Sandra Waldenberger*, Stefanie Dipper, and Ilka Lemke

Towards a broad-coverage graphemic analysis of large historical corpora

<https://doi.org/10.1515/zfs-2021-2037>

Abstract: This paper presents a method which we are developing to explore graphemic variation in large historical corpora of German. Historical corpora provide an amount of data at the level of graphemics which cannot be handled exhaustively using common methods of manual evaluation. To deal with this challenge, we apply methods from computational linguistics to pave the way for a broad-coverage graph(em)ic analysis of large historical corpora. In this paper, we show how our approach can be applied to the Reference Corpus of Middle High German. Illustrating our method and linguistic analysis, we present findings from our investigations into diatopic and/or diachronic variation as documented in 13th and 14th century charters (Urkunden) from the corpus.

Keywords: graphemic variation, Middle High German, corpus-based analysis, quantitative analysis

1 Introduction

The methods we present in this paper answer the call for semi-automatic means to analyze graphemic variation in historical texts (cf. Elmentaler 2018: 335). As the linguistic level of graphemics means that we are dealing with the smallest linguistic units, the amount of data provided by a corpus is the largest on this level. On the other hand, the graphemic level provides data sets that consist, on a basic level, of nothing else than character strings, which can be processed automatically. We use this fact to our advantage: The computational linguistic methods that we use are based on methods developed for normalizing historical spellings, i. e., for automatically mapping a historical spelling variant to a standardized (historical or modern) form (see, e. g., Jurish 2011; Pettersson 2016; Boll-

***Corresponding author: Sandra Waldenberger**, Germanistisches Institut, Ruhr-Universität Bochum, Universitätsstr. 150, 44801 Bochum, Germany, e-mail: sandra.waldenberger@rub.de
Stefanie Dipper, Sprachwissenschaftliches Institut, Ruhr-Universität Bochum, Universitätsstr. 150, 44801 Bochum, Germany, e-mail: stefanie.dipper@rub.de
Ilka Lemke, Westfälische Wilhelms-Universität Münster, Germanistisches Institut, Abteilung Sprachwissenschaft, Schlossplatz 34, Raum: VSH 146, 48143 Münster, Germany, e-mail: ilemke@uni-muenster.de

mann 2018; Mittmann 2020¹). The automatic systems learn which original character sequences typically correspond to which character sequences in the standardized form. For instance, historical word-final *-ey* typically corresponds to *-ei* in modern German. We call these correspondences ‘mappings’. In the current paper, the mappings are learned from mapping one historical form (from variety 1) to another historical form (from variety 2), and therefore highlight typical spelling variations between the two varieties. For instance, word-initial *ko-* from variety 1 might correspond to *cho-* in the other variety (as in *chomen* vs. *komen* ‘come’). These mappings form the basis for our graphemic investigations.

In Dipper and Waldenberger (2017), we applied the described methodology for the first time and examined mappings that were derived from a parallel corpus containing texts of different dialects from Early New High German, with large overlaps in vocabulary. The results of this pilot study were promising in that relevant variants could be automatically identified. As the next step, we now apply the same method to a large corpus of historical texts, namely the Reference Corpus of Middle High German (Referenzkorpus Mittelhochdeutsch, short: ReM, cf. Klein et al. 2016). Each token in ReM is annotated with a standardized form (Klein and Dipper 2016: 8) so that we can derive mappings from ReM. We begin our investigations with the corpus texts labeled ‘Urkunde’ in ReM (for a complete list, see the Appendix) and present results from these investigations in this paper. Charters (Urkunden) are a perfect starting point for investigations into diatopic and diachronic variation since they can be accurately dated and localised. This allows us to focus on one factor (time or space) at a time and, thus, reduce the parameters of variation to a minimum. Furthermore, charters from the Middle High German period are an interesting source for intra-personal graphemic variation, as models of the evolution of Middle High German writing systems have to build on models of individual writing systems: The individual charters of ReM each consist of a series of legal documents which are written by one single scribe. Hence, the charters from ReM allow us, in essence, to observe the writing habits of one scribe within a consistent setting.

We would like to point out that we do not intend our present study to reveal new insights into graphemic variation during the MHG period at this point of our

¹ Mittmann (2020) deals with a topic distantly related to ours: Mittmann exploits known variation regularities and manually translates them into replacement rules which he applies to the standardized forms. He thus artificially generates dialect-specific word forms and checks which of the word forms actually occur in the text. In our approach, the variation regularities do not need to be known in advance and the substitution rules are learned automatically in a data-driven way. In contrast to Mittmann (2020), our approach can discover new regularities. – We are not aware of any approaches comparable to ours.

work. Rather, we want to find out whether the methods presented in this study are able to replicate what has already been shown in past research in the area of historical graphemics of German. That is, in the current study, we want to show that the semi-automatic approach we are taking is able to deliver sound results. On the long run, our goal is to refine the maps and timelines of graphemic variation depicted in historical graphemics so far. In order to get there, we will – as part of a research project on a larger scale – conduct detailed and exhaustive investigations of texts taken from ReM, in order to densely trace the net of graphemic variation given in the Middle High German period. In this study, we give but a mere glimpse of what our method is able to bring to light.

