching in the corpus and Salt[11] (the internal model

[9] There is an ongoing discussion in corpus linguistics on what constitutes primary data (cf. Claridge 2008; Himmelmann 2012, the discussion involves the roles of originals, pictures (scans), transcriptions, and normalizations). Here, we focus on the technical features of a corpus and do not want to engage in this discussion. We will briefly come back to the different notions of 'text' in Section (3.5).

[10] In Sections (3.1) and (3.2) we will discuss the tokenization and normalization for historical German.

[11] http://corpus-tools.org/salt/ Accessed 8 June 2016.

behind Pepper) for transformations to or from other models. Salt allows us to align tokens by using a common timeline, a concept that has its origin in the annotation of speech data (Bird and Liberman 2001). A timeline is an ordered series of items with optional time-stamp information. This makes it conceptually very similar to a sequence of tokens, but a timeline does not encode any textual information by itself, though tokens can be connected to items in a common timeline. There is no theoretical limit in the number and granularity of items and their time codes. Thus, whenever one of the tokenizations needs a more fine grained segmentation, a new timeline item can be added as required, without influencing the other timeline items. PAULA XML and ANNIS use a very similar concept to implement multiple segmentations.[12]

Using the complex and powerful model described above, RIDGES has several normalization layers (see Section 3). One of them contains a diplomatic version of the text where the tokens correspond directly to words (sequences of characters between white spaces) that the author provided. Annotation layers that pertain to the rendering of the text refer to this layer. Another layer contains a modern German normalization. Annotation layers that pertain to part of speech or modern lemmas refer to this layer. Each tokenization layer can be the basis for one or more annotation layers. For example, each token can be assigned a pos tag or a tag describing typographical features (a category that is assigned to a token will be called a **token annotation**). A sequence of tokens can be categorized as a multi-word expression (an idiom, say), or a sentence type and we will call any category that is assigned to a sequence of tokens a **span annotation**. The *pos* annotation according to STTS in Figure (2) is a token annotation while the *syntax* annotation is a span annotation. Corpora can mix token annotations and span annotations as appropriate.

Fig. 2 Example for token and span annotations, loosely based on *Artzney Buchlein der Kreutter* (1532)
den ſamen trinck mit venchel waſſer
'drink the seed with fennel water'

tok	*den*	*ſamen*	*trinck*	*mit*	*venchel*
pos	ART	NN	VVFIN	APPR	NN
syntax	NP			PP	

The graph-based architectures we use (ANNIS)[13] is flexible enough to handle multiple segmentations and annotations. In our architecture, a corpus always has *1* to *n* tokenizations to which different annotations apply. Neither the number of tokenizations nor the number of annotations is restricted in our model. Annotation layers are technically independent of each other, following a stand-off annotation model (cf. Carletta et al. 2003; Chiarcos et al. 2009) in which each level of information is stored separately. As a result, new annotation layers can be added at any point in time: Each additional annotation layer enriches the corpus, and, conceptually speaking, needs not conflict with or replace another layer. It is also possible to retain multiple versions of annotations produced in earlier iterations of the corpus. As a consequence, it is possible that a corpus contains theoretically conflicting annotations. As an aside, such flexibility ensures that the corpus can be reused by others, since their analyses can be added more easily and searched for concurrently with existing stand-off annotations (cf. Kübler and Zinsmeister 2015, 33–36).

[12] Bird and Liberman (2001) proposed to use character offsets as a substitute for time-stamps in written texts, but since different tokenizations can have different base texts (unlike Figure (1), where the exact same character sequence is tokenized in different ways) this is not applicable to our model. But even without time-stamps, the structure of a timeline allows us to model the alignment between different tokenizations. In contrast to Salt, the PAULA and ANNIS data models do not have the explicit concept of a timeline and thus need a different way to encode it. The solution to this problem is an automatic creation of a single artificial minimal tokenization (cf. Krause et al. 2012), where each artificial token corresponds to a timeline item. The conceptual tokenizations are represented as annotations on top of these artificial tokens and are flagged as segmentation layers. Technically, a segmentation layer is just a normal annotation layer, but flagging it as a segmentation layer makes it behave like one of a set of alternative tokenization layers that the search engine, ANNIS, treats as the basic text of a document. This affects both the initial view of search results and the ability to define search context and distance between search elements.
[13] Other corpus projects using a similar corpus architecture are Falko (Reznicek et al. 2013), PCC (Stede and Neumann 2014), Referenzkorpus Altdeutsch (Donhauser 2015), or Coptic Scriptorium (Zeldes and Schroeder 2015).

Allowing conflicting annotations and thereby increasing the complexity of concepts and corpus architectures, requires an extensive documentation to enable the reuse of the RIDGES Corpus (see Odebrecht 2014). In addition to the full corpus documentation[14], the RIDGES corpus provides extensive annotation guidelines (Belz et al. 2015)[15], and we have formulated sample queries to make the search and analysis in ANNIS easier.

3. Annotation

In this section we will explain how we have pre-processed the corpus: Section (3.1) discusses the transcription, while Section (3.2) deals with multiple normalizations and multiple tokenizations. Based on the different normalizations, Section (3.3) presents the graphical annotations and Section (3.4) the linguistic annotations.

3.1 Transcription

A central issue that is discussed in the preparation of almost all historical corpora (Durrell et al. 2007; Rissanen 2008; Bollmann et al. 2011; Archer et al. 2015) is the tension between the desire for a narrow, diplomatic transcription on the one hand, and the need for a predictable, heuristic annotation of relevant features based on standardized representations on the other hand (cf. Baron et al. 2009). The RIDGES Herbology Corpus handles the problem by allowing for multiple normalizations, which are motivated by linguistic research questions. Depending on the research question, transcriptions vary in their diplomaticity regarding script usage, special characters, typesetting and encoding. Technically, each normalization layer can be regarded as a tokenization layer in the sense described in Section (2).

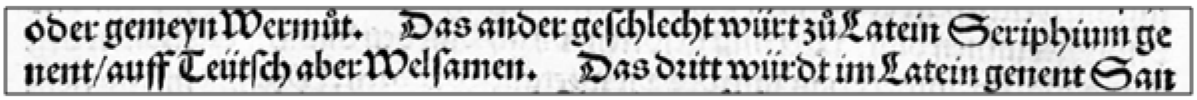
The transcription (called *dipl*) is narrowly diplomatic: We assign each glyph to a Unicode character[16]. Consider Example (1a). The transcription mirrors the historical spelling, spacing, and print space. All characters are taken from Unicode: in (1a), these are, for instance, *aͤ* (U+0061 U+0364), *ſ* (U+017F) and *⸗* (U+2E17). The Unicode standard provides characters for most of the glyphs needed for old German texts.[17] As we have motivated in Section (2.2), our corpus design and the multi-layer corpus architecture allow for multiple segmentations. The first segmentation is applied to the transcription on *dipl*. Separated 'words' at line breaks, be they with hyphenation, as in *Blaͤt⸗* and *lein* 'small leaf' in Example (1a), or without hyphenation, as in *ge* and *nent* 'called' in Example (1b), are treated as two separate tokens (see Section 3.2). Thus, we rely on graphical features for the diplomatic transcription and minimize the linguistic interpretation at this level (in the next examples, underlined words in the translation are hyphenated across a line break in the original).

> **Ex. 1a** Diplomatic transcription, Curioser Botanicus oder sonderbares Kräuterbuch (1675)
> *aber zart / gleich als wenn ſie aus vielen kleinen Blaͤt⸗*
> *lein zuſammen geſetzet waͤren / und wie die Vogelfe⸗*
> *dern auff beyden Seiten geordnet . Bluͤhet faſt wie*
> '... but gentle such as when they are comprised of many small leaves
> and how the bird-feathers are arranged
> from both sides. Blooms almost like...'

[14] For the corpus documentation see http://hdl.handle.net/11022/0000-0000-8253-F. Accessed 16 March 2016.
[15] http://korpling.german.hu-berlin.de/ridges. Accessed 16 March 2016.
[16] For the official Unicode table see www.unicode.org. Accessed 1 March 2016. An anonymous reviewer has asked why we have opted to use precomposed characters when possible and not to use combining diacritics. In principle, the TEI standard has taken an agnostic stance in this matter. Precomposed characters circumvent possible problems with regular expression engines that only have level 1 support for Unicode (e.g. when searching for a single grapheme cluster as described in http://unicode.org/reports/tr18/#Grapheme_Cluster_Mode). Not all glyphs have precomposed characters in Unicode and we use combining characters in this case.
[17] This is generally true even for incunabula which may contain rare glyphs. The Medieval Unicode Fonts Initiative (MUFI, http://folk.uib.no/hnooh/mufi/) is concerned with adding special characters represented in older texts to the Unicode standard. Accessed 1 March 2016.

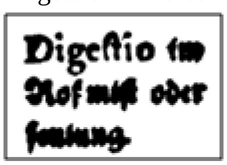

aber zart/ gleich als wenn ſie aus vielen kleinen Blät-
lein zuſammen geſetzet wären/ und wie die Vogelfe-
dern auff beyden Seiten geordnet. Blühet faſt wie

Ex. 1b Separate ‘words’ at a line break, New Kreüterbůch(1543)
oder gemeyn Wermůt . Das ander geſchlecht würt zů Latein Seriphium <u>*ge*</u>
<u>*nent*</u> */ auff Teütſch aber Welſamen . Das dritt würdt im Latein genent San*
‘or common Vermouth. The other kind came to be <u>called</u> Latin Seriphium
but in German Welsamen. The third is called in Latin San...’


oder gemeyn Wermůt. Das ander geſchlecht würt zů Latein Seriphium ge
nent/auff Teütſch aber Welſamen. Das dritt würdt im Latein genent San

The one difference between the original and the typographic layer concerns punctuation. Virgules, commas, full stops, and other punctuation signs are separated from words and treated as separate tokens. Unreadable or damaged text segments are represented as ‘unreadable’. Consider the last word in Figure (3). The final letters are *lung* but the letters before *lung* are not unambiguously readable. We mark this by an underscore (here: *_lung*).

Fig. 3 Margin, Alchymistische Practic (1603)
Digeſtio im Roſmiſt oder_lung.
‘Digestion in horse-manure or … [?]’


Digeſtio im
Roſmiſt oder
[illegible]lung.

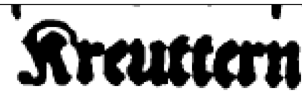
Since the insertion of margin or footnote text would prevent syntactic annotation of running sentences, margin texts are inserted before the paragraph containing them, whereas footnote text is inserted at the end of the paragraph, and their nature as notes and marginalia is annotated. The actual position of a footnote within the text is annotated on a layer called *ref* and referenced parallel to the *note* annotation on a layer called *xml_id*. Scripts and text characteristics, margins, and footnotes are marked in the graphical annotation, see Section (3.3). With a transcription of this kind, a more intuitive, visual access to the original historical text is provided (cf. Bartsch et al. 2011 for a similar approach). Such an approach is convenient for the implementation of a visualization in HTML, in applications such as ANNIS or in frameworks such as TEI[18] and allows for easy close reading of the text. To sum up, the transcription avoids a deep linguistic interpretation as far as possible and focuses on surface information, preserving most aspects of manuscript layout.

