Investigating Diatopic Variation in a Historical Corpus

Stefanie Dipper
Department of Linguistics
Ruhr-Universität Bochum
44780 Bochum, Germany
dipper@linguistics.rub.de

Sandra Waldenberger
German Department
Ruhr-Universität Bochum
44780 Bochum, Germany
sandra.waldenberger@rub.de

Abstract

This paper investigates diatopic variation in a historical corpus of German. Based on equivalent word forms from different language areas, replacement rules and mappings are derived which describe the relations between these word forms. These rules and mappings are then interpreted as reflections of morphological, phonological or graphemic variation. Based on sample rules and mappings, we show that our approach can replicate results from historical linguistics. While previous studies were restricted to predefined word lists, or confined to single authors or texts, our approach uses a much wider range of data available in historical corpora.

1 Introduction

In this paper we give an outline of our joint endeavor—combining computational and German historical linguistics—to develop a set of methods with the goal of uncovering and investigating the whole range of variation on the word level in a large scale corpus of historical texts. This is in contrast to traditional approaches in historical linguistics, who often use a predefined list of carefully-selected words for comparing linguistic variation.

In recent years, an increasing number of corpora of historical German has been built and published, including reference corpora of historical German, some still under construction (Donhauser, 2015; Klein et al., 2016; Schmitz et al., 2013; Peters and Nagel, 2014). Data from texts of historical and thus non-standard German is always strongly characterized by variation on every level of the language system. Hence, designing methods to gather and analyze the scope and scale of variation present in these corpora is a hot topic as well as a methodological challenge. Purely manual analysis is ruled out by the large amount of data provided by these corpora, necessitating the application of automatic methods.

We address the challenge of dealing with such data by way of systematic and exhaustive comparison of words that are variants of each other. To test and develop the comparative methods presented here we use the *Anselm* Corpus (Dipper and Schultz-Balluff, 2013).

The paper is organized as follows. Section 2 addresses prior work done in this area. In Section 3, we introduce the *Anselm* Corpus that we used in our comparison. Sections 4 and 5 present the comparison and its results, followed by a conclusion in Section 6.

2 Related Work

In recent years, spelling variation in non-standard data, such as historical texts or texts from social media, has come into focus in Natural Language Processing. Most often, variation is dealt with by normalization, i.e. mapping variants to some standard form (for historical data, see Piotrowski (2012, chap. 6)). The main focus of this research has been on how to automatize the normalization process, which is often a preparatory step to facilitate further processing of historical language data, e.g. by search tools or taggers (e.g. Jurish (2010), Bollmann (2012)). Some work addresses the extent of variance found in the data (e.g. Baron et al. (2009)). However, the derived mappings themselves that map historical to modern word forms are usually not in the focus of interest (but see Barteld et al. (2016)).

In contrast, theoretical linguists researching language evolution and language varieties are interested in these mappings, which highlight the

Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects, pages 36–45,
Valencia, Spain, April 3, 2017. ©2017 Association for Computational Linguistics

differences between the languages. Traditionally, historical linguistic research is mainly based on morphological and phonological properties. For instance, the relationships between the Indo-European languages have been established on the base of shared inflectional properties and phonetic relations, such as the first and second Germanic consonant shift. Similarly, dialect classification mainly depends on phonological and morphological features, with syntactic properties playing a minor role.

Language comparison in this spirit is based on specific language data: for sound-based comparison, lists of parallel words in different languages or language stages are usually used, such as the classical Swadesh list (Swadesh, 1955) or lists that have been compiled more recently for various languages (see, e.g., the data used in Jäger et al. (2017)). The challenge is then to identify related words, such as cognates and loan words, and unrelated words. The number of cognates between two languages serves as a measure of relatedness. In some approaches, no distinction is made between "real cognates", which are etymologically related, and words that are related due to some process other than strict inheritance.

In contrast to these approaches, we do not restrict our comparisons to single words from carefully-compiled word lists but aim at using as much data as possible from available corpora.

3 The Data

The data we use to test and refine our method has been extracted from the *Anselm* Corpus, which consists of about 50 versions of the medieval text *Interrogatio Sancti Anselmi de Passione Domini* ('Questions by Saint Anselm about the Lord's Passion'). The text is a dialogue between St. Anselm and the Virgin Mary, who recounts the events of the passion. The versions are from different language areas and time periods from Early New High German (1350–1600). Since they deal with the same topic, the overlap in content and vocabulary is large. Hence, the data provides a perfect basis for diatopic research. The map in Figure 1 gives an impression of the wide distribution of the different versions across the German language area.

