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 und 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

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 and Uwe Springmann and Thomas Krause 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 herbal texts, which has been constructed to study the emergence and change of scientific registers in a vernacular language of Europe (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 has a number of additional issues. The lexicon changes with the formation of terminology, spelling regularities emerge, word-formation, syntax, and text structure develop. All of this poses challenges to consistent annotation. At the same time we see changes in typesetting and printing methods which complicate automatic digitization. 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, or 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 a category 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)?
- Is 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 meet these challenges 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 required and the conceptual levels of analysis that together ensure an extensible, reusable and comparable corpus for the study of a register across time. Some 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 we talk about graphical and structural annotations such as line breaks and rendering (3.3) and different layers of linguistic annotation (3.4). We will discuss our decisions vis à vis other

[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.

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

The RIDGES corpus, version 4.1 contains 29 excerpts of 24 publications of herbal texts, ranging from 1478 to 1870, with approximately 30 years between the texts. New texts are added to the corpus at irregular intervals.[2] Herbal texts are chosen because they 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. Gloning 2007; Riecke 2004, 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 corpus contains excerpts of about 3000-4000 words each of prose texts such as advice books, lectures, and scientific texts (currently 154,267 tokens in total). Each document is stored with comprehensive bibliographic metadata such as title, author, editor, publication place, publisher and year. 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 (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 scans provided by Google Books[3] or research libraries[4] were chosen. The texts are digitized diplomatically (Section 3.1), normalized (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 SaltNPepper (Zipser and Romary 2010). The corpus is stored in the stand-off format PAULA XML (Dipper 2005), and its annotations are accessible via ANNIS[5], a browser based search and visualization platform (Chiarcos et al. 2008; Krause and Zeldes 2014). The corpus with all formats is archived long-term and extensively documented in the LAUDATIO-Repository (Odebrecht et al. 2015).[6]

2.2 Multi-layer Architecture

In this section we want to motivate the need for a multi-layer corpus architecture with the possibility for multiple tokenization. By tokenization we mean the segmentation of the primary data[7] into the smallest annotatable units (Schmid 2008), and annotation means 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.

[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 they could be published. The corpus is growing; we expect to publish version 5 in early 2016.

[3] https://books.google.de/

[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 .

[5] ANNIS, which stands for ANNotation of Information Structure, was originally designed to provide access to the data of the SFB 632 - Information Structure, http://annis-tools.org/.

[6] 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.

[7] 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).

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*). 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 tokenized 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. Table (1). While each of these decisions in Table (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.[8]

Tab. 1 Different tokenizations for *can't*

tok_a	*we*	*can't*	*do*	*that*		
tok_b	*we*	*can*	*t*	*do*	*that*	
tok_c	*we*	*can*	*'t*	*do*	*that*	
tok_d	*we*	*can*	*'*	*t*	*do*	*that*
tok_e	*we*	*can*	*n't*	*do*	*that*	

Especially in 'non-standard' texts such as historical texts it may be desirable to have different segmentations, in order to deal with different research questions. Thus, if there is conceptually more than one segmentation, the corpus architecture should also support this (Krause et al. 2012).

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 one token will be called a **token annotation**). A sequence of tokens can be categorized as multi-word expression (an idiom, say), or a sentence type (we will call this a **span annotation**). The pos annotation in Table (2) is a token annotation while the syntax annotation is a span annotation.

Tab. 2 Example for token and span annotation, 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	waſſer
pos	ART	NN	VVFIN	APPR	NN	
syntax	NP			PP		

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

The graph-based architecture we use (ANNIS)[9] 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) 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).

3. Annotation

Given the multi-layer standoff architecture described above we will now explain how we pre-processed the corpus: Section 3.1 discusses the transcription, Section 3.2 multiple normalizations and multiple tokenizations. Based on the different normalizations, Section 3.3 presents the graphical annotations and Section 3.4 the linguistic annotations.

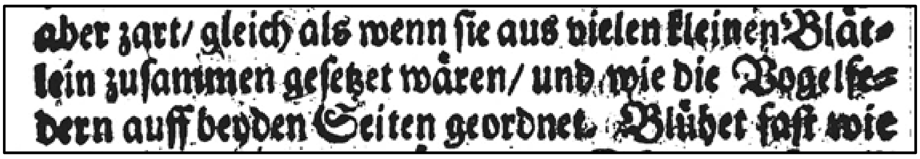
3.1 Transcription

A central issue that recurs in almost all historical corpora 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 precision regarding font usage, special characters, typesetting and encoding.

The first transcription (called *dipl*) is fairly narrowly diplomatic: We assign each glyph to a Unicode character[10]. 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+0364), *ſ* (U+017F) and ⸗ (U+2E17). The Unicode standard provides characters for most of the glyphs needed for old German texts.[11] As the multi-layer corpus architecture requires tokenization in each layer, the transcription *dipl* is tokenized. 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).

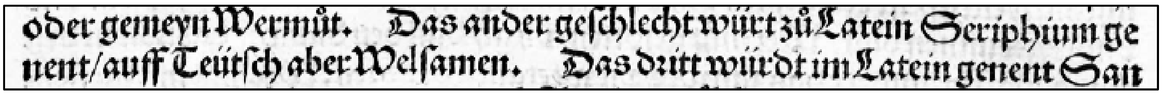
> **Ex. 1a** Diplomatic Transcription, Curioser Botanicus oder sonderbares Kräuterbuch (1675)
> *aber zart / gleich als wenn ſie aus vielen kleinen Blaͤt⸗*
> *lein zuſammen geſetzet waͤren / und wie die Vogelfe⸗*
> *dern auff beyden Seiten geordnet . Bluͤhet faſt wie*
> '... but gentle such as when they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

[9] Other corpus projects using a similar corpus architecture are Falko (Reznicek et al. 2013), DDD-AHD (Richling 2011), Coptic Scriptorium (Zeldes and Schroeder to appear) and PCC (Stede and Neumann 2014).
[10] For the official Unicode table see http://unicode-table.com/.
[11] 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.