3.2 Normalization

The spelling variation in historical documents is significant and to some extent unpredictable. The variance is even higher in a diachronic corpus (see Figure 4, for some of the variants we find in RIDGES for dative plural of *Kraut* ’herb’). For this reason, we need normalization in addition to the diplomatic layer. Normalized layers help us in (a) finding instances of ‘the same’ word, (b) making generalizations, and (c) enabling linguistic processing.

[18] TEI stands for *Text Encoding Initiative*, for an introduction see Romary (2009) and Section (3.3). http://www.tei-c.org. Accessed 10 May 2016.

Fig. 4 Spelling variation of the word form *Kräutern* (‘herb’, dative plural), RIDGES Corpus 4.1

Extract of the facsimile	**dipl**	**norm**	**year**	**historical document**
	Kraͤutern	Kräutern	1603	Alchymistische Practic
	Kreutern	Kräutern	1603	Alchymistische Practic
	Kreuttern	Kräutern	1603	Alchymistische Practic
	Kraͤuteren	Kräutern	1639	Pflantz-Gart

There can, in principle, be an infinite number of normalization layers for a given text. The question of what counts as ‘the same’ depends crucially on the research question. The standoff architecture we use allows for the insertion of as many normalization layers as needed, cf. Section (2.2). The current version of RIDGES has two normalized layers, called *clean* and *norm.*

The *clean* layer is generated automatically and requires only limited linguistic analysis and is built in the following way: The first normalization step in *clean* reduces some of the variation in an automatic and simple way according to the Modern German standard. All special characters used in historical German texts, e.g., ‘ſ’ (s) and ‘⸗’ (=), are automatically replaced with their modern equivalents.[19] The different character realizations for the German umlauts (*ä* and *aͤ*, *ü* and *uͤ*, *ö* and *oͤ*) are normalized uniformly to *ä, ö, ü*. Hyphenated words at line breaks are combined to one word form; e.g. a span over two or more tokens, see *Blaͤt⸗* and *lein*, cf. Figure (5).

Fig. 5 Normalizations, visualization in ANNIS of the Example (1a)[20]

dipl	aus	vielen	kleinen	Blaͤt⸗	lein	zuſammen	geſetzet	waͤren	/
clean	aus	vielen	kleinen	Blätlein		zusammen	gesetzet	wären	/
norm	aus	vielen	kleinen	Blättlein		zusammengesetzt		wären	/

Unreadable or otherwise uninterpretable text which is marked with an underscore in the *dipl* layer is marked as *unknown* in the *clean* layer, as shown in Figure (6).

Fig. 6 Uninterpretable Text, visualization in ANNIS of the example given in Figure (3)[21]

[19] See Voigt (2013) for guidelines, http://korpling.german.hu-berlin.de/ridges/download/v4/cleanV2README.txt. Accessed 1 March 2016.
[20] Match reference link: https://korpling.org/annis3/?id=2c8e7f8d-37d3-4d98-b251-ed02c8793e9c. Accessed 16 March 2016.
[21] Match reference link: https://korpling.org/annis3/?id=b5726999-a612-4605-a023-d90c6d615812. Accessed 16 March 2016.

dipl	Digeſtio	im	Roſmiſt	oder	_lung	.
clean	Digestio	im	Rosmist	oder	unknown	.
norm	Digestio	im	Rossmist	oder	unknown	.

Thus, *clean* can be interpreted both as an annotation on *dipl* as well as an independent segmentation. The *clean* layer is a robust and simple form of normalization because it affects the text primarily on a graphic and character level. In this way, it is predictable from *dipl*. However, the merging of word forms which are separated due to line breaks requires some interpretation (for example Figures 7 and 8a).

However, the normalization in *clean* is not sufficient to find all the different spellings of the 'same word', such as *Kräutern*, *Krauttern* and *Kraͤutteren* for *Kräutern* in Figure (4). Different capitalizations and double consonants such as *tt* as well as variants such as *eu* or *äu* or *aͤu* for /ɔɪ/ are not standardized and the potential types cannot be anticipated easily. It is therefore useful to have another, more abstract, annotation layer which we called *norm*, which maps these different forms to one form.[22] As stated above, the decisions concerning abstraction depend on the research question. One possible way of designing this normalized layer could be to map all possible spellings to a form from the language stage in question – the forms in a text from 1487 would then be mapped to a single historic word form. In a diachronic corpus such as RIDGES one can be even more abstract and map all word forms to a modern word form – the forms in a text from 1487 would then be mapped to Modern German word forms according to the standard Duden lexicon (Dudenredaktion 2016).[23]

Consider the different spellings of *Krankheit* 'illness' in Figure (7). The *dipl* layer represents the original spelling. As the clean layer operates automatically and does not impose any linguistic decisions, macrons (*ā*), which are used for either *an* or *am* in Early Modern German, are dissolved into both possible interpretations *kramckhait|kranckhait*, and *kranck* and *haít* are not combined because the original does not contain a hyphen. On *norm* all forms are mapped to the modern form (token annotation for the last two examples, span annotation for the first example). The mapping of historical spellings to modern word forms is by no means always unproblematic, and requires interpretation and linguistic decisions (Gévaudan 2002).[24]

Fig. 7 Examples for normalization of *Krankheit* ('illness'), selected from Gart der Gesundheit (1487)

dipl	*kranck*	*haít*	*krāckhaít*	*Kranckheit*
clean	*kranck*	*haít*	*kramckhait|kranckhait*	*Kranckheit*
norm	*Krankheit*		*Krankheit*	*Krankheit*

Depending on its syntactic use, the word form *dz* in Figures (8a) and (8b), underscored in the captions below, can be mapped onto the Modern German complementizer *dass* 'that' (8a) or the definite article *das* 'the' (8b).[25] The mapping thus needs a syntactic analysis. Another case in point is word formation. The spelling of compounds differs even for

[22] Note that there is a different way of dealing with the search problem, namely the mapping of different forms in the search itself, also known as fuzzy search. For further references on automatic normalization see Section (3.5).
[23] Duden (Dudenredaktion 2016) is the standard orthographic lexicon for German. Many other historical corpora follow modern reference lexicons in their normalization, cf. e.g. Rissanen (2012) and Donhauser (2015).
[24] Another problem of this approach is a conceptual one: Is it useful to map forms of one language to forms (and ultimately categories) of another language? Which interesting distinctions and properties are lost? This issue (similar to the debate about the comparative fallacy in second language acquisition research, see Bley-Vorman 1983) is interesting and needs to be discussed further.
[25] The text also contains the form *das* in both interpretations. The choice between *das* and *dz* seems to be driven by typographic needs. It seems that the correct alignment within the print space plays an important role for the early printers and that (at least sometimes) the choice of the shorter/longer form is driven by the need for less/more space rather than by linguistic considerations.

the same word and in the same text, and often it is unclear whether a word is a genitive form, a compound or a complex syntactic phrase (Perlitz 2014). Case and gender inflection are not normalized to Modern German forms in order to facilitate studies of the underlying synchronic morphology in each language stage.

Fig. 8a Normalization of a complementizer, visualization in ANNIS, Gart der Gesundheit (1487) [26]
der ge ſtalt / allaín dz beyfůſz braítere ble ter hat
'... of the form, only that Beifuss has broader leaves...'

dipl	der	ge	ſtalt	/	allaín	dz	beyfůſz	braítere	ble	ter	hat
clean	der	ge	stalt	/	allain	dz	beyfusz	braitere	ble	ter	hat
norm	der	Gestalt		/	allein	dass	Beifuß	breitere	Blätter		hat

Fig. 8b Normalization of a definite article, visualization in ANNIS, Gart der Gesundheit (1487) [27]
bluͤtend machē vn̄ darauff dz bulfer legen
'... make blossom and then lay the powder.'

dipl	bluͤtend	machē	vn̄	darauff	dz	bulfer	legen
clean	blütend	machem|machen	vnn	darauff	dz	bulfer	legen
norm	blütend	machen	und	darauf	das	Pulver	legen

If historical words such as *geheb* 'tight' are not represented in the Duden (Dudenredaktion 2016), we normalize only the spelling to conform to the modern orthography. Furthermore, we don't normalize old-fashioned or extinct word forms with respect to lexical or semantical change, cf. Figure (9). The historical phrase *Fuͤr boͤſe blattern* 'for bad pocks' is automatically normalized in the *clean* layer, where special characters are replaced with their modern equivalents without the consideration of lexical or semantic language change. On the *norm* layer the historical word form *blattern* 'pocks' is capitalized because it is a noun. The normalization does not cover that there are modern equivalents of *Blattern*, for instance *Pusteln* or *Pocken*.

Fig. 9 Normalization of historical word forms, visualization in ANNIS, Arznei der Kreutter (1532) [28]
Fuͤr boͤſe blattern.
'For bad pocks.'

dipl	Fuͤr	boͤſe	blattern	.
clean	Für	böse	blattern	.
norm	Für	böse	Blattern	.

The same goes for functional items: The change in meaning of the conjunction *wann* which first has a causal meaning ('because') and later a temporal meaning ('when') does not have an effect on the *norm* layer. Being aware of this problem in general, we normalize as described above and we annotate these phenomena in an additional layer called *erlaeuterung* ('explanation') which is published together with the next version 5.0 of RIDGES Herbology in 2016.

[26] Match reference link: https://korpling.org/annis3/?id=bfee80d4-e530-460f-b70c-e6993b979646. Accessed 16 March 2016.
[27] Match reference link: https://korpling.org/annis3/?id=652c8104-bf6d-4fcd-a9ab-66ddf5a414b0. Accessed 23 March 2016.
[28] Match reference link: https://korpling.org/annis3/?id=78a5b71a-9b9a-49dc-a49e-7d5a4efad0e3. Accessed 16 March 2016.

Thus, the word form *wann* will get the explanation *denn, weil* 'because', and the word form *Blattern* will get the explanation *Pusteln* (cf. for further discussion Gévaudan 2002; Klein 2013).

3.3 Graphical and Structural Annotation

This section gives an overview of the annotation layers that describe the graphical and structural properties of the text. By now we can make use of three different segmentations (*dipl*, *clean*, *norm*), a concept from which we will draw several advantages concerning our research questions (see Section 4). All graphical and structural annotations are based on *dipl* and assigned as spans, because they reflect the original layout and may cover multiple tokens. Linguistic annotations are mostly based on *norm* (see Section 3.4).