This paper is organized as follows: Section 2 presents the methods from computational linguistics we apply to the Reference Corpus ReM to get pre-analysed data sets;² we call those data sets ‘difference profiles’. The data sets provide the basis for qualitative analysis and interpretation of the graphemic variation that occurs in the charters of ReM, as we will show exemplarily in Section 3. Section 4 presents a quantitative analysis, with the goal of providing additional evidence for the qualitative observations, followed by a conclusion in Section 5.

2 Generating difference profiles

The present study builds on the approach introduced in Dipper and Waldenberger (2017). This approach uses pairs of so-called “equivalent word forms”, i. e., word forms whose standardized forms in ReM are identical, like the word forms shown in (1). We use the tool Norma to automatically align corresponding letters and letter sequences from the paired word forms (2) (for details, see Bollmann et al. 2011).

- (1) pair: *schlaht* (M345) – *flachte* (M348)
(common standardized ReM form: *slahte* ‘battle’)

- (2) Alignments:

M345	f	c	h	l	a		h	t	
M348	f			l	a	c	h	t	e

Using Norma, we derive patterns of correspondences from the alignments in two ways. First, we derive mappings of sequences of one to four characters based on weighted Levenshtein distance (Levenshtein 1966), a measure widely used in computational linguistics. Levenshtein (1966) proposed a method to calculate the dis-

² The data can be downloaded from: <https://github.com/rubcompling/zs2021>

tance between two strings (e. g., two words) by determining how many characters must be changed to map one string (like *fchlaht* in (2)) to the other (like *flachte*). Three different operations are possible: (i) a letter is deleted (in (2), the first ‘h’ of M345); (ii) a letter is inserted (in (2), the ‘c’ in M348); (iii) a letter is replaced (i. e., deletion and insertion together; no instance in (2)). Each operation “costs” 1 point, so the mapping in (2) would cost 4 points (2 deletions + 2 insertions). The more operations are necessary, the higher the score. A high score indicates a large distance and few similarities between the two strings. Weighted Levenshtein distance (WLD) distinguishes between differently “plausible” operations. For example, the substitution of ‘i’ by ‘j’ in historical word pairs is a very common operation, whereas the substitution of ‘i’ by ‘b’ should be rather rare. With WLD, common operations, which have been observed multiple times in the data, cost less than rare operations. Hence, cheap mappings indicate highly typical (frequent) writing differences between two texts.³ The WLD mappings we get from Norma not only map single characters but sequences of up to four characters (e. g., replace ‘fch’ by ‘f’).

For example, the alignments in (2) result in the mappings shown in (3), among others.⁴ The symbol ‘#’ indicates word boundaries, i. e., the beginning or end of a word. The mapping from ‘t#’ (M345) to ‘te#’ (M348) indicates an apocope (final ‘e’ has been deleted). According to the weights, this is the ‘cheapest’ mapping, i. e.,

3 The current study methodologically extends the one by Dipper and Waldenberger (2017) in that the Levenshtein mappings now include word boundaries. Since phonological and morphological changes frequently occur in the onset or in inflectional suffixes, explicit marking of word boundaries is essential. Two other extensions of the current study were: First, we experimented with deriving the patterns not only from types, but also from tokens. Type and token here refer to word pairs such as <fchlaht, flachte>. If, for example, such a word pair occurred very frequently, this would have an impact in the (alternative) token-based approach because the frequencies would increase accordingly, but it would have no impact in the (original) type-based approach because it is just one type. However, the results turned out rather not useful for the current study. The texts analysed in this study are highly formulaic and therefore contain a lot of reiterations (e. g., invocations and legal wording) which have a strong effect on a token-based approach. Due to their high frequencies, these idiosyncratic spellings (e. g. idiosyncratic capitalisations as in *Jngesiegel* ‘seal’; *IN/GODEs/NaMEN/aMEN* ‘in God’s name’; *Jnde* ‘and’ in M349) result in cheap mappings in the token-based approach, and would give the incorrect impression that these mappings indicate typical, general writing differences whereas in reality they are due to the text type specific formulaic language. Second, we experimented with Levenshtein-based mappings of character 5-grams instead of 4-grams. However, the results were not distinct from the approach using 4-grams.

4 The notation of the correspondences is directional, that is, mapping one form to another. Of course, this does not imply a corresponding ‘influence’ of one form on the other or the like but is a notation for (non-causal) correspondences.

it hints at a rather typical difference between these two texts. The mapping ‘fchl → fl’ is considerably more expensive.