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

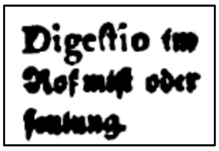
Ex. 1b Separate ‘words’ at a line break, New Kreüterbůch(1543)
oder gemeyn Wermůt . Das ander geſchlecht würt zů Latein Seriphium <u>*ge*</u>
<u>*nent*</u> */ auff Teütſch aber Welſamen . Das dritt würdt im Latein genent San*
‘or common Vermouth. The other kind came to be <u>called</u> Latin Sriphium
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 (1) The final letters are l*ung* but the letters before *lung* are not unambiguously readable. We mark this by an underscore (here: *_lung*).

Fig. 1 Margin, Alchymistische Practic (1603)
Digeſtio im Roſmiſt oder_lung
‘Digestion in horse-manure or … [?]’



Digeſtio im
Roſmiſt oder
[illegible]lung.

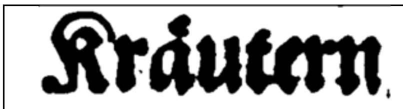
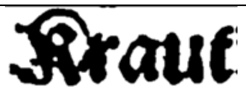
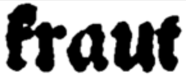
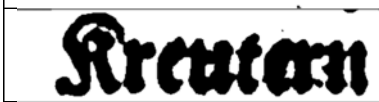
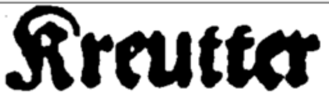
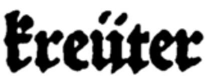
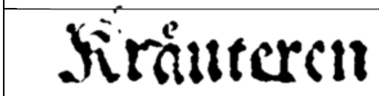
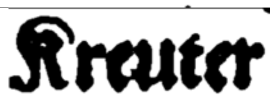
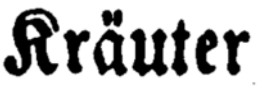
Since the insertion of margin or footnote text would prevent syntactic tagging, margin texts are inserted before the paragraph containing them, whereas footnote text is inserted at the end of the paragraph, though their nature as notes and marginalia is annotated. Font type and characteristics, margins, 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[12] 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

In addition to the diplomatic layer we need normalization. The spelling variation in historical documents is significant and to some extent unpredictable. The variance is even higher in diachronic documents (see Figure 2 for some of the variants we find in RIDGES for Kräuter ’herbs’). Normalized layers help us in (a) finding instances of ‘the same’ phenomenon, (b) making generalizations, and (c) further linguistic processing.

Fig. 2 Spelling variation of the lemma *Kraut* (herb), RIDGES Corpus 4.1

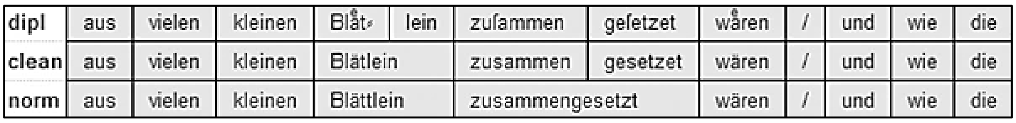
[12] TEI stands for Text Encoding Initiative, for an introduction see Romary (2009) and Section (3.3).

	Kraͤutern	Alchymistische Practic (1603)
	Kraut	Alchymistische Practic (1603)
	kraut	Alchymistische Practic (1603)
	Kreutern	Alchymistische Practic (1603)
	Kreutter	Alchymistische Practic (1603)
	kreüter	New Kreüterbůch (1543)
	Kraͤuteren	Pflantz-Gart (1639)
	Kreuter	Alchymistische Practic (1603)
	Kräuter	Deutsche Pflanzennamen (1870)

There can, in principle, be infinitely many 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.[13] In the current version of RIDGES, we have two normalized layers, called *clean* and *norm*.

clean is generated automatically and requires no deep linguistic analysis. The *dipl* layer is normalized in the *clean* layer in the following way: The first normalization step in *clean* merges 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.[14]

Fig. 3 Normalizations, visualization in ANNIS of the Example (1a)



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

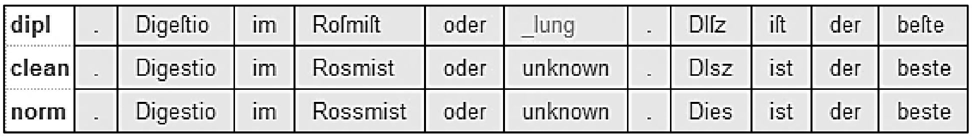
[13] Technically, a normalization layer is just an annotation layer, but flagging it as a segmentation layer makes it behave like one of a set of alternative tokenization layers that the search engine, ANNIS, treats as the basic text of a document. This affects both the initial view of search results and the ability to define search context and distance between search elements in term of *n* - *m* normalized or diplomatic tokens.

[14] See Voigt (2013) for guidelines, http://korpling.german.hu-berlin.de/ridges/download/v4/cleanV2README.txt.

The different character realization 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 (a span over two or more tokens, see *Blaͤt⸗* and *lein* in Figure 3).

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 (4).