The TEI framework provides crucial insights into text transcribing methodology (TEI Consortium 2015). TEI provides an extensive set of markup for the structural classifications of texts with the aim of describing textual layout positions.[29] Many projects use the TEI Guidelines to create digital critical editions which focus on the exact diplomatic markup of historical texts.[30] In contrast to critical editions, the RIDGES project uses only a few elements representing markup information, which are essential for linguistic analysis. In order to distinguish the running text from other textual elements in RIDGES, <head>, <note> (for footnotes) and <margin> (for marginal texts) have to be annotated. A transcription may cover line breaks and their markers (e.g., hyphens), which affects further annotations. We borrow the semantics for the conceptual annotations of these layers from TEI elements such as <lb>, <head> and @rend attributes, and implement them in our span annotations.

In Figure (10), *lb* (linebreak) reflects the original text form and allows to discriminate between hyphenation due to the end of the line on the one hand and hyphenated compound spelling on the other hand. The *lb* annotation span extends from the point at which a line begins and runs to the linebreak itself (in TEI XML, only the position of the line break is marked with a unary element,
). Without *lb*, we would have no heuristic to merge *Blaͤtlein* 'little leaf' on the *clean* layer. Having merged *Blät-* and *lein* to *Blätlein* in *clean*, the second normalization can easily be applied in the *norm* layer, cf. Figure (5). In this case, the *norm* segmentation interacts with the *lb* annotation in that it spans a *lb* boundary. The structural annotations *head* and *note* allow for specific decisions during a linguistic analysis. For example, one may decide to only include the continuous text and exclude the textual material in head, margin or footnote areas, because they may behave differently. A research question on marginals can then easily query only marginals. The different scripts (e.g. roman, blackletter) are annotated in a separate layer instead of transcribing them in *dipl* (varieties like roman and blackletter letters are not represented as distinct Unicode symbols).

Fig. 10 Annotation of line and page breaks, visualization in ANNIS of Example (1a)[31]

dipl	aus	vielen	kleinen	Blaͤt⸗	lein	zuſammen	geſetzt	waͤren	/	und	wie	die
lb	lb				lb							
pb	pb											

3.4 Linguistic Annotation

Many research questions require linguistic categorization of the data. In this section, we will describe just two areas that have been annotated in RIDGES. Further annotation layers can be added at any point in time. The first area

[29] There is also considerable work within the framework of the TEI relating to normalization and tokenization, as well as suggestions for multi-layer standoff approaches within the standard (see Heiden 2010; Pose et al. 2014).
[30] There are, among many others, Deutsches Textarchiv http://www.deutschestextarchiv.de/ (Geyken et al. 2012), the Duisburg-Leipzig Korpus romanischer Zeitungssprachen http://home.uni-leipzig.de/burr/CorpusLing/Korpusanalyse/default.htm (Burr et al. 2015), and Coptic Scriptorium http://copticscriptorium.org/, see Zeldes and Schroeder (2015). Accessed 1 March 2016.
[31] Match reference link: https://korpling.org/annis3/?id=d36c6622-9844-4b02-8f13-3699d6561e20. Accessed 23 March 2016.

concerns part-of-speech assignment and lemmatization, and is done automatically using the *norm* layer as input. The second area concerns the development of compounding and has been analyzed manually.

The *dipl* layer contains too much unpredictable variation for automatical part-of-speech tagging, cf. Figure (4). Having normalized the spelling variation of all word forms to Modern German forms (e.g. *Kraut*, *Kräuter*, *Kräutern*) in the *norm* layer, it is possible to automatically assign a lemma to a form, e.g. *Kraut*, see Figure (11). The tagging and lemmatization is done by the TreeTagger (Schmid 1994, using the STTS tagset of Schiller et al. 1999), which is trained on Modern German. In RIDGES Herbology Version 4.1, we checked the *pos* and *lemma* layer with the help of DECCA (Dickinson and Meurers 2003).

In Figure (11), part-of-speech does not change, regardless of the spelling, as the linguistic category *noun* (NN) remains the same. This is true even for cases, which we introduce above, where the transcription *dipl* is segmented into two tokens, cf. Figure (7) above. There, *Krank* and *heit* are normalized as one normalized token, and are thus given only one part-of-speech tag. Depending on its segmentation, the normalization layer may therefore influence the allocated *pos* categories.

Fig. 11 Example of the uniform linguistic annotations for a variety of historical word forms of the noun *Kraut*, selected from RIDGES Corpus 4.1

dipl	*Kraͤutern*	*Kraut*	*kraut*	*Kreuttern*	*Kreutter*	*kreüter*	*Kraͤuteren*	*Kreuter*	*Kräuter*
clean	*Kräutern*	*Kraut*	*kraut*	*Kreuttern*	*Kreutter*	*kreüter*	*Kräuteren*	*Kreuter*	*Kräuter*
norm	*Kräutern*	*Kraut*	*Kraut*	*Kräutern*	*Kräuter*	*Kräuter*	*Kräutern*	*Kräuter*	*Kräuter*
pos	NN	NN	NN	NN	NN	NN	NN	NN	NN
lemma	*Kraut*	*Kraut*	*Kraut*	*Kraut*	*Kraut*	*Kraut*	*Kraut*	*Kraut*	*Kraut*

In Figure (12), the to-infinitive *zubekommen* 'to receive' is transcribed as one token, but split up in the normalization. Thus, the split-up segments can be annotated separately on the *pos* layer, and can now be found with queries for all infinitives (VVINF) or the infinitive particle *zu* 'to' (PTKZU), cf. Figure (12).

Fig. 12 Split-up normalization and annotation, visualization in ANNIS, *Pflantz-Gart Capitel 4* (1639) [32]
Den Winter⸗ſpinet ſehr groſz zubekommen /
'To let grow very tall the winter spinach'

dipl	Den	Winter⸗ſpinet	ſehr	groſz	zubekommen		/
clean	Den	Winter-spinet	sehr	grosz	zubekommen		/
norm	Den	Winterspinat	sehr	groß	zu	bekommen	/
pos	ART	NN	ADV	ADJD	PTKZU	VVINF	$(
lemma	d	Winterspinat	sehr	groß	zu	bekommen	/

Since the TreeTagger is initially trained on Modern German newspaper texts and uses a fixed lexicon for lemmatization, there are a few performance issues. To evaluate the TreeTagger performance as well as its semi-automatic correction, we drew a sample of 1560 tokens and manually corrected the *pos* layer. We chose approx. 300 tokens for each century in the corpus. Using the manually corrected sample as a baseline, the TreeTagger shows a mean document accuracy of 93.80% with a standard deviation of 8.1%. Fairly similarly, the *pos* layer corrected with DECCA shows a mean document accuracy of 93.78% with a standard deviation of 5.6%. Of course this should neither be regarded as a detailed evaluation of the TreeTagger, nor of the DECCA method. In our sample, the pos tags ADV, FM and XY (for DECCA) and XY, FM and VVFIN (for the TreeTagger) are the most frequently corrected types. This

[32] Match reference link: https://korpling.org/annis3/?id=71b137b9-2a09-4dca-8d13-1a4998ac19d1 . Accessed 22 March 2016.

sample only provides a first impression of the TreeTagger performance using the STTS, which in turn has its own limitations with historical data.

Concerning the lemmatization with the help of the TreeTagger, some register-specific compounds such as *Kelchblätter* ‘sepals’, *Staubfäden* ‘filaments’ oder *Blumendecke* ‘flowerbed’ are not listed in the lexicon. Their lemmas are given as <unknown> but in many contexts they are tagged correctly with the part-of-speech category NN for common noun. The same tendency holds for verbs such as *destillieren* ‘distil’ and for adjectives such as *einblättrige* ‘one-leaved’ or *blattartige* ‘leaf-like’. We will further discuss this in Section (4.4), and we recently started to address this issue by evaluating and training NLP tools on our data, see Section (5).

A class of words with varying orthography as single or multiple tokens is found in the case of compounds. The development of compounding in German has been discussed in terms of a competition between lexicalized phrasal constructions and compositional syntactic constructions (cf. Paul 1995; Lindauer 1995; Splett 2000). Perlitz (2014) investigates the distribution of noun compounds and their phrasal equivalents in the scientific register of German in RIDGES, searching for connections between decisions of split and joint orthography and morphological forms consistent or inconsistent with a genitive attribute reading. For example, a form such as *Bauchflüsse* ‘stomach flows’ cannot represent a genitive attribute and head, since the genitive form of *Bauch* ‘stomach’ would require an *-s*: *Bauchs*. However, for a form such as *Teufelswurzel* (literally: devil's root, ‘hyoscyamus’, ‘devilsroot’) it is difficult to determine whether the *-s* represents a genitive or a compound linking element in its period, and much spelling variation is found (for a detailed discussion see Perlitz 2014). The annotation of the different spelling types of Perlitz (2014) has been integrated into the corpus (the layer *komp_orth*) for compounds (*k*) and syntactic genitive attributes (*attr_gen*), both based on the *norm* layer, see Figure (13).

Fig. 13 Annotation of compounds, visualization in ANNIS, Die Eigenschaften aller Heilpflanzen (1828) [33]
und andere Bauchfluͤſſe , das Naſenbluten und Erbrechen,
‘and other stomach flows, the nosebleeds and vomiting, ...’

norm	und	andere	Bauchflüsse	,	das	Nasenbluten
pos	KON	ADJA	NN	$,	ART	NN
lemma	und	ander	<unknown>	,	d	Nasenbluten
komp			k			k
komp_orth			zs			zs
prot			prot1			prot1

3.5 Related Work and Discussion

There are several corpora and corpus projects which deal with historical texts similar to the RIDGES corpus but focus on other research questions, goals, the sampling of registers and language periods. After having presented the architecture and pre-processing decisions we took for RIDGES, we want to briefly discuss some of these approaches to the construction of historical and diachronic corpora and further illustrate the advantages of a multi-layer architecture.

Many historical corpora only have one textual (or primary) layer on which annotations are based. In times before Unicode, the textual layer often could not or did not represent a diplomatic transcription. The decisions about normalization were built into the textual layer, which contained the normalized form as part of the running tokens representing the base text (one well-known and influential example is the Helsinki Corpus of Old English Texts[34]).

[33] Match reference link: https://korpling.org/annis3/?id=aa0086df-b15b-4447-817b-f00c63a2950a. Accessed 23 March 2016.
[34] http://www.helsinki.fi/varieng/CoRD/corpora/HelsinkiCorpus/. Accessed 1 March 2016.