(3) Mappings:

Text 1 (M345)	→ Text 2 (M348)	Weight
t#	te#	0.279632
t	Te	0.386421
ht	chte	0.502560
ht	hte	0.502560
fchl	fl	0.621775
#fch	#f	0.622387
fc	f	0.624207

Second, we derive replacement rules from the alignments, see (4) for some examples. For instance, the first rule ‘E → e / t_#’ specifies that ‘E’ is replaced by ‘e’ if it occurs between ‘t’ and ‘#’. ‘E’ is a special character and represents the empty string, ‘#’ indicates the word boundary. So, the first rule effectively inserts ‘e’ after word final *t*, see (1) and (2). The rules are ranked according to their absolute frequencies.⁵

(4) Rules:

Frequency	Rule
12	E → e / t_#
3	E → c / a_h
1	ch → E / f_l

Mappings and rules basically encode the same information, with small differences: The rule format highlights the differences between the two strings in that the left-hand-side of the rule shows the sequence from one variety and the right-hand-side of the rule shows the corresponding sequence from the other variety. This format is particularly transparent and easy to read and therefore especially suitable for qualitative analysis. Furthermore, rules are in general rather specific due to the context constraints and possibly miss certain generalizations. The mappings as implemented in Norma are more general and more flexible than rules with respect to length, mapping sequences of 1–4 characters. They are therefore better suited for the statistical analysis that is applied in Section 4. In other words: rules are well suited for qualitative investigations and mappings for quantitative ones. To give an example: Norma can produce complex mappings like ‘t# → te#’

⁵ We use here and below the following representation for the rules: first the frequency of the rule is specified, then the rule itself.

but also simple ones like ‘ $t \rightarrow te$ ’. Translated into rules, the complex mapping corresponds to the rule ‘ $E \rightarrow e/t_ \#$ ’ but there is no rule equivalent to the simple mapping because there is not enough context information.

Each pair of texts whose spellings we compare using this method yields its own characteristic mappings, depending on the type of differences (e. g., diatopic or diachronic, closely related spellings or not). We call the set of mappings that we derive from a text pair a ‘difference profile’ because it highlights the characteristic differences between these texts.

3 Interpreting difference profiles

The difference profiles generated by the methods described above allow for in-depth linguistic analyses of graphemic variation. We selected text pairings in such a way as to generate difference profiles along the diatopic as well as the diachronic dimension of the ReM (see the examples given in this section). The difference profiles show the range of variation documented in the paired texts and, consequently, the degree of graphemic distance between them.

Analyzing the difference profiles between two corpus texts mainly consists in categorizing the data (rules or mappings and their rankings) derived from the text pairings. We distinguish between two main categories: graphophonemic variation and graph(em)ic variation. The criteria we apply concern the type of character involved (capital or lower-case letters, abbreviations, graphic variants) and the position of the letter/grapheme. Of course, we do not interpret the rules and rankings without verifying the source word forms they are derived from. If there is sufficient evidence that the variation is linked to the phonemic level, i. e., if a difference in spelling is linked to a difference in the underlying phonemic systems, we infer the category graphophonemic variation. For this purpose, we draw on results from previous research.

As we show below, pre-analysed data in the form of difference profiles allows for a detailed and precise description of graphemic variation on different levels and ranges of variation corresponding to language areas and time periods. This variation spreads out along a continuum between the two poles graphemic and graphophonemic variation, which we illustrate in the following paragraphs.

Graphophonemic variation has been the main focus of traditional research on both diachronic and diatopic variation. To classify the variants at hand, it has to be established if they are either a later form that has evolved out of a previous form or if the cause for the variation is an underlying phonological variation that stems from characteristic diatopic differences. The diachronic and diatopic

dimension converge when considering ongoing processes with different spreading in the MHG dialects, e. g. diphthongization. Graphophonemic phenomena include among many others:

- High German consonant shift (‘Zweite Lautverschiebung’), e. g. *kain* vs. *cheȳn* in M345–M348
- apocope, the loss of word-final <e> representing unstressed vowel /ə/, e. g. *gerihtē* vs. *Geriht* in M344–M345
- syncope, the loss of /ə/ from the interior of a word, e. g. *Svns* vs. *funes* in M345–M347
- final obstruent devoicing (‘Auslautverhärtung’) and its graphemic representation, e. g. <k> and <g> in *ledik* vs. *ledig* in M344–M345
- New High German diphthongization, the combination of two adjacent vowels within the same syllable, e. g. *haufe* vs. *hufe* in M345–M347

Writing systems in Middle High German texts do not encode phonemic change as such. Instead, this needs to be hypothetically reconstructed from phonemic processes, which in turn usually have to be inferred from their graph(em)ic realizations (Wegera et al. 2018: 102). We address the challenge of tracing both phonemic and graphemic changes by comparing the variants in a corpus in an exhaustive and linguistically-informed way.