Fig. 4 Uninterpretable Text, visualization in ANNIS of the example given in Figure (1)



dipl	.	Digeſtio	im	Roſmiſt	oder	_lung	.	Dlſz	iſt	der	beſte
clean	.	Digestio	im	Rosmist	oder	unknown	.	Dlsz	ist	der	beste
norm	.	Digestio	im	Rossmist	oder	unknown	.	Dies	ist	der	beste

Thus, *clean* can be interpreted both as an annotation on *dipl* as well as an independent segmentation. The *clean* layers is a robust and simple form of normalization because it affects the text only on a character level without requiring definitions of words, word forms and or other linguistic concepts. It is also completely predictable from *dipl*.

However, the normalization in *clean* is not sufficient to find all the different spellings of the 'same word', such as *Kräutern*, *Krauttern*, *Kraͤutteren* for *Kräutern* in Figure (2). Different capitalizations, double consonants such as *tt* and 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 that maps these different forms to one form.[15] 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 Early Modern German (EMG) word forms.[16] In a diachronic corpus such as RIDGES one can be even more abstract and map all word forms to a modern word form.

Consider the different spellings of *Krankheit* 'illness' in Table (3). The *dipl* layer represents the original spelling. As the *clean* layer operates automatically and does not impose linguistic decisions, macrons (*ā*) used for either *an* or *am* in EMG are dissolved with both possible interpretations *kramckhait|kranckhait* and *kranck* and *haít* are not combined because the original does not contain a hyphen. In *norm* all forms are mapped to the modern form (token annotation for the last two examples, span annotation for the first example).

Tab. 3 Normalization of *Krankheit* ('illness'), Gart der Gesundheit (1487)

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

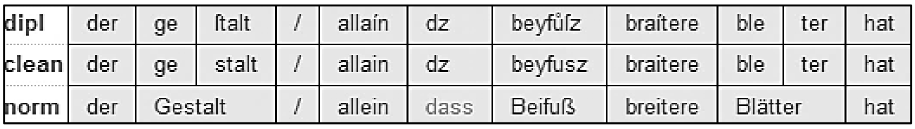
[15] Note that there is a different way of dealing with the search problem, namely the mapping of different forms in the search itself, a.k.a. fuzzy search. See the references on automatic normalization in Section (3.5).

[16] This is the decision taken by many historical corpora that cover one language stage, cf. Richling (2011) and Rissanen (2012).

The mapping of historical spellings to modern word forms is by no means always unproblematic and requires interpretation and linguistic decisions (Gévaudan 2002).[17]

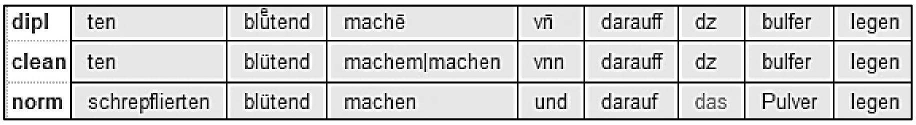
Depending on its syntactic use, the word form *dz* in Figures (5a) and (5b), underscored in the captions below, can be mapped onto the modern German complementizer *dass* (5a) or the definite article *das* (5b).[18] The mapping thus needs a syntactic analysis. Another case in point is word formation. The spelling of compounds differs even for the same word and in the same text, and often it is unclear whether a word is a genitive form, a compound or a syntactic combination (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. 5a Normalization of a complementizer, visualization in ANNIS, Gart der Gesundheit (1487)
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. 5b Normalization of a definite article, visualization in ANNIS, Gart der Gesundheit (1487)
ten bluͤtend machē vn̄ darauff dz bulfer legen
'... make blossom and then lay the powder.'



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

The construction of the *norm* layer affects different phenomena on different levels: spelling, morphological and morpho-syntactic phenomena such as inflection and compound spellings, as well as syntactic phenomena.

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 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. Many projects use the TEI Guidelines to create digital critical editions which focus on the exact

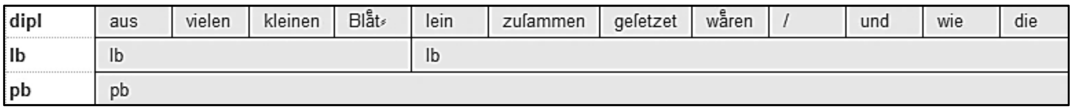
[17] Another problem of this approach is conceptual: Is it useful to map forms of one language to forms (and ultimately categories) of another language? Which interesting distinctions and properties are lost? This issue (similar to the debate about the comparative fallacy in second language acquisition research, see Bley-Vorman 1983) is interesting and needs to be discussed further.

[18] The text also contains the form *das* in both interpretations. The choice between *das* and *dz* seems to be driven by typographic needs.

diplomatic markup of historical texts.[19] In contrast to critical editions, the RIDGES project uses only a few elements representing markup information which is essential for linguistic reasons. 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 out of TEI elements such as <lb>, <head> and @rend attributes, and implement them in our span annotations.

In Figure (6), *lb* (linebreak) reflects the original text form and allows to discriminate between hyphenation due to the end of the line and hyphenated compound spelling. The *lb* annotation span extends from the point at which a line beings and up to the linebreak itself (in TEI XML, only the position of the line break is marked with a unary element, <br/>). 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 normalization can easily be applied in *norm*, cf. Figure (3). In this case, the *norm* segmentation interacts with the *lb* annotation in that it incorporates an *lb* boundary. The structural annotations *head* and *note* allow for specific decisions during a linguistic analysis. For example, one may decide on only including the continuous text and exclude the textual material in head, margin or footnote areas, because they may behave differently. Otherwise, a research question on margins can easily query only margins. The different fonts (e.g. Antiqua, Gothic) are annotated in the *hi_rend* layer instead of transcribing them in *dipl* (varieties like Antiqua and Gothic letters are also not considered distinct Unicode symbols).