Even in more richly annotated corpora, such as historical treebanks as pioneered by the Penn Parsed Corpora of Middle and Early Modern English (PPCME and PPCEME, see Kroch and Taylor 2000; Kroch et al. 2004), which also contain syntax tree annotations, limitations imposed by annotation formats meant that only one representation of the raw text can be used. Figure (14) illustrates the format:

Fig. 14 Fragment from the Penn Parsed Corpus of Early Modern English for 'The 5th of Feb. 1695'

```
( (NP-TMP (D Y=e=)
          (ADJ 5=th=)
          (PP (P of)
              (NP (NPR Feb.)))
          (, ,)
          (NUM $1695)
          (. .)))
```

The brackets in the Penn Treebank format express the syntactic phrases, the string at the left bracket is the syntactic category or part-of-speech, and the string at the right bracket is the actual token, cf. Figure (14). Typographical properties such as superscripts are expressed with '=' signs (see Kytö 1996), while letters such as the old Thorn represented as a capital Y (the abbreviation Y with superimposed superscript *e* standing for 'the') cannot be encoded in any special way. Formats such as TEI XML allow more verbose representation of rendering using tags such as <hi rend="...">, as well as <choice> tags to express alternate spellings or normalization. Using fully automatic normalization is an option to populate such tags, though usually the level of quality desired in a historical corpus for scholarly purposes will require semi-automatic methods (see Baron and Rayson 2008; Craig and Whipp 2010; Reynaert et al. 2012). The following encoding is a possible TEI rendition for the example above:

Fig. 15 Text encoding with TEI XML rendition

```
<w>
	<choice>
		<reg>The</reg>
		<orig>Þ<hi rend="superscript">e</hi></orig>
	</choice>
</w>
<w>
	<choice>
		<reg>fifth</reg>
		<orig>5<hi rend="superscript">th</hi></orig>
	</choice>
</w>
```

In Figure (15), encoding the Thorn as a thorn and not as a capital *Y* in the original is in itself a linguistic interpretation (this could be spelled *Ye* even in Modern English, as in intentionally archaic *Ye Olde Shoppe*).

Some more recent corpora have combined syntactic analysis, such as that found in PPCEME, with orthographic annotation within the framework of the TEI. For example, Höder (2012) describes HaCOSSA, the Hamburg Corpus of Old Swedish with Syntactic Annotations, which uses the TEI's clause, phrase and word elements (<cl>, <phr> and <w>) to build syntax trees, while representing orthographic contractions with <dipl> together with corrections, expanded abbreviations and continuations supplied by the editor (<corr>, <ex> and <supplied> elements), see Figure (16).

Fig. 16 Two examples of inline XML syntax and diplomatic annotation with contraction extension in HaCOSSA: ‘Where your treasure is, there your heart is...’ and ‘The Lord Jesus Christ’.

```
<cl>
        <cl>
                <w><dipl>hwar</dipl></w>
                <phr>
                        <w><dipl>tith</dipl></w>
                        <w><dipl>ligghiande</dipl></w>
                        <w><dipl>fææ</dipl></w>
                </phr>
                <w><dipl>ær</dipl></w>
        </cl>
        <punct><dipl>/</dipl></punct>
        <w><dipl>ther</dipl></w>
        <w><dipl>ær</dipl></w>
        <phr>
                <w><dipl>tith</dipl></w>
                <w><dipl>hyærta</dipl></w>
        </phr>
</cl>

<w><dipl>härra<ex>n</ex></dipl></w>
<w><dipl>Jhe<corr>sus</corr></dipl></w>
<w><dipl><supplied>Christ</supplied>us</dipl></w>
```

Although these annotations go a long way beyond what was possible when the Penn corpora were produced, the possibility of representing conflicting tokenizations is still not supported, as this would violate the XML hierarchy for <w> elements. A second example for a corpus architecture, which does not contain conflicting hierarchies is the Bonner Frühneuhochdeutschkorpus (Diel et al. 2002), which contains Early New High German text abstracts of approx. 400 words coming from different dialects and registers, such as private letters, official documents or grammars. The corpus was developed to investigate the inflectional morphology of Early New High German. The corpus architecture was built on a proprietary XML scheme with the help of a DTD which covers part-of-speech annotations and lemmatization, among other things.

Historical corpora with inline annotations, with or without XML tags, can be enormously useful for linguistic analysis but make cross-layer analyses of typographic and (to some extent) spelling properties difficult when these are cross-referenced with linguistic annotation. Even in corpora encoded in Unicode and using multi-layer architectures, we find that linguistic decisions strongly influence how the primary textual layers are interpreted. An example is the Tatian Corpus of Deviating Examples (T-Codex, version 2.1, Petrova et al. 2009)[35], which uses, among others, the ‘+’ to mark clitic constructions in Old High German, such as *n+ ist* (‘not+ is’) within the primary layer, thus mixing a diplomatic transcription and a linguistic analysis. At the same time, highly diplomatic editions of texts are sometimes built which do not allow for the inclusion of normalization, and these subsequently prevent linguistic searches, since users cannot predict all variant forms. It becomes clear that a corpus may contain several concepts of what a ‘text’ might be. A ‘text’ might be an annotation (e.g., *clean* or *dipl* in RIDGES) and at the same time an independent normalization concept above which further annotations might be applied. The RIDGES architecture allows as many ‘primary’ or ‘textual’ layers as are required for a given analysis: we can analyze a word as a clitic in its normalized realization, but as an independent linguistic unit when annotated above a diplomatic transcription layer. In this way

[35] Petrova, Svetlana; Donhauser, Karin; Odebrecht, Carolin; T-Codex (Version 2.1), Humboldt-Universität zu Berlin. https://korpling.german.hu-berlin.de/~annis/T-CODEX/corpus_description_tatian2.1.pdf, http://hdl.handle.net/11022/0000-0000-850C-D. Accessed 21 March 2016.

there is no loss of information and all layers can be used for the analysis, as envisioned by corpus creators. The corpus can be used for careful typographic studies as well as for abstract syntactic analyses, which are not intertwined with each other.

There have been several attempts for (semi-)automatic corpus processing which focus on the development and training of taggers and parsers for historical data. The German Manchester Corpus (GerManC, Durell et al. 2007)[36], for example, contains a wide range of text genres such as sermons, personal letters, drama, narrative prose and academic texts. The corpus project focuses on the training of tools which then automatically annotate these text samples, for example the normalization layer (Jurish 2010), the part-of-speech and lemmatization layers (Schmid 1994), as well as morphological tagging and dependency annotations (Bohnet 2010).

As far as we know, there are no freely available dictionaries for automatic normalization for the register and language period of the RIDGES corpus. Statistically learned rules for normalization have not worked well so far either, as the corpus is too small for statistical training as applied e.g. by Jurish (2010), Bollman et al. (2011, 2012), or Archer et al. (2015), for an overview see Piotrowski (2012). A key problem for a diachronic corpus is that orthography is changing across periods, and each text would require its own normalization rules. When turning to manual or semi-automatic normalization, different theoretical perspectives are argued for in the literature (cf. Baron and Rayson 2008; Pilz 2009; Ernst-Gerlach 2013). Rules for replacements may be applied for *ſ* and umlauts, but tend to get too complex when replacing unforeseeable spelling variations such as in Figure (4) for *Kraut* (herb). Instead, similar to the *clean* layer, the *norm* layer in RIDGES is based on the surface and graphematic characteristics of the modern target language, in order to facilitate searchability for users. In our view, a normalized layer of this nature is essential for any diachronic corpus to be accessible and the more so if a comparison to contemporary phases of the language with standardized orthography is planned.
Many recent corpus projects dealing with historical German are similar to RIDGES in so far that they all use multi-layer corpus architecture; for example the Early New High German and Modern German Fürstinnenkorrespondenzkorpus[37] containing private letters of aristocratic women, and the Old High German Altdeutschkorpus[38] (Donhauser 2015) as well as the Early New High German Anselm Corpus (Dipper and Schultz-Balluff 2013) containing medieval religious treatises. Due to the different research questions, language periods and text genres these corpora use different annotation schemes, different normalization rules. Only some of these corpora uses multiple segmentations, e.g. the Anselm Corpus for normalization similar to the RIDGES Corpus.

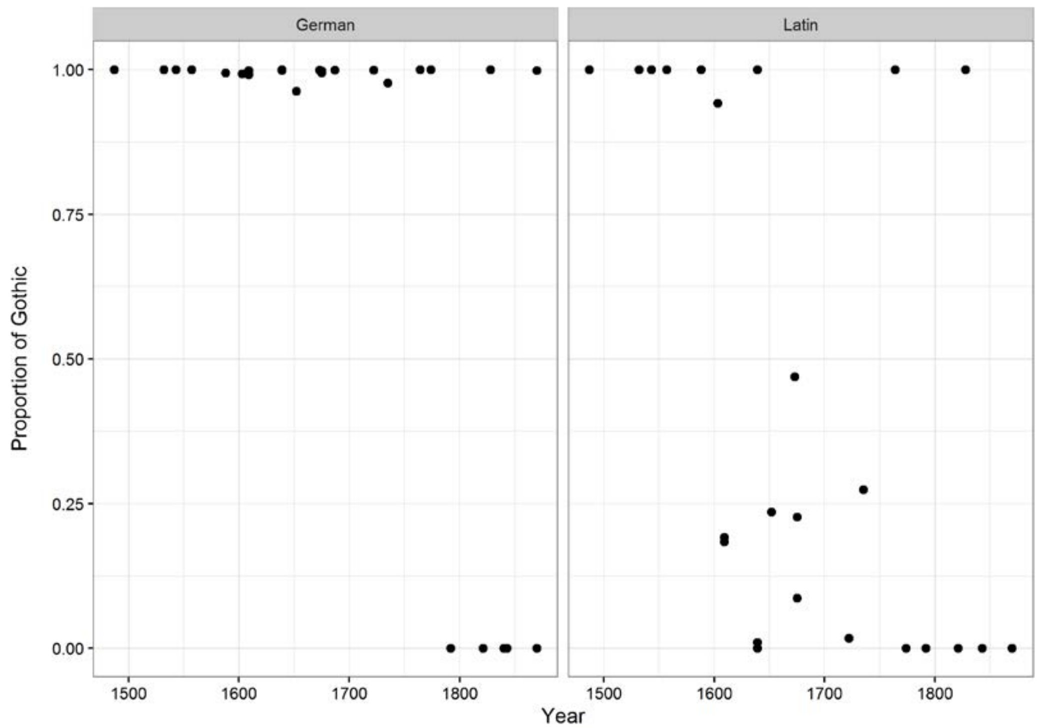
4. Case Studies
In the following section, we will briefly illustrate how the multi-layer architecture with multiple tokenizations is useful for answering research questions. We will present studies based on structural markup annotation (Section 4.1), on graphematic information (Section 4.2 and 4.3), on linguistic annotation (Section 4.4), and on register-specific annotation (Section 4.5). The case studies described here might serve as a starting point for more thorough and extensive investigations using the RIDGES corpus, as the corpus is freely available.