Within writing systems, scribes show preferences in selecting specific characters to encode certain phonemes, sometimes by additionally marking features such as length and brevity. In our texts, these preferences may result in character-based variation such as:

- different s-spellings <s>, long-s <f> or <z>, e. g. *des*, *def* or *dez* in M345 in word-final position. In addition, there are form confusions possible between capitalized S and the different forms of the s-spellings, e. g. *Swie* vs. *fwie* in M345–M347
- <v> or <f> in word-initial position, e. g. *fūmf* vs. *vūnf* in M345–M347
- <j>, <y> or <i>, e. g. *in* vs. *yn* and *drj* vs. *drī* in M344–M345
- affricate, e. g. <ch> or <h> representing /χ/, e. g. *acht* vs. *ah̄t* in M352–M352
- <v>, <u> representing either consonants (labiodental fricative /f/) or vowel /u/, e. g. in *vnd* vs. *und* and *Zu* vs. *zv* in M352–M353
- double vs. single consonants, e. g. <tt> vs. <t>, *Gerih̄t* vs. *Gerih̄tt* in M345–M351 (cf. Wegera et al. 2018: 75)
- German Umlaut, e. g. *Dīv* vs. *dū* representing /y:/ in M345–M347

Graphemic variation reveals the inventory of characters that is available to a certain scribe at a certain time and place. We use the term graph(em)ic variation (Lemke 2020) to refer to such variants as listed above and also to refer to surface

Table 1: 14th century charters in ReM.

M345	14_1-bairalem-PUV-G	Augsburger Urkunden
M347	14_1-alem-PU-G	Freiburger Urkunden
M351	14_1-bair-PUV-G	Landshuter Urkunden
M348	14_1-omd-PU-G	Jena-Weidaer Urkunden
M350	14_1-mfrk-PUV-G	Kölner Urkunden
M352	14_1-rhfrhess-PU-G	Mainzer Urkunden
M353	14_1-ofrk-PU-G	Nürnberger Urkunden

and form properties as the following (see also Russ 1978; Simmler 2003; Elmentaler 1993; 2018, Glaser 1985; Wegera et al. 2018):

- diacritic marks, e. g. < ’ > in *dīe* vs. *d̄ie* in M345–M347 (cf. Gärtner 1991)
- ligatures, e. g. *ftæten* vs. *ftāten* and *ftaeten* in M344–M345
- marks encoding nasals, e. g. *hūdert* vs. *hūndirt* in M350–M352
- capitalisation, e. g. *Graben* vs. *graben* in M345–M347 (cf. Bergmann 1999)
- <er>-abbreviations, e. g. *vnz’brochen* vs. *vnzerbrochñ* in M344–M345 (cf. Dülfer and Korn 2006; Cappelli 2006)

As we will show in the following paragraphs, the difference profiles generated by the automatic methods can be translated into profiles of graphemic divergence between pairs of corpus texts, reflecting the degree of graphemic distance between them. Since graphemic variation can be linked to the factor space (diatopic variation) as well as the factor time (diachronic variation), a sound methodological approach is to reduce the parameters of variation to one factor (space or time).

In Section 3.1, we present selected findings that illustrate how our methods produce results which comply with previous findings in historical graphemics on the level of diatopic variation. To do so, we analyze text pairings of charters dated to the first half of the 14th century (to exclude the factor time from the equation; see Table 1). The following examples show in which ways well-known cases of diatopic variation such as New High German diphthongization or consonantal variation due to the High German consonant shift are reflected in the difference profiles.

In Section 3.2, we present an in-depth analysis of one single difference profile derived from two texts yielding diachronic variation.

3.1 Text pairings reflecting diatopic variation

The findings we elaborate on in this paragraph are derived from pairing the corpus texts shown in Table 1.

Table 2: Top 10 entries of the difference profile M345–M347.

Frequency	Rule	Phenomenon	Mapping		Weight
46	E → e/t_#	apocope, final	#G	#g	0.0753611
43	E → e/g_#	apocope, final	G	g	0.0756565
25	E → e/h_#	apocope, final	'#	er#	0.0863456
23	s → f/e_#	s-spelling, final	#Go	#go	0.111989
21	i → i/r_c	diacritic mark, medial	Go	go	0.111989
17	' → er/d_#	<er>-abbreviation, final	i	í	0.125067
15	E → e/b_#	apocope, final	'	er	0.131312
14	ñ → en/b_#	nasal, final	'#	r#	0.150607
13	E → t/o_t	double consonant, medial	s#	f#	0.151205
12	í → i/l_c	diacritic mark, medial	s	f	0.157837

We start with pairs of texts from the language area 'Oberdeutsch' (Upper German), which encompasses Alemannic (Freiburger Urkunden), Bavarian (Landshuter Urkunden) and a transition area between the two (Augsburger Urkunden). The relatively small graphemic distance between the texts in this group becomes evident when looking at the top 10 level rules and Levenshtein rankings (see Table 2).