Fig. 6 Annotation of line and page breaks, visualization in ANNIS of Example (1a)



dipl	aus	vielen	kleinen	Blåͤt⸗	lein	zuſammen	geſetzet	wåͤren	/	und	wie	die
lb	lb				lb							
pb	pb											

After having assembled all the necessary information about document structure, the next section will give some examples for a linguistic analysis based on linguistic annotations, such as part-of-speech tagging, lemmatization and the analysis of compound nouns.

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 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 to be part-of-speech tagged automatically, cf. Figure (2). 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, e.g. *Kraut*, see Table (3). 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 Table (4), 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 where the transcription *dipl* is segmented into two tokens, cf. Table (3). Here, *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.

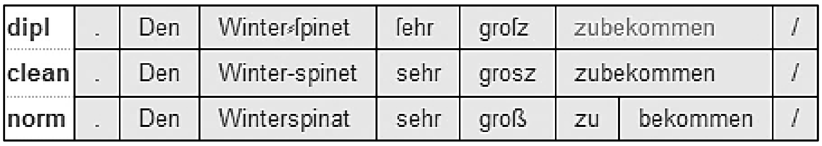
[19] There are, among many others, the Duisburg-Leipzig Korpus romanischer Zeitungssprachen http://home.uni-leipzig.de/burr/CorpusLing/Korpusanalyse/default.htm (Burr et al. 2015), Deutsches Text Archiv http://www.deutschestextarchiv.de/ (Geyken et al. 2012), Coptic Scriptorium http://copticscriptorium.org/, see Zeldes and Schroeder (to appear).

Tab. 4 Example of linguistic annotations for *Kraut*, RIDGES Corpus 4.1

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

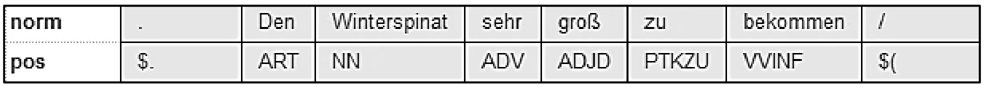
In Figure (7a), *zubekommen* is transcribed as one token, but split up in the normalization. Thus, the split-up segments can be annotated separately on the *pos* layer, and can now be found with queries for all infinitives (VVINF) or the infinitive particle *zu* (PTKZU), cf. Figure (7b).

Fig. 7a Split-up normalization, visualization in ANNIS, Pflantz-Gart Capitel 4 (1639)
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	/

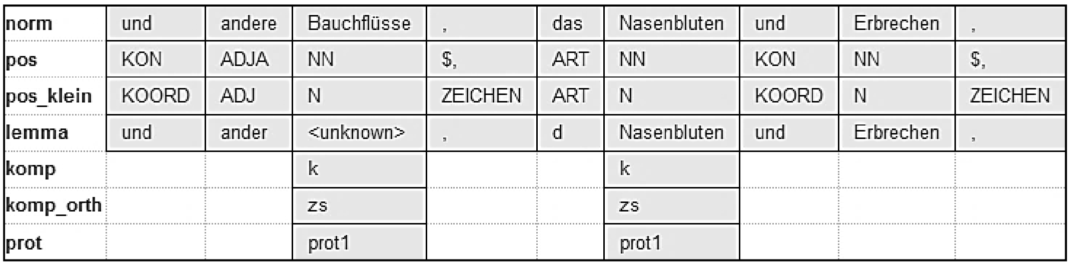
Fig. 7b Part-of-Speech annotation, visualization in ANNIS, Pflantz-Gart Capitel 4 (1639)



norm	.	Den	Winterspinat	sehr	groß	zu	bekommen	/
pos	$.	ART	NN	ADV	ADJD	PTKZU	VVINF	$(

Another 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; Splett 2000; Lindauer 1995). 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* would require an -*s*: *Bauchs*. However, for a form such as *Teufelswurzel* (literally: devil's root, 'hyoscamus', 'Devilsroot') it is difficult to determine whether the -*s* represents a genitive or a compound linking element in its synchronic period, and much spelling variation is found (for a detailed discussion see Perlitz 2014). Her annotation of the different spelling types 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 (8).

Fig. 8 Annotation of compounds, visualization in ANNIS, Die Eigenschaften aller Heilpflanzen (1828)
und andere Bauchfluͤſſe , das Naſenbluten und Erbrechen , befeſtiget ,
‘and other stomach flows, steadies nosebleeds and vomiting, ...’



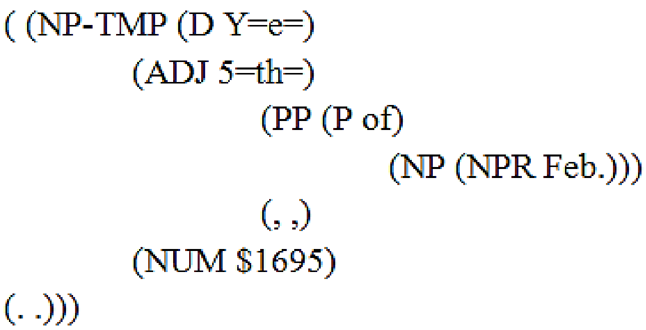
norm	und	andere	Bauchflüsse	,	das	Nasenbluten	und	Erbrechen	,
pos	KON	ADJA	NN	$,	ART	NN	KON	NN	$,
pos_klein	KOORD	ADJ	N	ZEICHEN	ART	N	KOORD	N	ZEICHEN
lemma	und	ander	<unknown>	,	d	Nasenbluten	und	Erbrechen	,
komp			k			k			
komp_orth			zs			zs			
prot			prot1			prot1			