4.1 Scripts Depending on Language
(German) historical texts differ, among other things, with respect to their use of scripts, which typically include many fonts. The interaction between the scripts used and the language which is printed may give a first insight into the

[36] Bennett, Paul; Durrell, Martin; Ensslin, Astrid; Scheible, Silke; Whitt, Richard; GerManC (Version 1.0), University of Manchester. http://www.llc.manchester.ac.uk/research/projects/germanc/. http://hdl.handle.net/11022/0000-0000-2D1B-1. Accessed 21 March 2016.
[37] Fürstinnenkorrespondenzkorpus. Lühr, Rosemarie; Faßhauer, Vera; Prutscher, Daniela; Seidel, Henry; Fuerstinnenkorrespondenz (Version 1.1), Universität Jena, DFG. http://www.indogermanistik.uni-jena.de/Web/Projekte/Fuerstinnenkorr.htm. http://hdl.handle.net/11022/0000-0000-82A0-7. Accessed 21 March 2016.
[38] Donhauser, Karin; Gippert, Jost; Lühr, Rosemarie; ddd-ad (Version 0.1), Humboldt-Universität zu Berlin. https://referenzkorpusaltdeutsch.wordpress.com/. http://hdl.handle.net/11022/0000-0000-7FC2-7. Accessed 21 March 2016.

function and distribution of object and meta language in scientific texts. The RIDGES corpus contains both necessary types of annotation, for script and language. Both annotation concepts were inspired by the TEI Guidelines.[39] The use of the two scripts roman and blackletter is annotated based on the diplomatic transcription *dipl*. The language is annotated with the ISO 639-2 language codes, e.g. *deu* for German[40], *lat* for Latin and *eng* for English.[41] Figure (17) shows the correlation between the script distribution within a text and the two most frequent languages, namely German and Latin.[42] For German, there is a change from the predominantly used blackletter to roman, starting around 1800. Interestingly, we observe that all German sequences in a text are either printed completely in roman or blackletter script. Latin terms or descriptions seem to be marked by roman, beginning around 1600, as can be seen in the right panel of Figure (17). However, the change observed here is not categorical, but rather varies to differing degrees until 1750.

Fig. 17 Distribution of the scripts roman (tag *antiqua*) and blackletter (tag *gothic*) for German and Latin in RIDGES 4.1



4.2 Punctuation

In written Modern German, the distribution and function of punctuation is regulated in the orthography (see for example Duden 4, 1072–1073). In former stages of German, there was no binding orthographic norm for punctuation in the written language (Höchli 1981; Simmler 2003; Nerius 2007). Thus, there is variation in punctuation in addition

[39] The element <lang> http://www.tei-c.org/release/doc/tei-p5-doc/de/html/ref-lang.html and attribute xml:lang, and the element <hi> which can be attributed information about the font http://www.tei-c.org/release/doc/tei-p5-doc/de/html/ref-hi.html. Accessed 1 March 2016.

[40] Due to the ongoing history of the corpus and the evolving annotation guidelines, not all texts contain annotation for German. If a document does not have an explicit annotation *deu*, we counted each *dipl* token without any annotation in the *lang* layer as German in the post-hoc analysis.

[41] http://www.loc.gov/standards/iso639-2/php/code_list.php. Accessed 1 March 2016.

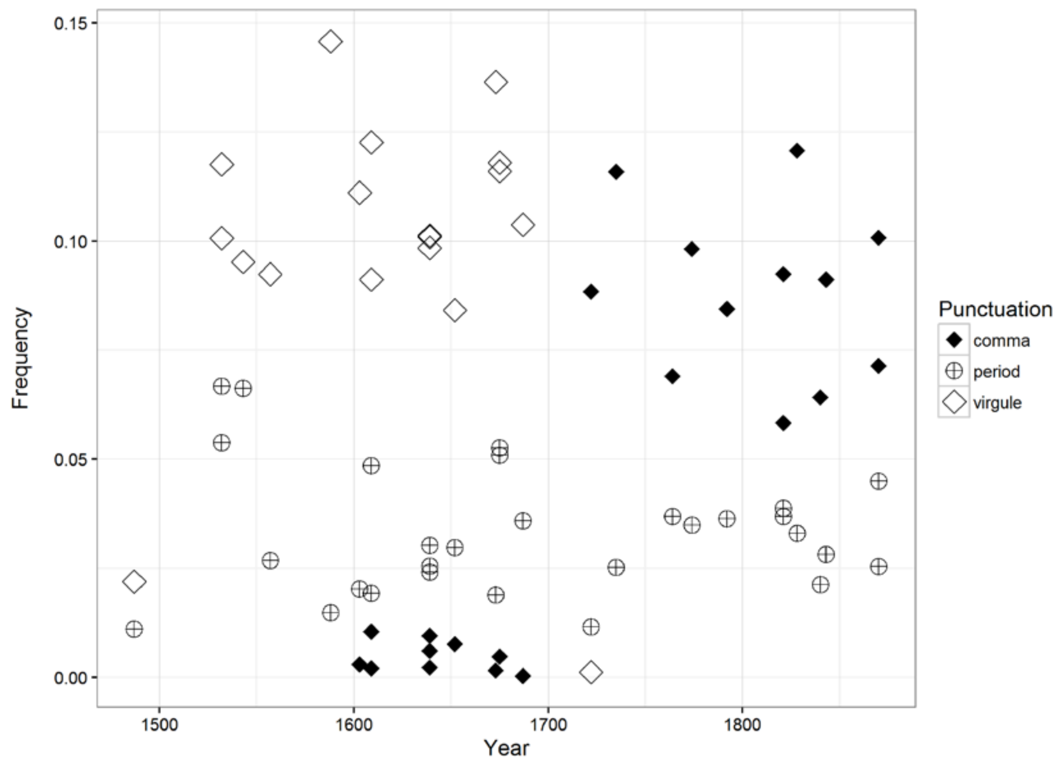
[42] Many of the texts contain Latin passages, ranging from words (often translations of the name of a herb or an illness, as in Example (1b)) to phrases and, sometimes, whole paragraphs. The texts also contain information (also often translations of the names) in other languages, such as Greek, French, or English.

to variation in the spelling of word forms (see Section 4.3). In the following case study, we investigate the distribution of the three punctuation types: period, comma and virgule (slashes) in order to gain empirical insights into their potential functions. We base the analysis on *dipl*, as all punctuation instances are already segmented during the transcription.

Figure (18) shows the distribution of period, comma and virgule for each text. The prevalent slashes or virgules used in documents before 1700 show roughly the same relative frequencies as commas after 1700. Between 1500 and 1700, only a marginal number of commas were used. The frequencies of periods do not vary much (note that this gives us no information as to their function and distribution, which might have changed considerably).

To start a first interpretational attempt, Figure (18) shows a tendency which is described and discussed as a change in the use of punctuation, or text structuring characters (Höchli 1981; Reichmann & Wegera 1993). The RIDGES corpus can provide empirical evidence: After being used only marginally over a hundred year span, commas abruptly rise in use, indicating that slash replacement has not evolved gradually, but may have been conventionalized by the writing community rapidly. Adding further annotation to the data might reveal interesting differences in the use of punctuation over the centuries.

Fig. 18 Punctuation frequencies per text in RIDGES 4.1



4.3 Spelling Variation

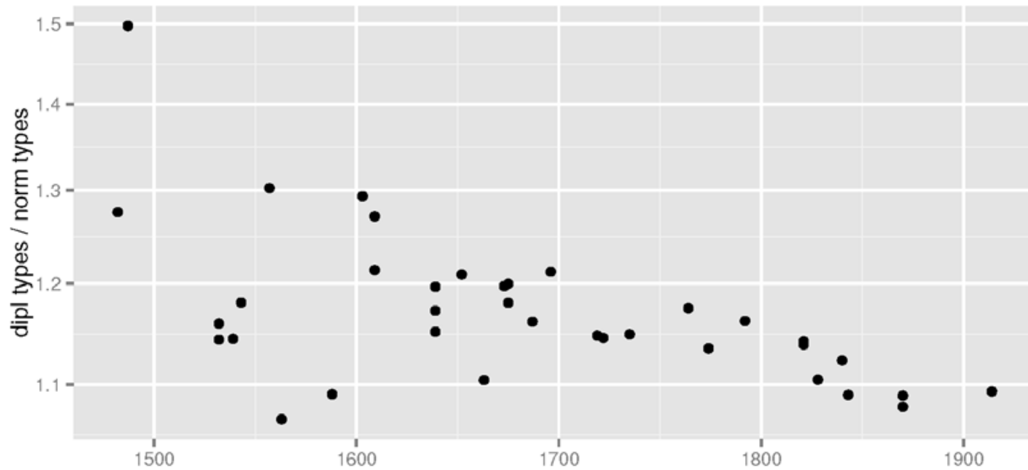
It is interesting to investigate whether standardization is only influenced by extrinsic forces or whether there is some inherent trend to reduce variation in a system, which then facilitates an extrinsic standardization of the remaining varieties (Reichmann and Wegera 1993; Besch 2003; Nerius 2003; Nerius 2007; Wolff 2009). Since the late 19th century, Modern German is highly regulated, influenced by a sequence of standardization committees[43] and decisions

[43] Currently it is the 'Rat für deutsche Rechtschreibung' http://www.rechtschreibrat.com/. Accessed 1 March 2016.

about teaching materials and standards taught in schools. There were, of course, standardization initiatives in earlier times but they were typically locally and functionally confined (due to the political and educational situation), such as, e.g., the Kanzleisprachen (the use of language in government offices, cf. Bentzinger 2000) or the influence of Luther's Bible translation (and following texts), for an overview see Nerius (2007). Thus, we would assume that the variation between the historic *dipl* and the Modern German *norm* decreases over time.

Figure (19) shows the mean of different spelling variants (*dipl*) per normalized word form (*norm*) for each document (y-axis) over time in the RIDGES Corpus (x-axis): We searched for the word forms in the *dipl* layer which are normalized to the same word form in the *norm* layer in each document in the corpus. For example, in earlier texts such as Gart der Gesundheit (1487) (cf. Figure 7), there exist several spelling variants of the lemma *Kraut* ('herb') whereas in later texts, there is only a single variant of the same lemma in a document.
We calculated the within-document spelling variation and observed this variation over time. Figure (19) shows that the spelling variance of the *dipl*-token seems in fact to decrease gradually as expected. The results are based on surface information only and do not allow conclusions about the cause of this variation without further study.