The difference profile (Table 2) shows variation on the spectrum of differences in spelling that concern surface and form properties: The difference profile consists mainly of capitalization, use of the <er>-abbreviation, diacritic <'> and different s-spellings.

However, there are two distinct differences between M347 (Freiburg) and M345 (Augsburg) which can be identified as indicators of diatopic variation: Firstly, M345 is more prone to apocope than M347. That is consistent with past research on apocope: Lindgren (1953: 178) has shown that Bavarian texts tend to show apocope earlier to a larger extent than Alemannic texts. Secondly, there is a difference in diphthong spelling <ai ~ ay ~ ei>: while M345 (Augsburg) leans towards Bavarian <a> as first part of the digraph (preferring <ai> over <ay>, cf. Paul, Mhd.Gr.: § L45), M347 (Freiburg) opts for <ei>. The latter is documented in a series of rules of the type 'a → e...' in front of *i* or *í* and *y* respectively, derived from such pairings of word forms as *aigen* ~ *aygen* vs. *eigen*, *ain* vs. *eín*, *zwaín* / *zway* vs. *zweín*.

We continue with diatopic variation on the graphophonematic level: as an example, instances of early diphthongization /i:/ > /ae/ become apparent when looking at the difference profile M345 (Augsburg) – M351 (Landshut), see (5). Digraphic spellings <ei> reflecting diphthongization are more prevalent in Bavarian charters (cf. Paul, Mhd.Gr.: 75) at the time. This can be seen in our material when looking at the following selected rules:

(5) **M345–M351**

- 15 E → e/r_i
- 11 i → eí/m_n
- 11 E → e/m_i
- 9 E → e/f_í
- 9 E → e/m_í
- etc.

While M345 mainly preserves the monograph <i> or <í>, M351 opts for <ei> / <eí>, e. g. *dri*, *drízzíg*, *mín*, *fí*, *fin* (Augsburg) → *Drei*, *Dreizichk*, *mein*, *fei*, *fein* (Landshut).

We now take a look at Middle German texts as well, contrasting M350 (Köln) with M352 (Mainz). Representing different parts of the ‘Rhenish fan’, those two texts differ in properties linked with the High German consonant shift: While M350 holds on to Germanic */d/ <d>, not realizing the shift to High German /t/ (*dach* ~ *dag*, *deil*, *dún*), M352 shifts to <t> in initial position, thus pre-empting the spread of this Upper German feature into West Middle German writing practices (*tag*, *teíl*, *tún*) (cf. Paul, Mhd.Gr.: § L62).

The following rules, which are derived from the examples listed above, point to this variation:

(6) **M350–M352**

- 10 d → t/#_a
- 2 D → t/#_a
- 2 d → t/#_e
- 4 d → t/#_u

Some specific traits of Riparian writing become visible as well, such as unshifted Germanic /t/ in word final position (*dat* vs. *das* ~ *daz*, *et* vs. *es* ~ *ez*, *wat* vs. *was* ~ *waz*) and fricative <v ~ u> instead of MHG /b/ (cg. Paul, Mhd.Gr.: § L98), e. g. *erue* ~ *erve* (M350) vs. *erbe* (M352).

(7) **M350–M352**

- 3 v → b/a_e *haven* vs. *haben*
- 2 u → b/a_e *hauen* vs. *haben*
- 6 v → b/e_e *geven* vs. *geben*; *geſchreven* vs. *geſchreiben*
- 7 u → b/e_e *geuen* vs. *geben*; *geſchreuen* vs. *geſchreiben*
- 2 v → b/l_e *felve* vs. *felbe*
- 2 u → b/l_e *feluen* vs. *felben*
- 2 v → b/r_e *erve* vs. *Erbe*
- 7 u → b/r_e *erue* vs. *erbe*

Table 3: Top entries of the difference profile M344–M345.