3.5 Discussion

After having presented the architecture and pre-processing decisions we took for RIDGES we want to briefly present and discuss other approaches to the construction of historical and diachronic corpora and further illustrate the advantages of a multilayer 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 (decisions about normalization were built into the textual layer (one well-known and influential example is the Helsinki Corpus of Old English Texts[20]). 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 could be used. Figure (9) illustrates the format:

Fig. 9 Fragment from the Penn Parsed Corpus of Early Modern English for ‘The 5th of Feb. 1695’



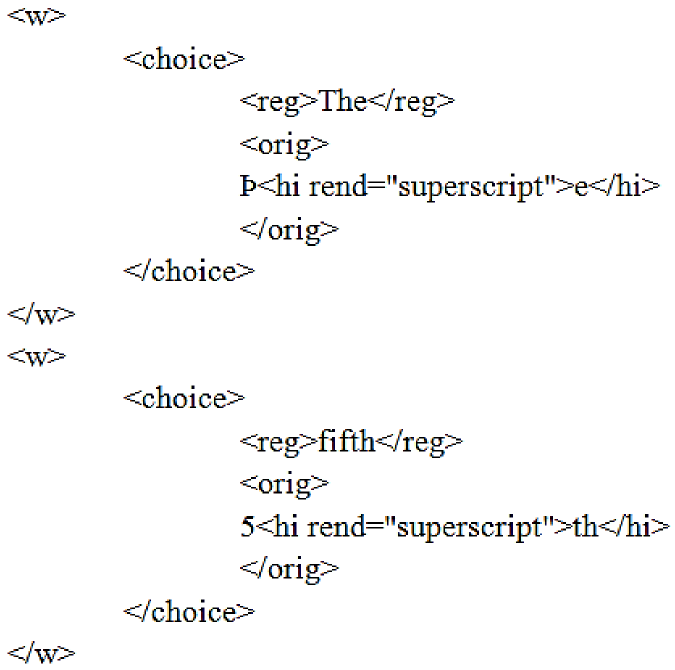
```
( (NP-TMP (D Y=e=)
          (ADJ 5=th=)
                   (PP (P of)
                            (NP (NPR Feb.)))
                   (, ,)
          (NUM $1695)
(. .)))
```

The brackets in 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 (9). 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

[20] http://www.helsinki.fi/varieng/CoRD/corpora/HelsinkiCorpus/

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. 10 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 (10), 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*). However adding syntactic information as in PPCEME, and other linguistic information, becomes cumbersome at best, and depending on the annotations, impossible at worst.

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) 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 being built which cannot be normalized and therefore prevent a linguistic search without being able to 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 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.

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 hand have not worked well so far either, as the corpus is too small for statistical training as applied e.g. by Bollman et al. (2011), Jurish (2010), or

Archer et al. (2015). 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. Pilz 2009; Baron and Rayson 2008; 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 (2) 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 are planned.

4. Case Studies

In the following section, we will briefly illustrate how the multi-layer and multiple tokenization architecture 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.