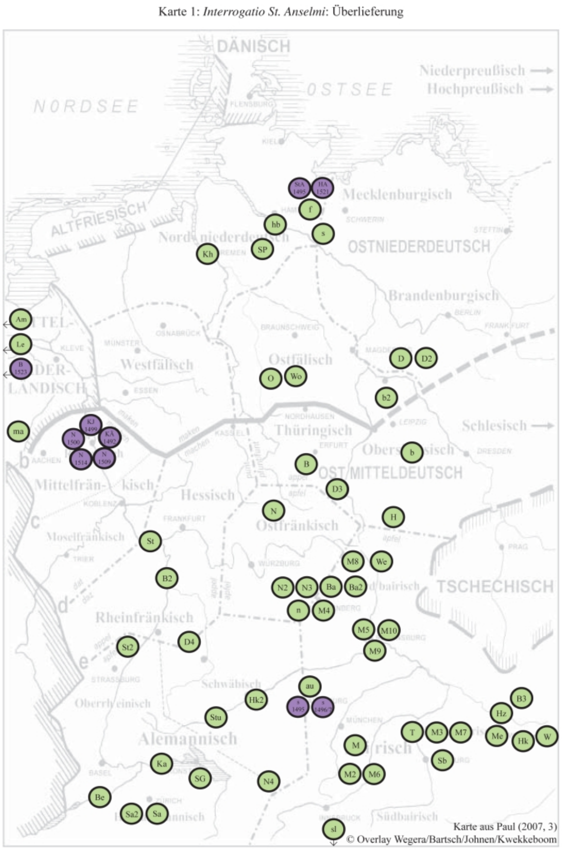

Figure 1: Distribution of the Anselm texts across the German-speaking area. Each marker represents one text (map taken from `https://www.linguistics.rub.de/anselm/corpus/map.html`).

Each word form in the *Anselm* Corpus has been manually annotated by its modern German translation (Bollmann et al., 2012). We define as *shared* or *equivalent* all historical word forms whose modern translations are identical. For instance, *vffston* in an Alemannic text and *vpstain* in a Ripuarian text are considered equivalent because they both correspond to modern German *aufstehen* 'stand up'. The investigations we present in this paper are based on such shared, equivalent word forms occurring in different texts.

Table 1 gives an overview of the temporal and regional distribution of shared words in the *Anselm* data.[1] The table shows that the *Anselm* Corpus has a good coverage of the 15th century, and that *mbair* is the best-documented language area.

We selected seven texts from different language areas for diatopic comparison. The comparison starts with texts written in the same language area

[1] *14* means '14th century', *14.1* means 'first half of the 14th century' (i.e. 1300–1350), etc. The language areas are: *alem*: Alemannic, *hchalem*: High Alemannic, *mbair*: Central Bavarian, *nbair*: North Bavarian, *obs*: Upper Saxon, *rhfrk*: Rhine-Franconian, *rip*: 'Ripuarian', *schwaeb*: 'Swabian'; *thuer*: 'Thuringian'.

			14	14.1	14.2	15	15.1	15.2	16	16.1
Upper German	East	nbair					4143:2	12840:7		
		mbair			2119:1		7995:4	9915:6	2236:2	1535:1
	West	alem	2166:1			1991:1				
		hchalem		2497:1				6653:2		1976:1
		schwaeb					2102:1	4404:4		
Central German	West	rip						3507:2		5510:3
		rhfrk						4203:2		
	East	thuer				1250:1		1713:1		
		obs			777:1					2039:1

Table 1: Temporal and regional distributions of shared, equivalent words (number of types) in the *Anselm* Corpus. The numbers after the colon represent the number of texts that have been compared.

	nbair		mbair		schw	rhfrk	
	M4	*Ba2*	*M3*	*B3*	*D4*	*St*	*B2*
M4	–	1572	2407	1856	1732	2058	2141
Ba2		–	1744	1475	1552	1614	1585
M3			–	1954	1884	2220	2315
B3				–	1611	1734	1765
D4					–	1865	1779
St						–	2300

Table 2: Number of pairwise shared words (types) for diatopic comparison, all texts dating from 15.2.