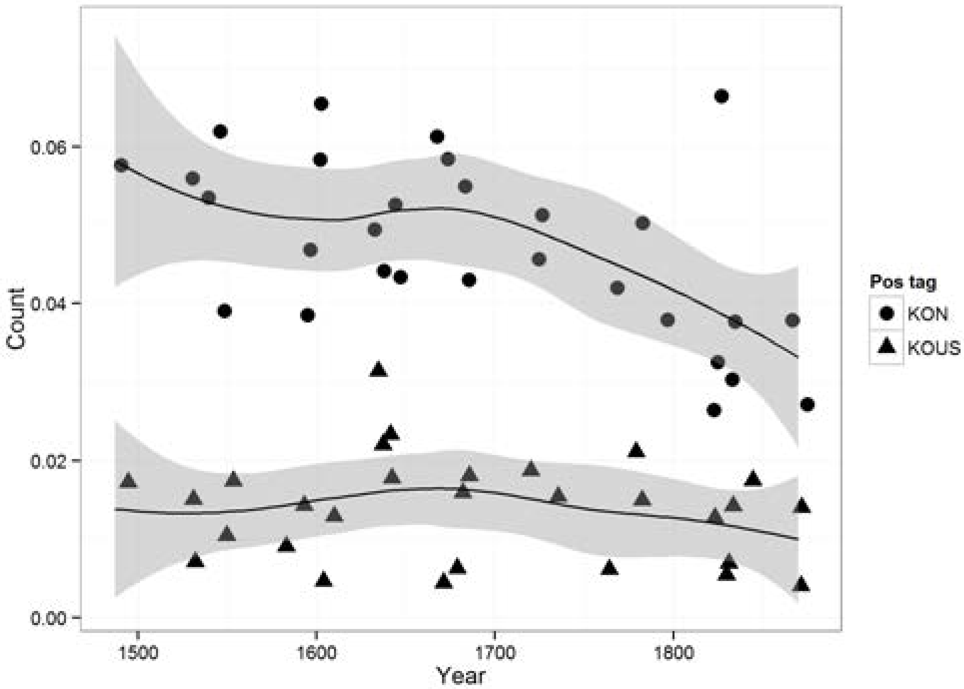
Fig. 19 Spelling variations in RIDGES 4.1 per document



4.4 Part-of-speech Variation
Text coherence and complexity is a relevant feature in the development of scientific registers (cf. Biber and Gray 2011a; Biber and Gray 2011b). Clausal coordination, therefore, might be a point of interest for research on RIDGES (cf. Admoni 1990). There are coordinating connectors and subordinating connectors (cf. Ebert 1978; Hartweg & Wegera 2005). In the STTS tagset, coordinating connectors and subordinating connectors are marked with different tags: coordinating connectors (*pos* layer, tag: KON, e.g., *und* 'and', *oder* 'or', *aber* 'but') and subordination connectors (*pos* layer tag: KOUS, e.g., *weil* 'because', *dass* 'that', *damit* 'so that', *wenn* 'if', *ob* 'whether'). As the TreeTagger used for the part-of-speech annotation in RIDGES is not trained for Early New High German, it is applied to the *norm* layer.

As Figure (20) shows, we are able to gain a quick overview of the relative frequencies of these variants. The frequency of the *pos* tag for coordination (KON) seems to decrease, while the frequency of the *pos* tag for subordination (KOUS) remains constant. A first interpretation might be that coordination structures get more and more infrequent in the emerging scientific register. In this study, we ignore other cohesive elements, e.g., adverbs, which might replace both types of coordination. Additionally, note that a more detailed analysis should look more closely at the correction of false negatives. False positives have been corrected during the semi-automatic correction of the *pos* layer, cf. Section (3.4). A further restriction for conclusions might be that KON also coordinates simple phrases like nominal or prepositional phrases, whereas KOUS tends to be used for subordinating clauses. Thus, the context needs to be considered in further research.

Fig. 20 Frequencies of the *pos* tags KON and KOUS in RIDGES 4.1



4.5 Expected Term Frequencies

The same principles of multi-layer architectures apply in the same way to research on the content of documents and not just to linguistic forms. Just as in other languages, several aspects of text organization and presentation have developed in German scientific literature over the course of time. For example, in order to study the development of technicality in scientific writing, we can look at the use of technical terms for herbs, diseases and other technical terms annotated in the corpus. If we assume that term types will be distributed independently of document dates, we can measure the deviation from this assumption in terms of observed versus expected frequencies based on the relative frequency of each term type in the whole corpus (for a short introduction to overuse and underuse see Lüdeling 2011). Figure (22) shows an association plot with rectangles representing the size of this difference (black and above the line for higher frequency than expected, grey and below for less). In RIDGES, we distinguish between three kinds of terms, herbs (*h*), diseases (*d*) and technical terms (*t*) in the *term* layer, see Figure (21):

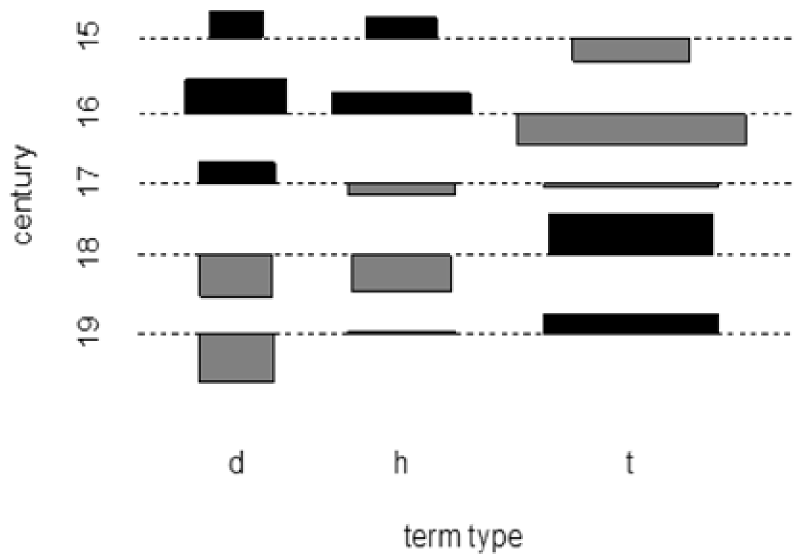
Fig. 21 Annotation examples of *term* layer; technical terms, herbs and diseases in RIDGES 4.1

Diſz nennet man einen **Kolben** /	*Kolben* (‘flask’) as a technical term (t)[44]
/ die trincke von **Camillen blumen**	*Camillen blumen* (‘chamomile flowers’) as a term for a herb (h)[45]
/ vnd wer einen **boͤſen Magen** hat	*boͤſen Magen* (‘bad stomach’) as a term for a disease (d)[46]

[44] Match reference link: https://korpling.org/annis3/?id=a220ec13-3043-4fdc-8849-ef23f3f1ad31. Accessed 23 March 2016.
[45] Match reference link: https://korpling.org/annis3/?id=efbb91f4-0e07-42fa-a23a-7fa33f2c53ac. Accessed 23 March 2016.
[46] Match reference link: https://korpling.org/annis3/?id=d0a863e9-f64f-47a0-a856-72c12650a082. Accessed 23 March 2016.

Figure (22) shows a non-linguistic, content-related fact about the corpus: there is a clear trend in the documents included to move from discussing a lot of herbs and diseases (*h* and *d* respectively) to mentioning much fewer of these compared to other technical terms. This is to do with the kinds of texts under inspection: The texts change over the time and there seems to develop a variety of texts, from medical compendiums and lists of herbs and their effects in earlier texts, to scholarly discussions developing technical terms that go beyond actual specific herbs etc.

Fig. 22 Association of term categories with centuries in RIDGES 4.1



5. Summary and Outlook

In this paper we presented the RIDGES corpus, a freely available corpus charting the development of German as a language of science. The development of a scientific register in a vernacular (language) as an alternative to Latin was a non-trivial step that had to be repeated across Europe in the Middle Ages and the Renaissance, and studies of this process cannot be carried out without corpora of this kind. Key considerations in designing such a corpus include evenly spaced out samples (in 30 year bins in our case) and maximal comparability of the domain across time (here using the relatively stable botanical domain, but of course homogeneity is always only partial).

In encoding the corpus we have learned many lessons about the natures and conflicting needs of manuscript-near diplomatic and spelling analyses versus normalized, linguistic analyses geared towards identifying content and constructions across time. We view the presence of at least one primary division of diplomatic representation and normalized representation as essential to any diachronic corpus that is geared towards (re-)usability for a variety of research questions and fields. Our work with the RIDGES data has led us to adopt a stand-off annotation model which allows the encoding of multiple, even conflicting base text layers, each possibly carrying its own annotations independently of the others. Thus, part-of-speech analysis can build on top of normalized word forms, while structural descriptions of manuscripts or prints can exist above a separate textual representation. The number and nomenclature of the annotations is not constrained, including such corpus specific layers as the annotation of terminological reference in term types across time. The case studies presented here are meant to illustrate the feasibility and utility of the multi-layer approach: all data was extracted directly from the ANNIS search engine without the need for complex scripts analyzing the structure of the annotations to derive the necessary information.

The RIDGES corpus architecture and preparation focus on manual annotation, as well as surface-oriented and consistent interpretations. An exciting avenue of research is to improve Optical Character Recognition (OCR) on older

German typefaces (Fraktur, Schwabacher etc.) to the point where manual correction becomes easy enough to increase the order of magnitude of the data (see Springmann and Lüdeling, submitted). Further on, the RIDGES corpus shows, in line with other approaches, that it is necessary to evaluate and train NLP tools to achieve a better and solid base for further analysis, at a point where the RIDGES corpus is big enough and can serve as a gold-standard. The cooperation project called LangBank[47] will start to address these issues. The analysis of the corpus is also ongoing, with some first results e.g. on compounding in the German scientific register, becoming available now (Perlitz 2014).

We believe that the architecture and design choices employed in the corpus put it in a position to be expanded on and studied for a variety of philological and linguistic research questions. The data presented here is freely available, but does not represent the final version of the RIDGES corpus: we will continue to collect data and annotate it further.

6. References

Admoni, W. (1990). *Historische Syntax des Deutschen*. Tübingen: Niemeyer.

Archer, D., Kytö, M., Baron, A., Rayson, P., et al. (2015). Guidelines for Normalising Early Modern English Corpora. Decisions and Justifications. *ICAME Journal*, 39(1), 5–24. doi: 10.1515/icame-2015-0001

Baron, A., Rayson, P., Archer, D., et al. (2009). Word Frequency and Key word Statistics in Historical Corpus Linguistics. *International Journal of English Studies*, 20(1), 41–67.

Baron, A., & Rayson, P. (2008). VARD 2: A Tool for Dealing with Spelling Variation in Historical Corpora. In *Proceedings of Postgraduate Conference in Corpus Linguistics*, PCCL 2008. Birmingham. http://acorn.aston.ac.uk/conf_proceedings.html. Accessed 4 August 2015.

Belz, M., Odebrecht, C., Perlitz, L. & Voigt, V. (2015). Annotationsrichtlinien zu Ridges Herbology Version 4.1, Humboldt-Universität zu Berlin. http://korpling.german.hu-berlin.de/ridges/download/pubs/annotationGuidelines_v4.1.pdf. Accessed 16 March 2016.