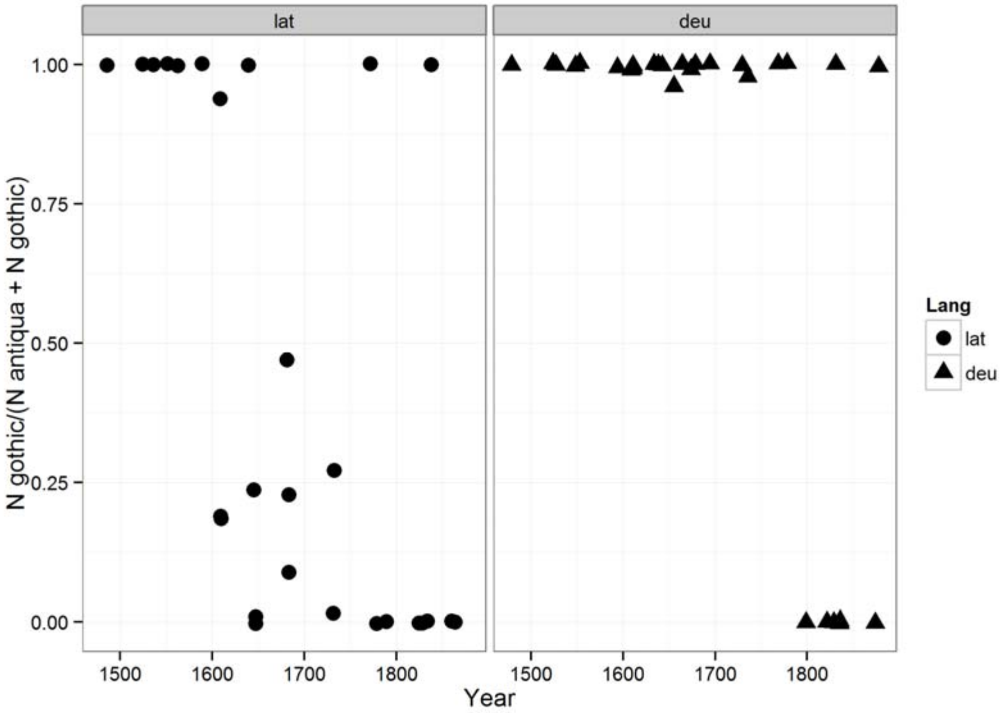
4.1 Fonts Depending on Language

(German) historical texts differ, among other things, with respect to their typeface, which typically has many fonts representing it. The interaction between the typeface used and the language which is printed may give a first insight into the function and distribution of object and meta language in scientific texts. The RIDGES corpus contains both necessary types of annotation: typeface (*font*) and language (l*ang*). Both annotation concepts were inspired by the TEI Guidelines.[21] The use of the two typefaces Antiqua and Gothic is annotated in RIDGES on the graphical annotation layer *font*, which is based on the diplomatic transcription *dipl*. The language is annotated in the layer *lang* with the ISO 639-2 language codes, e.g. *deu* for German, *lat* for Latin and *eng* for English.[22] Figure (9) shows the correlation between the font distribution within a text and the two most frequent languages, namely German and Latin. For German, there is a change from the predominantly used Gothic to Antiqua, starting around 1800. Interestingly, we observe that texts are either printed completely in Antiqua or Gothic typeface. For Latin, the typeface Antiqua seems to mark Latin terms or descriptions beginning around 1600, as can be seen on the right panel of Figure (11). However, the change observed here is not categorical as it is for German, but varies to differing degrees until 1750. The first results may imply that Latin is used as what might nowadays be described as object language in linguistics. Thus, we can assume that the vernacular German widens to a scientific language.

[21] 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.

[22] http://www.loc.gov/standards/iso639-2/php/code_list.php

Fig. 11 Distribution of the typefaces Antiqua and Gothic for German (*deu*) and Latin (*lat*) in RIDGES 4.1



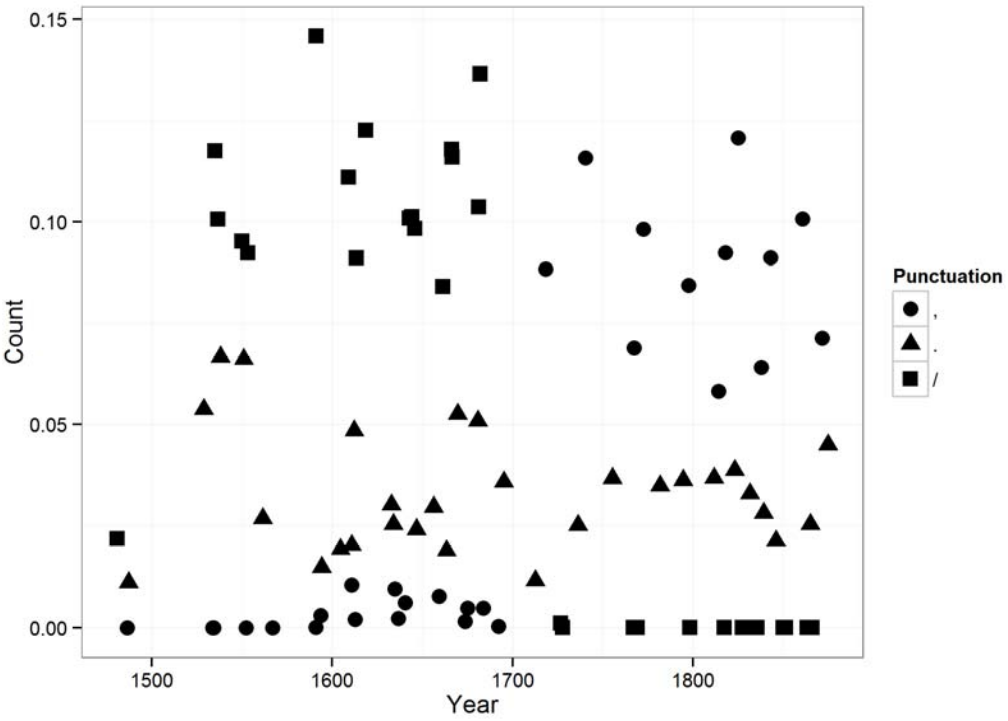
4.2 Punctuation

In written Modern German, the distribution and function of punctuation is regulated in the orthography (see for example Duden 2005, 1072-1073). In former stages of German, there was no binding orthographic norm for punctuation in the written language (Simmler 2003; Nerius 2007; Höchli 1981). Thus, there is variation in punctuation next to the spelling of word forms (see Section 4.3). In the following case study, we investigate the distribution of the three punctuation types: . , and / in order to gain empirical insights into their potential functions. Therefore, we use *dipl*, as all punctuation instances are already segmented during the transcription.

Figure (12) shows the distribution of . , and / 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 have been 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 (12) shows a tendency which is described and discussed as a change in the use of punctuation, or text structuring characters (Reichmann & Wegera 1993, Höchli 1981). With the help of the multi-layer corpus architecture, we are able to 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. Further annotation might reveal interesting differences in the use of punctuation over the centuries.