Text 1	Text 2	Normalization
M4: bedürfen	*M3*: pedurffen	bedürfen ‘require’
Ba2: pitten	*B3*: biten	bitten ‘ask’
St: uch	*D4*: aüch	euch ‘you (pl)’

Table 3: Examples of equivalent word forms.

(e.g. *mbair*) and proceeds with the comparison of adjacent language areas that belong to the same major dialect (*mbair* → *nbair*). Finally, texts from different dialects are compared, which are separated by the *Speyer line*, an isogloss separating the language areas called ‘Central German’ and ‘Upper German’ (*rhfrk* → *schwaeb*/*bair*).

Table 2 shows the overlap between the texts that we compared.[2]

4 Diatopic Comparison

As mentioned above, the diatopic comparisons are based on *equivalent* word forms. This section describes how these forms are found and how they form the base of comparison.

4.1 Finding Equivalent Word Forms

All original word forms in the *Anselm* Corpus have been manually normalized to the corresponding modern German word forms (Bollmann et al., 2012). All word forms with identical normalizations are considered equivalent.[3] For each pair of texts, equivalent word forms were collected and paired. Table 3 shows some sample pairs.

4.2 Deriving Rewrite Rules and Levenshtein-based Mappings

Similarities and differences between the equivalent word forms are modeled by means of ‘Rewrite rules’ and Levenshtein-based mappings (for detailed description and comparison of both methods, see Bollmann (2012)).

Rewrite rules Given a pair of equivalent word forms, both forms are first aligned at the character level, see (1a) which aligns the equivalent word forms *biten* and *pitten* (*bitten*, ‘ask’) (for details, see Bollmann et al. (2011)). (1b) is an alternative representation of the alignments. In the following,

[2] The identifiers used here (*M4* etc.) are the sigla of the individual texts, as defined in the *Anselm* corpus, see `https://www.linguistics.ruhr-uni-bochum.de/anselm/corpus/german.html` for a list. The sigla are derived from the respective repositories, e.g. *M4* is stored in Munich (Bayerische Staatsbibliothek).

[3] In order to diminish data sparseness, word forms with identical normalization but differing morphological properties were also considered equivalent, e.g. *chrawcze* (*M3*, mostly dative case) and *creucz* (*B3*, any case), both normalized as *kreuz* ‘cross’.

the format of (1b) is used in the presentation of examples.

(1) a.

B3	b	i		t	e	n
Ba2	p	i	t	t	e	n

b. |b=p|i=i|t=t|=t|e=e|n=n|

From these character alignments, rewrite rules are derived that replace characters from the first word to arrive at the second word. The word pair in (1) gives rise to the context-aware replacement rules shown in (2). '#' indicates word boundaries, 'E' represents the empty string.

(2) a. b → p | # _ i
"Replace word-initial 'b' by 'p', if followed by 'i'"

b. E → t | i _ t
"Insert 't' between 'i' and 't'"

In addition to the replacement rules, "identity rules" are derived, recording the characters that are identical in both word forms, see (3) and (4) for the identity rules derived from (1).[4]

(3) a. i → i | p _ t

b. t → t | t _ e

c. e → e | t _ e

d. n → n | e _ #

(4) a. E → E | # _ p

b. E → E | p _ i

c. E → E | t _ e

d. E → E | e _ n

e. E → E | n _ #

The rules derived from a text pair are collected and counted. Table 4 shows the top five identity and non-identity rules with their frequencies, as derived from the equivalent word forms of *B3* and *Ba2*. The interpretation of these rules is addressed below.

Freq	Rule
419	E → E | n _ #
312	E → E | e _ n
281	n → n | e _ #
265	E → E | t _ #
240	E → E | e _ r
26	c → E | # _ z
19	E → e | r _ n
17	n → E | a _ n
16	j → i | # _ o
14	j → i | # _ u

Table 4: Most frequent rewrite rules derived from *B3* → *Ba2* (top: identity rules; bottom: non-identity rules).

Weight	Seq 1 → Seq 2
0.125245	nn → n
0.195926	j → i
0.202549	ei → ai
0.220936	te → tte
0.227544	enn → en

Table 5: Least weighted mappings derived from *B3* → *Ba2*.

Levenshtein-based mappings Another way of modeling the relation between both word forms is by means of weighted Levenshtein-based mappings, which map character sequences of varying length. The more often a certain mapping has been observed in the data, the smaller its weight or cost. According to Levenshtein, identity mappings are the cheapest mappings with zero costs.