Bartsch, N., Dipper, S., Herbers, b., Kwekkeboom, S., Wegera, K.-P., Eschke, L., Klein, T., & Weber, E. (2011). Annotiertes Referenzkorpus Mittelhochdeutsch (1050–1350). In *33. Jahrestagung der Deutschen Gesellschaft für Sprachwissenschaft*, DGfS-CL Poster Session 2011. Göttingen.

Bentzinger, R. (2000). Die Kanzleisprachen. In W. Besch, A. Betten, O. Reichmann, & S. Sonderegger (Eds.), *Sprachgeschichte. Ein Handbuch zur Geschichte der deutschen Sprache und ihrer Erforschung* (pp. 1665–1673). Vol. 2, Second Edition. Berlin i.a.: de Gruyter.

Besch, W. (2003). Die Entstehung und Ausformung der neuhochdeutschen Schriftsprache/Standardsprache. In W. Besch, A. Betten, O. Reichmann, & S. Sonderegger (Eds.), *Sprachgeschichte. Ein Handbuch zur Geschichte der deutschen Sprache und ihrer Erforschung* (pp. 2252–2296). Vol. 3, Second Edition. Berlin i.a.: de Gruyter.

Biber, D., & Gray, B. (2011a). Grammar emerging in the Noun Phrase: The Influence of Written Language use. *English Language and Linguistics*, (15), 223–250. doi: 10.1017/S1360674311000025

Biber, D., & Gray. B. (2011b). The Historical Shift of Scientific Academic Prose in English towards less explicit Styles of Expression: Writing without Verbs. In V. Bathia, P. Sánchez, & P. Perez-Paredes (Eds.), *Researching Specialized Languages* (pp. 11–24). Amsterdam: John Benjamins. doi: 10.1075/scl.47

Biber, D., & Conrad, S. (2009). *Register, Genre, and Style*. Cambridge: Cambridge University Press.

Bley-Vroman, R. (1983). The Comparative Fallacy in Interlanguage Studies: The Case of Systematicity. *Language Learning*, 33(1), 1–17.

Bird, S. & Liberman, M. (2001). A formal framework for linguistic annotation. *Speech Communication*, 33(1–2), 23–60. doi: 10.1016/S0167-6393(00)00068-6

Bohnet, B. (2010). Top Accuracy and Fast Dependency Parsing is not a Contradiction. In *Proceedings of the 23rd International Conference on Computational Linguistics* (pp. 89–97), Coling 2010. Beijing.

Bollmann, M., Dipper, S., Krasselt, J., & Petran, F. (2012). Manual and Semi-automatic Normalization of Historical Spelling — Case Studies from Early New High German. In *Proceedings of the First International Workshop on Language Technology for Historical Text(s)*, KONVENS 2011. Wien.

[47] http://sfs.uni-tuebingen.de/langbank/de/index.html Accessed 16 March 2016.

Bollmann, M., Petran, F., & Dipper, S. (2011). Applying Rule-Based Normalization to Different Types of Historical Texts: An Evaluation. In *Proceedings of the 5th Language and Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics*, TLC 2011. Poznan.

Burr, E., Burkhardt, J., Potapenko, E., Sierig, R., & Concepción Durán, A. (2015). Das Duisburg-Leipzig Korpus romanischer Zeitungssprachen und sein Textmodell. In *Proceedings Von Daten zu Erkenntnissen. 2. Jahrestagung des Verbandes der Digital Humanities im deutschsprachigen Raum*, DHd 2015, Graz. http://gams.uni-graz.at/o:dhd2015.abstracts-gesamt. Accessed 22 march 2016.

Carletta, J., Evert, S., Heid, U., Kilgour, J., Robertson, J., Voormann., H., et al. (2003). The NITE XML Toolkit: Flexible annotation for multi-modal language data. *Behavior Research Methods, Instruments, and Computers*, 35(3), 353–363. doi: 10.3758/BF03195511

Chiarcos, C., Dipper, S., Götze, M., Leser, U., Lüdeling, A., Ritz, J. & Stede, M. (2008). A Flexible Framework for Integrating Annotations from Different Tools and Tag Sets. *Traitement Automatique des Langues*, 49(2), 271–293.

Claridge, C. (2008). Historical Corpora. In A. Lüdeling, & M. Kytö (Eds.), *Corpus Linguistics. An International Handbook* (pp. 242–259). Volume 1. Berlin i.a.: de Gruyter.

Craig, H., & Whipp, R. (2010). Old Spellings, New Methods: Automated Procedures for Indeterminate Linguistic Data. *Literary and Linguistic Computing*, 25(1), 37–52. doi: 10.1093/llc/fqp033

Dickinson, M., & Meurers, W. D. (2003). Detecting Errors in Part-of-Speech Annotation. In *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 107–114), EACL-03, Budapest. http://decca.osu.edu/publications/dickinson-meurers-03.html. Accessed 22 March 2016.

Diel, M., Fisseni, B., Lenders, W., & Schmitz, H. (2002). *XML-Kodierung des Bonner Frühneuhochdeutschkorpus*. IKP-Arbeitsbericht NF 02, Bonn. https://korpora.zim.uni-duisburg-essen.de/Fnhd/ikpab-nf02.pdf. Accessed 22 March 2016.

Dipper, S. (2005). XML-Based Stand-Off Representation and Exploitation of Multi-Level Linguistic Annotation. In *Proceedings of Berliner XML Tage* (pp. 39–50), BXML 2005. Berlin.

Dipper, S., & 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* (pp. 27–42), NEALT, Oslo. http://www.ep.liu.se/ecp/article.asp?issue=087&article=003. Accessed 22 March 2016.

Donhauser, K. (2015). Das Referenzkorpus Altdeutsch: Das Konzept, die Realisierung und die neuen Möglichkeiten. In J. Gippert, & R. Gehrke, (Eds.), *Historical Corpora. Challenges and Perspectives* (pp. 25–50). Tübingen: Narr.

Dudenredaktion (2016). Duden online Wörtbuch. Berlin: Bibliographisches Institut GmbH. http://www.duden.de/woerterbuch. Accessed 23 March 2016.

Dudenredaktion (Ed.) (2005). *Dudengrammatik*. Band 4. 7. Auflage. Mannheim i.a.: Dudenverlag.

Durrell, M., Ensslin, A., & Bennett, P. (2007). The GerManC project. *Sprache und Datenverarbeitung* 31, 71–80.

Ebert, R. P. (1978). *Historische Syntax des Deutschen*. Stuttgart: J.B. Metzlersche Verlagsbuchhandlung.

Ernst-Gerlach, A. (2013). *Retrievalmethoden für historische Korpora mit nicht standardisierten Schreibweisen*. PhD thesis. Universität Duisburg. http://duepublico.uni-duisburg-essen.de/servlets/DerivateServlet/Derivate-33270/Ernst-Gerlach_Diss.pdf. Accessed 4 August 2015.

Gévaudan, P. (2002). Klassifikation der lexikalischen Entwicklungen. Semantische, morphologische und stratische Filiation. PhD Thesis, Universität Tübingen.

Geyken, A., Haaf S., & Wiegand F. (2012). The DTA-base Format: A TEI-Subset for the Compilation of Interoperable Corpora. In *Proceedings of the Conference of the 11th Conference on Natural Language Processing (KONVENS) – Empirical Methods in Natural Language Processing* (pp. 383–391), LThist 2012 Workshop. Vienna.

Gloning, T. (2007). Deutsche Kräuterbücher des 12. bis 18. Jahrhunderts. Textorganisation, Wortgebrauch, funktionale Syntax. In A. Meyer, & J. Schulz-Grobert (Eds.), *Gesund und krank im Mittelalter* (pp. 9–88). Leipzig: Eudora-Verlag.

Habermann, M. (2001). *Deutsche Fachtexte der frühen Neuzeit: naturkundlich-medizinische Wissensvermittlung im Spannungsfeld von Latein und Volkssprache*. Studia linguistica Germanica (61). Berlin i.a.: de Gruyter.

Hartweg, F., & Wegera, K. (2005). *Frühneuhochdeutsch. Eine Einführung in die Sprache des Spätmittelalters und der frühen Neuzeit*. 2. Auflage. Tübingen: Niemeyer.

Heiden, S. (2010) The TXM Platform: Building Open-Source Textual Analysis Software Compatible with the TEI Encoding Scheme. In R. Otoguro, K. Ishikawa, H. Umemoto, K. Yoshimoto & Y. Harada (Eds.), *24th Pacific Asia Conference on Language, Information and Computation* (pp. 389–398). Sendai, Japan.

Himmelmann, N. P. (2012). Linguistic Data Types and the Interface between Language Documentation and Description. *Language Documentation and Conservation*, (6), 187–207.

Höchli, S. (1981). *Zur Geschichte der Interpunktion im Deutschen. Eine kritische Darstellung der Lehrschriften von der zweiten Hälfte des 15.Jahrhunderts bis zum Ende des 18. Jahrhunderts*. Studia Linguistica Germanica (17). Berlin i.a.: de Gruyter.

Höder, S. (2012). Annotating Ambiguity: Insights from a Corpus-based Study on Syntactic Change in Old Swedish. In T. Schmidt & K. Wörner (Eds.), *Multilingual Corpora and Multilingual Corpus Analysis* (pp. 245–271). Hamburg studies on multilingualism (14). Amsterdam i.a.: Benjamins.

Jurish, B. (2010). More than Words: Using Token context to improve Canonicalization of Historical German. *Journal for Language Technology and Computational Linguistics*, 25(1), 23–39.

Klein, T. (2013). Verknüpfung digitaler Lemmalisten historischer Sprachstufen des Deutschen. Wie und wozu? Talk at *Arbeitsgespräch zur historischen Lexikographie. Universität Trier*. https://www.uni-trier.de/fileadmin/forschung/maw/MWB/Arbeitsgespraech2013/Vortrag_Bullay_Klein.pdf Accessed 22 March 2016.

Klein, W. P. (1999). *Die Geschichte der meteorologischen Kommunikation in Deutschland. Eine historische Fallstudie zur Entwicklung von Wissenschaftssprachen*. Postdoctoral thesis, Freie Universität Berlin.

Krause, T., & Zeldes, A. (2016). ANNIS3: A new Architecture for Generic Corpus Query and Visualization. *Digital Scholarship in the Humanities* 31(1), 118–139. doi: http://dx.doi.org/10.1093/llc/fqu057. Accessed 22 March 2016.

Krause, T., Lüdeling, A., Odebrecht, C., & Zeldes, A. (2012) Multiple Tokenizations in a Diachronic Corpus. In *Exploring Ancient Languages through Corpora Conference*, EALC 2012. Oslo.