Fig. 12 Punctuation frequencies per text in RIDGES 4.1



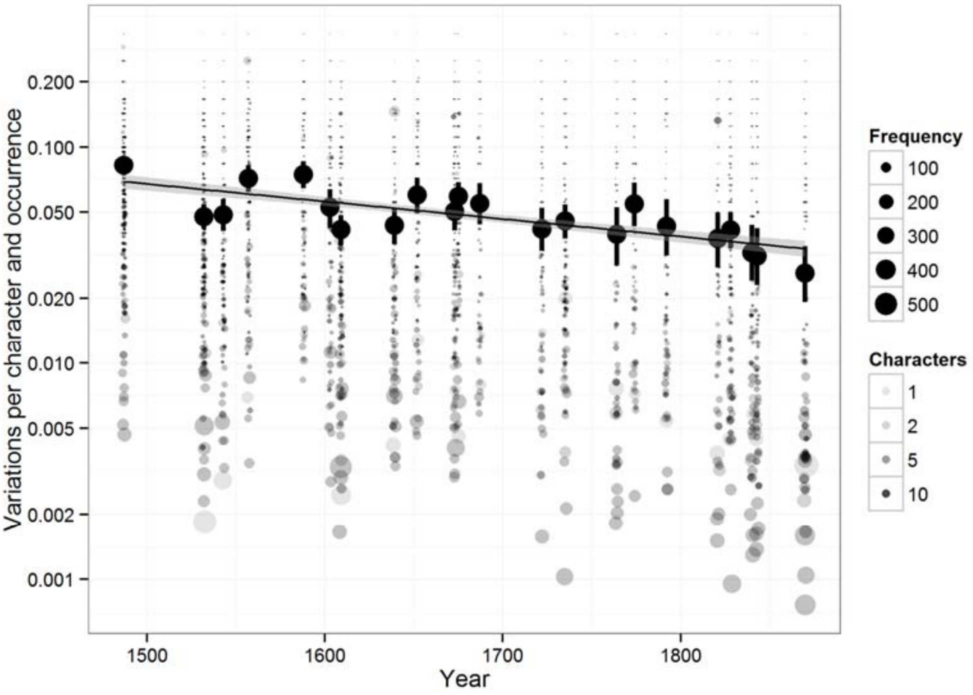
4.3 Spelling Variation

It is interesting to investigate whether standardization is only carried out 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 (Nerius 2003; Besch 2003 Nerius 2007; Reichmann and Wegera 1993; Wolff 2009). Since the late 19th century, Modern German is highly regulated, influenced by a sequence of standardization committees[23] and decisions 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), see for an overview Nerius (2007). Thus, we would assume that the variation between the historic *dipl* and the modern German *norm* decreases over time. This, then, would reflect a systematic convergent trend in spelling.

Here, spelling variation reflects the variance per character and occurrence between *dipl* and *norm* over time. We can measure spelling variation across the corpus texts by measuring the spelling variation on the earliest text (1487) and then using this baseline of variance for comparing all word forms in the later texts. Consider the word *iſt* (*dipl* 'is') and its normalized equivalent *ist* (*norm* 'is'). For earlier texts, there is a variation in writing, whereas for later texts with the spelling *ist* on *dipl*, there is no variation. We considered the length and frequency of each *dipl*-token by integrating the amount of characters and their occurrence in this corpus. Figure (13) shows that the spelling variance of the *dipl*-token seems in fact to decrease gradually. The results are based on surface information only and do not allow conclusions about the cause of this variation without further study.

[23] Currently it is the ‚Rat für deutsche Rechtschreibung' http://www.rechtschreibrat.com/.

Fig. 13 Spelling variations in *dipl* vs. *norm* with the earliest text (1487) as baseline in RIDGES 4.1

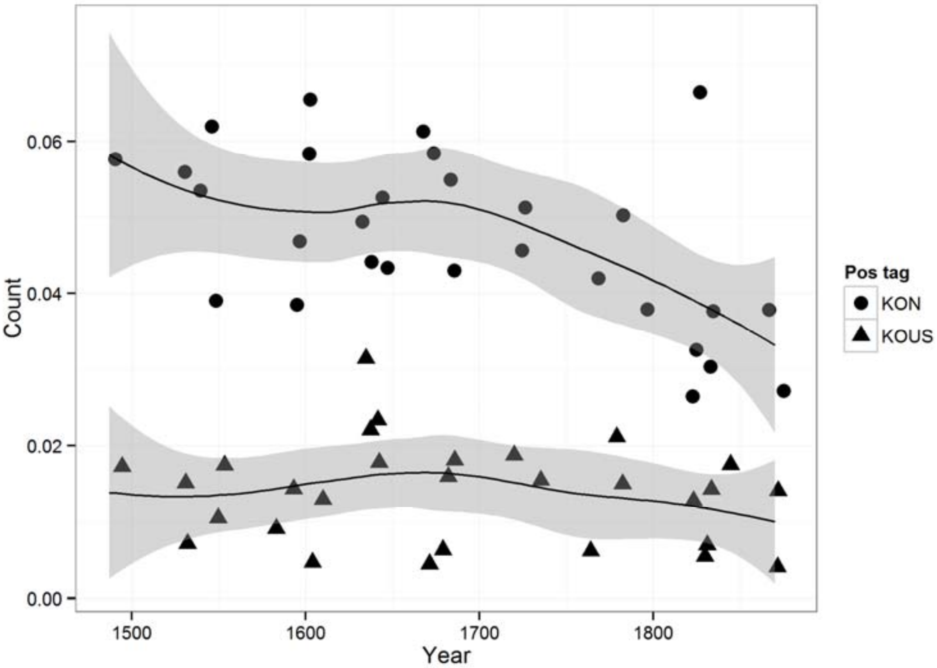


4.4 Part-of-speech Variants

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). Clausal coordination can be thought of as either coordinating conjunctions or subordinating conjunctions (cf. Ebert 1978; Hartweg & Wegera 2005). Coordination and subordination are marked with specific parts-of-speech. We are giving an example of two contrasting forms of one variable for clausal coordination: coordination (*pos* layer, tag: KON, e.g. *und* 'and'*, oder* 'or'*, aber* 'but') and subordination (*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 does not account for the specific coordination possibilities in Early Modern German, we used the *norm* layer as input for the tagger. Thus, we benefit from the normalization, as this layer merges difficult cases such as words separated by line breaks or differing spelling variants (see Section 3.2), which can then be allocated to a *pos* tag.

As Figure (14) 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 did not account for 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 positives. A further restriction for conclusions might be that KON also coordinates simple phrases like nominal or prepositional phrases, whereas KOUS only subordinates clauses. Thus, the context needs to be considered in further research.