Some sample mappings derived from the example pair in (1a) are provided in (5). Table 5 shows the top five cheapest mappings derived from *B3* and *Ba2*.

(5) a. b → p

b. bi → pi

c. te → tte

d. t → tt

4.3 Interpreting the Rules and Mappings

The notation of the rules and mappings makes use of '→', implying that there is a directed relation between the two word forms, which takes one of the forms as the input and produces the other form

[4] The left context is checked against the target word form, the right context against the source form. The rules can also map sequences of characters, thus considering larger context. For details see Bollmann et al. (2011).
The rules in (4) prevent the insertion of characters at specific positions.

Rule	Analysis
c → E | # _ z	Graphemic variation: <cz> or <z> representing /ts/ in initial position
E → e | r _ n	Syncope (loss) of <e> representing /ə/ before final <n>
n → E | a _ n	<n> or <nn> representing /n/
j → i | # _ o	Graphemic variation: <j> or <i> in initial position
j → i | # _ u	

Table 6: Top non-identity rewrite rules derived from *B3* → *Ba2*, along with a linguistic analysis.

as the output. This interpretation may seem adequate for diachronic changes where we can say that the later form evolves out of the former form. For diatopic relations, a bidirectional interpretation seems more sensible, simply stating that a certain character (or character sequence) in one language area corresponds to another one in the other language area.

The (non-identity) rules and mappings often encode interesting relations, such as b → p, which indicates (de)voicing of plosives. In the next section, we go through a set of selected rules and mappings, discussing the range of phenomena that can be observed.

Ultimately we aim at using the rules and mappings for automatic clustering of texts relating to the crucial factors in language variation, language area and time, as well as other parameters—if they are included in the metadata the corpus provides—such as text type/function. Speaking from the perspective of historical linguistics, we hope to further enhance methodology by facilitating exhaustive analyses of larger corpora. Of course, this approach must be able to bear comparison to previous non-exhaustive approaches. It should be able to reflect previous, well-substantiated findings, such as the results of the High German consonant shift, but it should also be able to allow for new insights and eventually to draw a more detailed picture. The examples discussed in the next section were selected in a way to show that our approach will be able to satisfy both criteria.

5 First Results

Before discussing some results in detail, we would like to begin this section by giving an impression of how to interpret the replacement rules extracted by the method described above.

Table 6 gives linguistic analyses for the top non-identity rules of the pairing *B3* → *Ba2* (listed in Table 4).

The interpretation of the rewrite rules has to take into account which texts have been paired, in particular their spatial and temporal relation. In the example, we have paired two texts from the same period and the same area (Bavarian), but from different regions: *Ba2* is a North Bavarian text, and *B3* a Central Bavarian text, so we do not expect to see any diachronic variation here, and diatopic variation only to some extent.

The rules derived from the corpus show variants which are related to different levels of linguistic variation on the word level: to morphological, phonological and graphemic variation. To classify the rules as morphological, phonological or graphemic, the underlying word forms have to be consulted. As an example, see the list of 26 alignments in Table 7 that the rule in (6) has been derived from. The list of alignments shows all word forms starting with an inital affricate /ts/, which is encoded by <cz> in *B3* on the one hand and by <z> in *Ba2* on the other hand. As can be seen, the graphematic variation <cz>/<z> concerns a variety of different lemmas but becomes visible as a pattern through the rewrite rule.

(6) *B3* → *Ba2*: c → E | # _ z

In some cases (9 instances), *Ba2* also uses <cz>, like *B3*, triggering an identity rule, (7).

(7) *B3* → *Ba2*: c → c | # _ z

Morphological variation When pairing a Central German text, *St* (from the Rhine-Franconian area (Mainz)) with any of the Bavarian Upper German texts (*Ba2*, *M4*, *M3*, *B3*) from the same time period—the latter in order to rule out diachronic variation—the rule shown in (8) sticks out in all comparisons, see Table 8. To give an impression of the type of rules and their frequencies that have been derived, the table provides the three top-ranked (non-identity) rules for each pairing.