Frequency	Rule	Phenomenon	Mapping		Weight
24	e → E/t_#	apocope, final	e#	#	0.0816302
17	e → E/g_#	apocope, final	í	i	0.118916
16	er → ‘/n_#	<er>-abbreviation, final	er#	‘#	0.128063
14	e → E/b_#	apocope, final	’	er	0.156097
11	f → s/n_#	s-spelling, final	er	’	0.158187
10	E → e/l_n	syncope in M344, medial	i	í	0.16799
10	E → ^e /o_l	superscript, medial	ín	in	0.179793
10	E → ^e /v_n	superscript, medial	te	t	0.187742
10	en → ñ/b_#	nasal, final	f#	s#	0.195347
9	í → i/a_n	diacritic mark, medial	’#	er#	0.203591
9	i → í/r_c	diacritic mark, medial	te#	t#	0.207653
9	’ → er/d_#	<er>-abbreviation, final	#ch	#k	0.234775
9	f → s/e_#	s-spelling, final	r#	’#	0.235377
8	ch → k/#_a	High German consonant shift, initial	’	r	0.235409
8	er → ‘/d_#	<er>-abbreviation, final	#ph	#pf	0.236286
8	h → f/p_e	ph/pf, medial	ph	pf	0.236286
8	í → i/m_n	diacritic mark, medial	ge	g	0.238273
8	z → s/r_#	s-spelling, final	v	u	0.238322
7	g → G/#_e	capitalization, initial	#t	#T	0.240639
			’	e	0.244531
			in	ín	0.248136
			ge#	g#	0.25562
			ch	k	0.255671
			r#	#	0.260164
			ner#	n’#	0.263805
			ner	n’	0.263805

3.2 Text pairings reflecting diachronic variation

The findings we focus on in this paragraph are derived from texts pairings of charters from Augsburg dated in the second half of the 13th century (M344) and the first half of the 14th century (M345). Both texts belong to an area of Upper German which is located between Alemannic and Bavarian. That is, we pair two texts from the same area, but from different time periods.

The rules derived from pairing M344 and M345 show variants that are related to different levels of linguistic variation: to graphophonemic variation as well as to graph(em)ic variation (see the rules and our categorization in Table 3).

The difference profile shows variation on the graphophonematic level as well as on the graphemic level. The variation -Ø vs. -e in word-final position concerns

different lemmas and becomes visible as a pattern through highly ranked rules and Levenshtein rankings:

(8) **M344–M345**

24 $e \rightarrow E/t_ \#$

17 $e \rightarrow E/g_ \#$

14 $e \rightarrow E/b_ \#$

M344 uses ‘full’ forms whereas in the later text, M345, the corresponding word forms strongly tend to apocoptation. These rules are triggered mainly by nouns, such as *baumgarte* vs. *Bavngart*, *geburte* vs. *gebürt*; *geziuge* vs. *geziug*, *clage* vs. *clag*, *phenninge* vs. *pfennig* but also show up in *vmbe* vs. *vmb*. The rules with frequencies 24, 17 and 14 and the top-ranked mapping replicate results from historical linguistics: Starting in the 13th century, apocope of final vowels is considered to have occurred first in Bavarian and subsequently also in Alemannic. As is to be expected, our data concurs with the known fact that it is not until the first half of the 14th century that Augsburger Urkunden begin to noticeably reflect apocope. So this can be identified as an indicator of diachronic variation.

In addition, there is a devoicing of consonants in word-final position. The rules derived from pairing M344 and M345 encode relations such as <k> vs. <g> which are associated with final-obstruent devoicing of plosives, e. g.

(9) **M344–M345**

7 $k \rightarrow g/i_ \#$

While M344 leans towards <k> in word-final position, M345 opts for <g>, e. g. *ledik* vs. *ledig*, and is particularly evident in numerals, e. g. *drizzik* vs. *drizzig*, *funfzik* vs. *fünfzig*, *zwainzik* vs. *zwaintzig*. These findings on g-spellings are in line with past research on the graphemic representation of final devoicing reflecting its decrease and the increase of , <d> and <g> in the 14th century (Wegera et al. 2018: 125; cf. Brockhaus 1995, Goblirsch 1994, Mihm 2004).

The graphophonemic variation <ch> vs. <k> in initial positions becomes visible as a pattern through the rules (e. g. *chain* vs. *kain*, *chomen* vs. *komen*) which can be interpreted as indicators of the High German consonant shift. M344 shows shifted ch-forms, whereas the latter text, M345, already prefers unshifted k-forms, which corresponds with past research on decline of Upper German <ch> in Middle High German (Paul, Mhd.Gr.: §§ L 59–62).

(10) **M344–M345**

Rule: 8 $ch \rightarrow k/\#_a$

Levenshtein: #ch #k 0.234775

ch k 0.255671

On the level of graph(em)ic variation, there is a difference in s-spellings: M344 prefers word-final <f> and <z> and M345 is prone to consistent <s>-spellings, e. g.