Fig. 14 Frequencies of the *pos* tags KON and KOUS in RIDGES 4.1

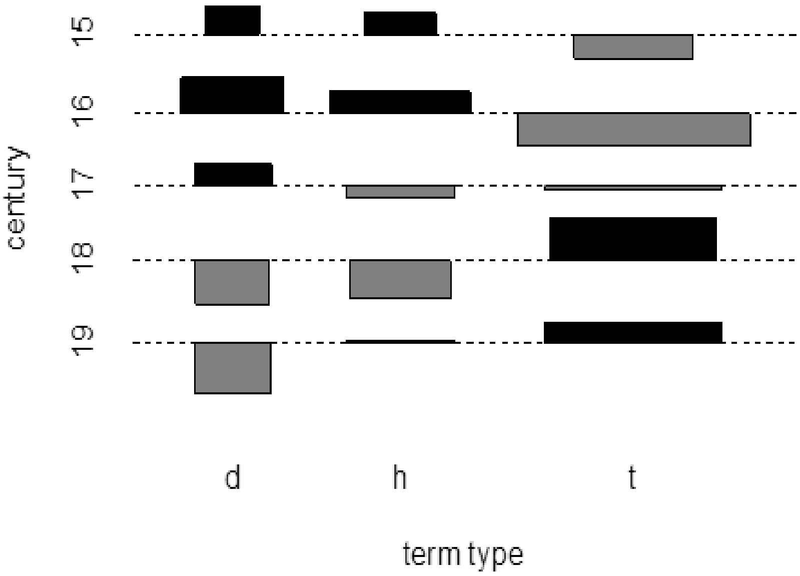


4.5 Expected Term Frequencies

The same principles of multi-layer architectures apply to research on the content of documents and not just to linguistic forms, i.e. to non-linguistic research as well. 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 (13) 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 term (*t*) in the *term* layer.

Figure (15) 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: 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. 15 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 an indigenous 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 multilayer 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 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. 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). 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.

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.

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 the Postgraduate Conference in Corpus Linguistics*, PCCL 2008. Birmingham. http://acorn.aston.ac.uk/conf_proceedings.html. Accessed 4 August 2015.

Bartsch, N., Dipper, S., Herbers, b., Kwekkeboom, S., Wegera, K.-P., Eschke, L., Klein, T., & Weber, E. (2011). Annotiertes Referenzkorpus Mittelhochdeutsch (1050-1350). In *Jahrestagung der Deutschen Gesellschaft für Sprachwissenschaft*, DGfS-CL Poster Session 2015. Göttingen.

Bentzinger, R. (2000). Die Kanzleisprachen. In W. Besche, 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: Walter 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: Walter 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.

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.

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.

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 *Von Daten zu Erkenntnissen. 2. Jahrestagung des Verbandes der Digital Humanities im deutschsprachigen Raum*, DHd 2015, Graz.

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.

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. *Traitment automatique des langues*, 49, 271-293.

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

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.

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

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 o.a.: Walter de Gruyter.

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

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*. Berlin, New York : de Gruyter.

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, 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. (2014). ANNIS3: A new architecture for generic corpus query and visualization. *Digital Scholarship in the Humanities*. http://dsh.oxfordjournals.org/cgi/content/abstract/fqu057?ijkey=GJBr0LhNfKW1g8i&keytype=ref. Accessed 4 August 2015.

Krause, T., Lüdeling, A., Odebrecht, C., & Zeldes, A. (2012) Multiple Tokenization 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, New York: 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 o.a.: Oxford University Press.

Kytö, M. (2011). Corpora and historical linguistics. *Revista Brasileira de Linguística Aplicada*, 11(2), 417-457. http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1984-63982011000200007&lng=en&tlng=en. 10.1590/S1984-63982011000200007. Accessed 04 August 2015.

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 Unter-suchung 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 o.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: Walter 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.

Pahta, P., & Taavitsainen, I. (2010). Scientific Discourse. In A. H. Jucker, & I. Taavitsainen (Eds.), *Historical Pragmatics* (pp. 549-586). Vol. 8. Berlin: 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 rechnergestutzte Variationsverarbeitung*. PhD Thesis. Universität Duisburg-Essen.

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: Walter de Gruyter.

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.

Richling, J. (2011). ‘Referenzkorpus Altdeutsch' (Old German Reference Corpus): Searching in Deeply Annotated Historical Corpora. In *New Methods in Historical Corpora Conference des Projekts GerManC*, 2011. Manchester.

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: Mouton 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. https://www.google.de/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&ved=0CCIQFjAAahUKEwiMyvGnjJfHAhWI3SwKHRKnCWA&url=http%3A%2F%2Fwww.sfs.uni-tuebingen.de%2Fresources%2Fstts-1999.pdf&ei=I7XEVcysPIi7swGSzqaABg&usg=AFQjCNGYlbMcTszTnXG0lbTYszoazGvhug&sig2=1CMBK8V3SkHGHUg0uY2W8w&bvm=bv.99804247,d.bGg&cad=rja. Accessed 7 August 2015.

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: Mouton 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: Walter de Gruyter.

Splett, J. (2000). Wortbildung des Althochdeutschen. In W. Besche, 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: Walter de Gruyter.

Springmann, Uwe & Lüdeling, Anke (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 a.o.: 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*, 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 a.o.: Gruyter.

Voigt, V. (2013) Phyton Script for the normalization layer *clean*. Documentation : http://korpling.german.hu-berlin.de/ridges/download/v4/cleanV2README.txt. Accessed 13 August 2015.

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

Zeldes, A., & Schroeder, C. T. (to appear). Computational Methods for Coptic: Developing and Using Part-of-Speech Tagging for Digital Scholarship in the Humanities. *Digital Scholarship in the Humanities*.

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.