(8) t → E | n _ #

Alignments	Lemma
|c=|z=z|e=e|c=|h=h|e=e|r=r|=e|n=n| |c=|z=z|e=e|c=c|h=h|e=e|r=r|=e|n=n|	*zeheren* ‘(to) weep’
|c=|z=z|a=a|r=r|t=t|e=e|n=n|	*zart* ‘sweet’
|c=|z=z|u=u|h=h|a=a|n=n|t=t| |c=|z=z|u=u|h=h|a=a|n=n|=d|t=t| |c=|z=z|u=u|h=h|a=a|n=n|t=d| |c=|z=z|u=u|h=h|a=a|n=n|n=|t=t| |c=|z=z|u=u|h=h|a=a|n=n|n=d|t=t| |c=|z=z|u=u|h=h|a=a|n=n|n=|t=d|	*zuhand* ∼ *zehant* ‘at once’
|c=|z=z|e=e|h=h|e=e|n=n|	*zehn* ‘ten’
|c=|z=z|e=a|i=i|c=c|h=h|e=e|n=n|	*Zeichen* ‘sign’
|c=|z=z|e=a|i=i|g=g|e=e|n=n|	*zeigen* ‘(to) show’
|c=|z=z|e=e|i=i|t=t|	*Zeit* ‘time’
|c=|z=z|u=e|s=r|c=|h=s|l=l|a=a|h=g|e=e|n=n|	*zerschlagen* ‘(to) break’
|c=|z=z|u=e|s=s|p=p|i=i|=e|l=l|t=t|	*ze(r)spalten* ‘(to) split’
|c=|z=z|e=e|r=r|s=s|t=t|o=e|r=r|e=e|r=r|	*Zerstörer* ‘destroyer’
|c=|z=z|u=u|g=c|k=h|t=t| |c=|z=z|u=o|g=c|k=h|t=| |c=|z=z|o=u|c=c|h=h|=t| |c=|z=z|o=o|c=c|h=h| |c=|z=z|u=u|g=g|e=e|n=n|	*ziehen* ‘to pull’
|c=|z=z|o=o|r=r|=e|n=n|	*Zorn* ‘anger’
|c=|z=z|u=u|	*zu* ‘to’
|c=|z=z|w=w|u=e|=n| |c=|z=z|w=w|u=o|	*zwen* ∼ *zwo* ‘two’
|c=|z=z|e=e|r=r|t=t|l=l|i=i|c=c|h=h|e=e|n=n|	*zertlich* ‘gentle’

Table 7: All 26 alignments underlying the replacement rule in (6).

Text pair	Rule	Freq
St → *Ba2* (nbair)	t → E | n _ #	51
	e → E | t _ #	33
	d → t | # _ o	25
St → *M4* (nbair)	t → E | n _ #	45
	e → E | t _ #	34
	e → E | d _ #	27
St → *M3* (mbair)	t → E | n _ #	47
	E → h | c _ r	40
	E → e | l _ i	31
St → *B3* (mbair)	t → E | n _ #	44
	e → E | t _ #	32
	i → E | o _ s	24

Table 8: Three top-ranked replacement rules, as derived from pairing a Central German text (*St*) with different Upper German texts.

This rule is triggered mainly by varying inflectional verb forms, such as *gaben* vs. *gabent* '(they) gave', *haben* vs. *habent* '(they) have', *kommen* vs. *komment* '(they) come', *glauben* vs. *glaubent* '(they) believe', etc.

Rule (8) reflects a well-known case of diatopic morphological variation in the Early New High German period: Upper German strongly tends towards *-ent* as inflectional marker for plural verb forms, whereas Central German prefers *-en* (Dammers et al., 1988, §74ff.).

The Levenshtein-based mappings confirm the picture. Table 9 shows the top mappings for three of the pairings in Table 8. Only with the pairing *St* → *B3* (mbair), there is no respective mapping among the top-ranked ones.

B2 is another text from the Rhine-Franconian area but has been located further south than *St* (see Figure 2). If *B2* is paired with the same Upper German texts (*Ba2*, *M4*, *M3*, *B3*), the results do not contain rule (8) at all, or their frequency is much lower. This also reflects the findings presented in Dammers et al. (1988, §76ff.), who show that the distribution of the variants *-ent* vs. *-en* does not coincide completely with the isoglosse(s) separating Upper from Central German, and *-ent* is instead common farther to the north.

These examples show that the method proposed in this paper is able to confirm results of previous research, i.e. it is possible to derive constraints on the localization of these texts by means of their 'linguistic footprint' as mirrored in these rules.