(11) **M344–M345**

- 11 f → s/n_#
- 9 f → s/e_#
- 8 z → s/r_#
- 7 z → s/a_#

M344 prefers variant s-spellings <f> and <z> in word-final position, whereas M345 already tends to transcribe <s>, e. g. in forms like *aigenf* vs. *aigens*, *minf* vs. *mins*, *vnf* vs. *vns*, *Def* vs. *Des*, *Gotef* vs. *Gotes*, *Swaz* vs. *fwas*, *Daz* vs. *Das*, and in proper nouns, e. g. *Jacobf* vs. *Iacobs* and *Johanf* vs. *Iohans*. This can be identified as an indicator of s-Zusammenfall, i. e. the reduction of s-variants in diachronic development (Wegera et al. 2018: 89, cf. Michel 1959). One interesting finding reflecting the diachronic development is that M345 is more prone to capitalisation than M344. The difference profile also includes the use of the <er>-abbreviation, diacritic marks <˘> and superscripts. M345 opts for superscripts, e. g. *folte* vs. *f^olte*, *getvn* vs. *get^ovn*, *vnf* vs. *v^ons*.

(12) **M344–M345**

- Rules: 10 E → ^e/o_l
 10 E → ^e/v_n
- Levenshtein: #v #^ev 0.265871
 o ^o 0.334745

When it comes to proper names (eg. *Konrad*), we observe a high degree of graph(em)ic variation: *Chvnrat/Chunrat/chvnrat* (M344) vs. *C^ovnrat/C^uvnrat/C^uvnrat/Ch^uvnrat* (M345). Proper names tend towards specific, even idiosyncratic variation (the tendency towards a high degree of variation in proper names and toponyms is observable till the Early New High German period, cf. Mihm 2000). The rules and mappings abstract from the concrete word forms and do not show whether they come from idiosyncratic variation. Hence, it is important to look behind the rules and rankings and at the word form pairs that yield the results. At a later stage, we will develop and introduce practices to handle this issue consistently. A clear requirement of all our efforts is to be – at all times – able to refer to the data underlying the difference profiles we work with to make sure our interpretations are not misguided.

We hope to have shown with these examples of diachronic and diatopic variation that the rules as a part of our computational linguistic methods provide a sound basis for a qualitative exhaustive graphemic analysis. In combination with

categorizations of the variables, we are able to draw a fine-grained overall picture of graphematic variation in Middle High German based on the automatically pre-analysed data.

4 Statistically determined graphemic similarities

The difference profiles, which were examined manually in the previous section, can also be compared to each other quantitatively at a macro level in order to obtain information about graphemic similarities of whole texts to each other. We illustrate how the statistically determined similarities of entire texts mirror known similarities between neighboring dialects. So here again, our statistical approach is able to produce results that are in line with results known from research.

For the quantitative analysis, we represent each text by the set consisting of all Levenshtein mappings from that text to all other texts. To focus on typical, frequently-seen mappings, only the first 500 mappings of each pairing are included in the union set. For example, the text M345 is represented by the first 500 mappings of M345–M347, merged with the first 500 mappings of M345–M348, and so on. In total, there are 7 texts in the diatopic subcorpus (see Table 1), i. e., 6 mappings per text. That is, the union set can contain a maximum of 3000 mappings. However, there are of course many identical mappings from different pairs of texts, so the union sets ultimately contain between 1820 and 2122 mappings (on average: 1985). Next, we compare the texts (i. e., their union sets U1 and U2) pairwise, and determine the intersection S. From this, we compute the similarity score as follows: $\text{Simil} = |S|/|U1|$, i. e., the proportion of shared mappings in all mappings of text 1. (13) illustrates with some made-up examples how the similarity score works. Columns U1 and U2 show the union sets of two texts. ‘abc’ represents three mappings ‘a’, ‘b’, and ‘c’, which together form the union set. Column |U1| specifies the number of elements (mappings) in U1. |S| specifies the size of the intersection, and Simil provides the similarity scores.

(13)

	U1	U2	U1	S	Simil
1.	abc	abc	3	3	1
2.	abc	def	3	0	0
3.	abc	abd	3	2	2/3
4.	abc	a	3	1	1/3
5.	abcde	a	5	1	1/5
6.	a	abc	1	1	1
7.	a	def	1	0	0

Example 1 is a case of complete overlap between U1 and U2, whereas example 2 shows a case of no overlap, yielding scores of 1 (= 100 % overlap) and 0, respectively. Example 3 compares two equal-sized sets with a partial overlap and a score of $2/3$ (which would also apply to U2, in the reverse direction). Examples 4 and 5 show the effect that occurs for sets of unequal size: If U2 consists of only one mapping, S can contain at most 1 element, which is reflected by a small Simil score. In the opposite case, if U1 consists of only one mapping and this mapping is in S, $\text{Simil} = 1$ holds (example 6); if not, $\text{Simil} = 0$ (example 7). Since our union sets are roughly of equal size and show partial overlap, our comparisons are most similar to example 3.

The aim of the quantitative comparison is to show that similarity of neighboring dialects is reflected in statistical similarity of the writing system. Therefore, we present the results in a table (in form of a heatmap, see Fig. 2) in such a way that texts from neighboring dialects are placed in proximity in the table. The arrows in Fig. 1 arrange the texts in such a way that texts linked by the arrows (e. g., M351 and M347) come from neighboring dialects (Bavarian and Alemannic).