Kroch, A., Santorini, B., & Delfs, L. (2004). *The Penn-Helsinki Parsed Corpus of Early Modern English (PPCEME)*. Department of Linguistics, University of Pennsylvania. CD-ROM, first edition.

Kroch, A., & Taylor, A. (2000). *The Penn-Helsinki Parsed Corpus of Middle English (PPCME2)*. Department of Linguistics, University of Pennsylvania. CD-ROM, second edition.

Kübler, S., & Zinsmeister, H. (2015). *Corpus Linguistics and Linguistically Annotated Corpora*. London i.a.: Bloomsbury.

Kytö, M., & Pahta, P. (2012). Evidence from Historical Corpora up to the Twentieth Century. In T. Nevalainen, & E. C. Traugott (Eds.), *The Oxford Handbook of the History of English* (pp. 123–133). Oxford i.a.: Oxford University Press.

Kytö, M. (2011). Corpora and Historical Linguistics. *Revista Brasileira de Linguística Aplicada*, 11(2), 417–457. doi: 10.1590/S1984-63982011000200007. Accessed 22 March 2016.

Kytö, M. (1996). *Manual to the Diachronic Part of the Helsinki Corpus of English Texts: Coding Conventions and Lists of Source Texts*. Third edition. Helsinki: University of Helsinki, Department of English.

Lindauer, T. (1995). *Genitivattribute. Eine morphosyntaktische Untersuchung zum deutschen DP/NP-System*. Tübingen: Niemeyer.

Lüdeling, A. (2011). Corpora in Linguistics: Sampling and Annotation. In K. Grandin (Ed.), *Going Digital. Evolutionary and Revolutionary Aspects of Digitization*. Nobel Symposium 147 (pp. 220–243). New York: Science History Publications.

Lüdeling, A.; Poschenrieder, T., Faulstich, L. C., et al. (2005). DeutschDiachronDigital – Ein diachrones Korpus des Deutschen. *Jahrbuch für Computerphilologie* 2004, 119–136.

Nerius, D. (2007). *Deutsche Orthographie*. 4th revised Edition. Hildesheim i.a.: Olms.

Nerius, D. (2003). Graphematische Entwicklungstendenzen in der Geschichte des Deutschen. In W. Besch, A. Betten, O. Reichmann, & S. Sonderegger (Eds.), *Sprachgeschichte. Ein Handbuch zur Geschichte der deutschen Sprache und ihrer Erforschung* (pp. 2461–2472). Vol. 3, Second Edition. Berlin i.a.: de Gruyter.

Odebrecht, C., Krause, T., & Lüdeling, A. (2015). Austausch von historischen Texten verschiedener Sprachen über das LAUDATIO-Repository. In *37. Jahrestagung der Deutschen Gesellschaft für Sprachwissenschaft*, DGfS-CL Poster Session 2015. Leipzig.

Odebrecht, C. (2014). Modeling Linguistic Research Data for a Repository for Historical Corpora. In *Proceedings of Digital Humanities 2014 Conference, DH Conference 2014* (pp. 284–285), Université de Lausanne, Lausanne. https://dh2014.files.wordpress.com/2014/07/dh2014_abstracts_proceedings_07-11.pdf. Accessed 22 March 2016.

Pahta, P., & Taavitsainen, I. (2010). Scientific Discourse. In A. H. Jucker, & I. Taavitsainen (Eds.), *Historical Pragmatics* (pp. 549–586). Vol. 8. Berlin: Mouton de Gruyter.

Paul, H. (1995). *Prinzipien der Sprachgeschichte*. 10. Auflage. Tübingen: Niemeyer.

Perlitz, L. (2014). *Konkurrenz zwischen Wortbildung und Syntax: Historische Entwicklung von Benennung*. Bachelorarbeit, Humboldt-Universität zu Berlin.

Petrova, S., Solf, M., Ritz, J., Chiarcos, C., Zeldes, A., et al. (2009). Building and using a Richly Annotated Interlinear Diachronic Corpus: The Case of Old High German Tatian. *Traitement automatique des langues*, 50(2), 47–71.

Pilz , T. (2009). *Nichtstandardisierte Rechtschreibung – Variationsmodellierung und rechnergestützte Variationsverarbeitung*. PhD Thesis. Universität Duisburg-Essen.

Piotrowski, M. (2012). *Natural Language Processing for Historical Texts*. Synthesis Lectures on Human Language Technologies; 17. San Rafael: Morgan & Claypool.

Pörksen, U. (2003). Deutsche Sprachgeschichte und die Entwicklung der Naturwissenschaften – Aspekte einer Geschichte der Naturwissenschaftssprache und ihrer Wechselwirkung zur Gemeinsprache. In W. Besch, A. Betten, O. Reichmann, & S. Sonderegger (Eds.), *Sprachgeschichte. Ein Handbuch zur Geschichte der deutschen Sprache und ihrer Erforschung* (pp. 193–210). Vol.1, Second Edition. Berlin i.a.: de Gruyter.

Pose, J., Lopez, P. & Romary, L. (2014). *A Generic Formalism for Encoding Stand-off Annotations in TEI*. INRIA Technical Report. hal-01061548. Accessed 22 March 2016.

Reichmann, O., & Wegera, K.-P. (1993).Schreibung und Lautung. In Reichmann, O., & Wegera, K.P. (Eds.) (1993). *Frühneuhochdeutsche Grammatik* (pp. 13–163). Tübingen: Niemeyer.

Reznicek, M., Lüdeling, A., & Hirschmann, H. (2013). Competing Target Hypotheses in the Falko Corpus: A Flexible Multi-Layer Corpus Architecture. In A. Díaz-Negrillo (Ed.), *Automatic Treatment and Analysis of Learner Corpus Data* (pp. 101–123). Amsterdam: John Benjamins.

Reynaert, M., Hendricks, I., & Marquilhas, R. (2012) Historical Spelling Normalization. A Comparison of Two Statistical Methods: TICCL and VARD2. In *Proceedings of the Second Workshop on Annotation of Corpora for Research in the Humanities*, ACRH 2012. Lisbon.

Riecke, J. (2007). Beiträge zum mittelalterlichen deutschen Wortschatz der Heilkunde. In A. Meyer, & J. Schulz-Grobert (Eds.), *Gesund und krank im Mittelalter. Marburger Beiträge zur Kulturgeschichte der Medizin* (pp. 89–106). Leipzig: Eudora-Verlag.

Rieke, J. (2004). *Die Frühgeschichte der mittelalterlichen medizinischen Fachsprache im Deutschen. Band 1: Untersuchungen, Band 2: Wörterbuch*. Berlin, New York: Walter de Gruyter.

Rissanen, M. (2012). Corpora and the study of English historical syntax. In M. Kytö (Ed.), *English Corpus Linguistics: Crossing Paths* (pp. 197–220). Amsterdam, New York: Rodopi.

Rissanen, M. (2008). Corpus Linguistics and Historical Linguistics. In A. Lüdeling, & M. Kytö (Eds.), *Corpus Linguistics. An International Handbook* (pp. 53–68). Vol 1. Berlin i.a.: de Gruyter.

Romary, L. (2009). Questions & Answers for TEI Newcomers. *Jahrbuch für Computerphilologie, 10. Digital Libraries*. doi: http://arxiv.org/abs/0812.3563.

Schiller, A., Teufel, S., Stöckert, C., & Thielen, C. (1999). *Guidelines für das Tagging deutscher Textcorpora mit STTS (Kleines und großes Tagset)*. Universität Stuttgart und Tübingen. For STTS Tag Table (1995/1999) see http://www.sfs.uni-tuebingen.de/resources/stts-1999.pdf. Accessed 1 March 2016.

Schmid, H. (2008). Tokenization and Part-of-speech Tagging. In A. Lüdeling, & M. Kytö (Eds.), *Corpus Linguistics. An International Handbook* (pp.527–551). Vol 1. Berlin i.a.: de Gruyter.

Schmid, H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. In *Proceedings of International Conference on New Methods in Language Processing*, 1994. Manchester.

Simmler, F. (2003). Geschichte der Interpunktionssysteme im Deutschen. In W. Besch, A. Betten, O. Reichmann, & S. Sonderegger (Eds.), *Sprachgeschichte. Ein Handbuch zur Geschichte der deutschen Sprache und ihrer Erforschung* (pp. 2472–2504). Vol. 3, Second Edition. Berlin i.a.: de Gruyter.

Splett, J. (2000). Wortbildung des Althochdeutschen. In W. Besch, A. Betten, O. Reichmann, & S. Sonderegger (Eds.), *Sprachgeschichte. Ein Handbuch zur Geschichte der deutschen Sprache und ihrer Erforschung* (pp. 1213-1222). Vol. 2, Second Edition. Berlin i.a.: de Gruyter.

Springmann, U., & Lüdeling, A. (submitted). *Progress of OCR of Early Printings exemplified by the RIDGES Herbology Corpus*.

Squires, C. (2010). Konstantes und Variables im Aufbau von deutschen mittelalterlichen heilkundlichen Texten und angrenzenden Textsorten In A. Ziegler (Ed.), *Diachronie, Althochdeutsch, Mittelhochdeutsch 1: Historische Textgrammatik und Historische Syntax des Deutschen* (pp. 561–588). Berlin i.a.: de Gruyter.

Stede, M., & Neumann, A. (2014). Potsdam Commentary Corpus 2.0: Annotation for Discourse Research. In *Proceedings of the Language Resources and Evaluation Conference* (pp. 925–929), LREC 2014, Reykjavik.

TEI Consortium (Eds.) (2015) *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. Version 2.8.0. 2015-04-06. TEI Consortium. http://www.tei-c.org/Guidelines/P5/. Accessed 13 August 2015.

Vikør, L. (2004). Lingua Franca and International Language. Verkehrssprache und Internationale Sprache. In U. Ammon (Ed.) *Sociolinguistics. An international handbook of the science of language and society* (pp. 328–334). Berlin i.a.: de Gruyter.

Voigt, V. (2013) Python Script for the Normalization Layer *clean*. Script and Documentation see point 6 at http://korpling.german.hu-berlin.de/ridges/documentation_v4_en.html. Accessed 1 March 2015.

Wolff, G. (2009). *Deutsche Sprachgeschichte von den Anfängen bis zur Gegenwart*. 6. Edition. Tübingen and Basel: Narr Francke.

Zeldes, A., & Schroeder, C. T. (2015). Computational Methods for Coptic: Developing and Using Part-of-Speech Tagging for Digital Scholarship in the Humanities. *Digital Scholarship in the Humanities* 31(1), 164–176. doi: http://dx.doi.org/10.1093/llc/fqv043. Accessed 22 March 2016.

Zipser F., & Romary, L. (2010). A Model oriented Approach to the Mapping of Annotation formats using Standards. In *Proceedings of the Workshop on Language Resource and Language Technology Standards*, LREC 2010. Malta.