Text pair	Se1 → Seq2	Weight
St → *Ba2* (nbair)	y → i	0.136881
	yn → in	0.155339
	nt → **n**	0.167918
	d → t	0.171744
St → *M4* (nbair)	y → i	0.13741
	yn → in	0.15568
	yn → ein	0.194489
	nt → **n**	0.213094
St → *M3* (mbair)	cr → chr	0.117811
	b → p	0.146198
	nt → **n**	0.161911
	ent → **en**	0.168727

Table 9: Top four Levenshtein-based mappings of the Central text *St* with three texts from Upper German. The mappings corresponding to the replacement rule t → E | n _ # have been highlighted.

Phonological variation We next look at a rule that is related to the High German consonant shift, see (9).

(9) *St* → *D4*: E → f | #p _ e
example: *penning* vs. *pfenni(n)g*

The rule in (9)[5] has been derived from pairing *St* with a Swabian text, *D4*. *D4* is a borderline case, i.e. located on the border between Upper and Central German, which is indicated by the isoglosse called *Germersheim Line*. This line marks the shift of Germanic /p/ to affricate /pf/ in initial position, see Figure 2. Rule (9) locates *D4* south of the *Germersheim Line*.

Another example of phonologically-based variation is the rule in (10)[6]. This rule clearly identifies *St* as a Rhine-Franconian text, showing /d/ instead of Upper German /t/ in initial or medial position, see Table 10.

(10) *St* → *D4*: t → d | # _ o
examples: *tochter* vs. *dochter*, *todes* vs. *dodes*

Graphemic variation The above examples confirmed results already known from the literature. The next examples illustrate that our new method also enables us to refine the picture of historical

[5] Rule rank: 30; rule frequency: 8.

[6] Rule rank: 8; rule frequency: 16.

<table>
<tr><th>Freq</th><th>Rule</th><th>Phonemes</th></tr>
<tr><td>26</td><td>E → c | # _ z</td><td>Initial affricate /ts/, <cz> vs. <z>
e.g. <i>czehen</i> vs. <i>zehen</i>; <i>czu</i> vs. <i>zu</i></td></tr>
<tr><td>15</td><td>u → ü | f _ r</td><td>Umlaut vowels with or without trema <¨>
e.g. <i>für</i> vs. <i>fur</i>; <i>fürst</i> vs. <i>furst</i>; <i>füren</i> vs. <i>furen</i></td></tr>
<tr><td>13</td><td>t → E | d _ #</td><td>Final alveolar stop <dt> vs. <t>
e.g. <i>gesundt</i> vs. <i>gesund</i>; <i>kindt</i> vs. <i>kind</i></td></tr>
<tr><td>11</td><td>z → E | s _ #</td><td>Final alveolar fricative <sz> or <s>
e.g. <i>bisz</i> vs. <i>bis</i>; <i>dasz</i> vs. <i>das</i>; <i>schosz</i> vs. <i>schos</i></td></tr>
</table>

Table 11: Selected rules and their frequencies, as derived from *Ba2* → *M4*, both from North Bavaria, along with a description of the phonemes that are represented by the respective graphemes.

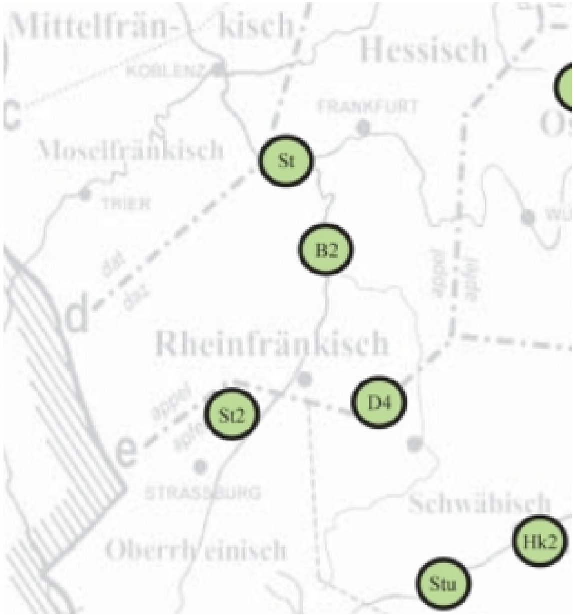

Figure 2: Localization of the texts. The *Speyer Line* is indicated by the letter 'e' on the left side. It coincides with the *Germersheim Line* in the Western part of the German language area. North of this line, Germanic */p/ is retained, south of the line, /pf/ is used instead.

Text pair	Freq	Rank
St → *Ba2* (nbair)	25	3
St → *M4* (nbair)	16	12
St → *M3* (mbair)	21	8
St → *B3* (mbair)	14	13

Table 10: Absolute frequencies and ranks of the replacement rule t → d | # _ o, as derived from pairing a Central German text (*St*) with different Upper German texts.

variation, especially when it comes to graphemic variation.

Suitable examples come from pairing neighboring texts, e.g. *Ba2* → *M4*, two texts from North Bavaria (*nbair*). This pairing generates rules which correspond mainly to graphemic variation, in contrast to pairings of different language areas, as in the previous section, see the examples in Table 11.

In applying the method proposed in this paper systematically and exhaustively, a highly nuanced picture of graphemic variation will become observable. In systematically assessing the replacement rules derived from a balanced corpus of historical texts we hope to be able to ascertain a complete picture of graphemic variation, i.e. which variants were available and were preferred by scribes in different areas.

6 Conclusion

We hope that our approach will help filling research gaps in historical linguistics. Thus far, research had to cope with a lack of corpora on the one hand, and the restrictedness of retrieval methods on the other hand. Therefore, previous studies in historical graphematics were inevitably restricted and of a merely exemplary nature.

Of course, the exemplary segments these studies have been focussing on—copies of one text in Glaser (1985), texts by one and the same author in Wiesinger (1996), German prints of the bible translated by Martin Luther in Rieke (1998), texts originating from one scribal office in Moser (1977) and texts from one specific place (Duisburg) in Mihm (2004) and Elmentaler (1998; 2001; 2003)—have been selected applying expedient criteria. The studies have been able to provide insight into a very small, if significant area, leaving the rest of the map to remain blank.

This is where our approach comes in. The characteristics of the bundle of methods described above is that we aim at capturing the whole range of variation documented in historical corpora, and that we do so by 'joining forces' and mustering expertise from NLP as well as from German historical linguistics. In this way we make sure that the results delivered by the computational methods fit the requirements of actual variation analysis and are therefore to be considered not only usable, but beneficial for future corpus-based historical linguistics. Our approach will be applicable to corpora with a normalization layer—which is the case for the reference corpora of historical German.

As Table 1 shows, the *Anselm* Corpus does not allow for comprehensive diachronic analyses. When applied to a corpus which covers a larger time period than the *Anselm* Corpus, we expect the proposed method to discover both diachronic and diatopic variation. Language change never occurs as a sudden change or replacement of one variant by the other but involves a period of coexistences of multiple variants. Hence, language change will become visible as changes in frequency of the variants involved (cf. Wegera and Waldenberger (2012, 25)), starting out with an increasing number of instances of the new variant and—if the process is successful—resulting in a decrease of the older variant. Such changes in frequency will translate into the rewrite rules generated by our method, specifically into the ratio between non-identity rules and their corresponding identity rules.

Acknowledgments

We would like to thank the anonymous reviewers for helpful comments. The research reported in this paper was supported by Deutsche Forschungsgemeinschaft (DFG), Grant DI 1558/4.

References

Alistair Baron, Paul Rayson, and Dawn Archer. 2009. Word frequency and key word statistics in historical corpus linguistics. *Anglistik: International Journal of English Studies*, 20(1):41–67.

Fabian Barteld, Ingrid Schröder, and Heike Zinsmeister. 2016. Dealing with word-internal modification and spelling variation in data-driven lemmatization. In *Proceedings of the 10th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH)*, pages 52–62.

Marcel Bollmann, Florian Petran, and Stefanie Dipper. 2011. Rule-based normalization of historical texts. In *Proceedings of the RANLP-Workshop on Language Technologies for Digital Humanities and Cultural Heritage*, pages 34–42, Hissar, Bulgaria.

Marcel Bollmann, Stefanie Dipper, Julia Krasselt, and Florian Petran. 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) (LThist2012), KONVENS*, Wien, Austria.

Marcel Bollmann. 2012. (Semi-)automatic normalization of historical texts using distance measures and the Norma tool. In *Proceedings of the Workshop on Annotating Corpora for Research in the Humanities (ACRH-2)*, Lisbon.

Ulf Dammers, Walter Hoffmann, and Hans-Joachim Solms. 1988. *Grammatik des Frühneuhochdeutschen 4: Flexion der starken und schwachen Verben*. Winter, Heidelberg.

Stefanie Dipper and Simone Schultz-Balluff. 2013. The *Anselm Corpus*: Methods and perspectives of a parallel aligned corpus. In *Proceedings of the NODALIDA Workshop on Computational Historical Linguistics*, pages 27–42, Oslo, Norway.

Karin Donhauser. 2015. Das Referenzkorpus Altdeutsch: Das Konzept, die Realisierung und die neuen Möglichkeiten. In Jost Gippert and Ralf Gehrke, editors, *Historical Corpora. Challenges and Perspectives*. Narr, Tübingen.

Michael Elmentaler. 1998. Die Schreibsprachgeschichte des Niederrheins. Ein Forschungsprojekt der Duisburger Universität. In Dieter Heimböckel, editor, *Sprache und Literatur am Niederrhein*, pages 15–34. Pomp, Bottrop.

Michael Elmentaler. 2001. Der Erkenntniswert der schreibsprachlichen Variation für die Sprachgeschichte. Überlegungen zu den Erkenntnissen eines Duisburger Graphematikprojektes. *Rheinische Vierteljahrsblätter*, 65:290–314.

Michael Elmentaler. 2003. *Struktur und Wandel vormoderner Schreibsprachen*. de Gruyter, Berlin, New York.

Elvira Glaser. 1985. *Graphische Studien zum Schreibsprachwandel vom 13. bis 16. Jahrhundert. Vergleich verschiedener Handschriften des Augsburger Stadtbuches*. Winter, Heidelberg.

Bryan Jurish. 2010. More than words: Using token context to improve canonicalization of historical German. *Journal for Language Technology and Computational Linguistics*, 25(1):23–39.

Gerhard Jäger, Johann-Mattis List, and Pavel Sofroniev. 2017. Using support vector machines and state-of-the-art algorithms for phonetic alignment to identify cognates in multi-lingual wordlists. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*.

Thomas Klein, Klaus-Peter Wegera, Stefanie Dipper, and Claudia Wich-Reif. 2016. Referenzkorpus Mittelhochdeutsch (1050–1350), Version 1.0. https://www.linguistics.ruhr-uni-bochum.de/rem/. ISLRN 332-536-136-099-5.

Arend Mihm. 2004. Zur Neubestimmung des Verhältnisses zwischen Schreibsprachen und historischer Mündlichkeit. In Franz Patocka and Peter Wiesinger, editors, *Morphologie und Syntax deutscher Dialekte und historische Dialektologie des Deutschen. Beiträge zum 1. Kongress der Internationalen Gesellschaft für Dialektologie des Deutschen, Marburg/Lahn, 5.-8. März 2003*, pages 340–382, Wien. Praesens.

Hans Moser. 1977. Die Kanzlei Kaiser Maximilians I. Graphematik eines Schreibusus. Univ. Innsbruck.

Robert Peters and Norbert Nagel. 2014. Das digitale 'Referenzkorpus Mittelniederdeutsch / Niederrheinisch (ReN)'. In Vilmos Ágel and Andreas Gardt, editors, *Paradigmen der Sprachgeschichtsschreibung*. de Gruyter, Berlin, Boston.

Michael Piotrowski. 2012. *Natural Language Processing for Historical Text*. Morgan & Claypool.

Ursula Rieke. 1998. *Studien zur Herausbildung der neuhochdeutschen Orthographie. Die Markierung der Vokalquantitäten in deutschsprachigen Bibeldrucken des 16.-18. Jahrhunderts*. Winder, Heidelberg.

Hans-Christian Schmitz, Bernhard Schröder, and Klaus-Peter Wegera. 2013. Das Bonner Frühneuhochdeutsch-Korpus und das Referenzkorpus 'Frühneuhochdeutsch'. In Ingelore Hafemann, editor, *Perspektiven einer corpusbasierten historischen Linguistik und Philologie. Internationale Tagung des Akademienvorhabens "Altägyptisches Wörterbuch" an der Berlin-Brandenburgischen Akademie der Wissenschaften, 12.–13. Dezember 2011*.

Morris Swadesh. 1955. Towards greater accuracy in lexicostatistic dating. *International Journal of American Linguistics*, 21:121–137.

Klaus-Peter Wegera and Sandra Waldenberger. 2012. *Deutsch Diachron. Eine Einführung in den Sprachwandel des Deutschen*. Erich Schmidt, Berlin.

Peter Wiesinger. 1996. *Schreibung und Aussprache im älteren Frühneuhochdeutschen. Zum Verhältnis von Graphem – Phonem – Phon am bairisch-österreichischen Beispiel von Andreas Kurzmann um 1400*. de Gruyter, Berlin.