Figure 2 shows the statistical results in form of a heat map in which text 1 is plotted on the x-axis (columns) and text 2 is plotted on the y-axis (rows). The order of the texts corresponds to the order shown in Fig. 1.

Figure 1: Texts from the diatopic subcorpus, arranged by pairwise similarity.

The cells contain the similarity scores that apply between the text in this column and the text in this row. The darker a cell, the higher the similarity between the two texts. The diagonal would contain the identity mapping, so all cells of the diagonal would be filled in black with 1.0 (100 % overlap). The upper and lower triangles contain the similarity scores of text comparisons in both directions. For

Figure 2: Heat map of similarity scores (created by the R package ggplot2, Wickham 2016).

instance, the top left cell labeled ‘0.29’ (1st row, 2nd column) shows the similarity of M345 (= U1) with M351 (= U2). In contrast, the top left cell labeled ‘0.31’ (2nd row, 1st column) shows the similarity of the reversed order, i. e., comparing M351 (= U1) with M345 (= U2). The differences between both scores are rather small and can be attributed to differences in size, as explained in (13).

The following observations stand out:

- M353 shows the highest similarities overall, including 3 pairings with > 30 % overlap (see the rather dark column and row labeled as ‘M353’). This result reflects very well the central position of M353 in Fig. 2. Other texts with similarities to many texts are M352 and M345, which also have rather central positions.
- In contrast, M350 is dissimilar to all texts (only values < 20 %). M348 is also rather dissimilar overall. Both texts, M350 and M348, are located in the margins of the map shown in Fig. 2.
- If we want to compare the directly adjacent texts, we have to look at the diagonals directly above and below the central diagonal. Here, the diagonal becomes lighter from the upper left to the lower right, i. e., the Upper German texts (beginning with M345) are more similar to each other than the Middle German texts.

Table 4: All texts with their average similarity scores.

M350	M348	M347	M351	M352	M345	M353
0.17	0.18	0.21	0.22	0.24	0.25	0.26

Finally, Table 4 shows all texts with their average similarity scores, sorted according to the scores. The average scores confirm the observations made above with regard to individual text pairings in that M350 is least similar to all others and M353 is most similar.

This type of statistical analysis could be applied, for example, when studying a new collection of texts whose graph(em)ic properties are not yet known. With the help of such heatmaps, these texts can be automatically presorted based on their graphemathical similarities and differences.

5 Conclusion

In this paper, we have showcased our analysis workflow which combines methods from both computational and historical linguistics as a potent method for historical graphemics, mainly for investigations into graph(em)ic variation. We could give but a glimpse at the vast range of phenomena to be considered on this level of linguistic inquiry with the exemplary presentation of some results in Section 3. This paper serves as a progress report in order to inform the scientific community about our ongoing efforts in this area. On the level of historical graphemics, our next goal is to map out the above-mentioned continuum of different ‘levels’ of variation in detail. As a starting point, we use pairings of texts with identical properties, starting with pairings of one text with itself to determine kind of the ‘background noise’ of grap(em)ic variation, then pairing texts that differ in time, followed by language area. This will result in an ever more tightly woven net of graph(em)ic similarities and differences that will cover, in the end, the whole range of graphemic divergence during the MHG period and show its spatial and temporal distribution. In close collaboration between computational linguistics and historical linguistics, we strive to combine qualitative methods with quantitative methods, as we have demonstrated in Section 4. We see great added value in the proposed combination of qualitative and quantitative methods: So far, most studies focused on single graphemic phenomena or on specific authors or areas or time periods. These individual studies must necessarily remain fragmentary and can only provide a slice of the overall picture. We think that our approach,

based on the difference profiles, makes it possible to evaluate historical data exhaustively and allows us to study for the first time the writing system as a whole and thus, to relate different phenomena, authors, areas and time periods to each other.

Acknowledgment: We would like to thank the reviewers and the editors of this special issue for helpful comments. This work was partially funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project-IDs 200609649 and 89085660.

Appendix

Table A.1: Reference Corpus of Middle High German: Complete list of Charters; for a detailed description cf.: <https://www.linguistics.rub.de/rem/corpus/details.html>.

sigle	abbreviation	time	text	# tokens
M344	13_2-bairalem-PV-G	13.2	Augsburger Urkunden	4,910
M346	13_2-alem-PU-G	13.2	Freiburger Urkunden	12,960
M349	13_2-mfrk-PUV-G	13.2	Gottfried Hagen: Kölner Urkunden	9,361
M345	14_1-bairalem-PUV-G	14.1	Augsburger Urkunden	11,482
M347	14_1-alem-PU-G	14.1	Freiburger Urkunden	10,621
M351	14_1-bair-PUV-G	14.1	Landshuter Urkunden	9,940
M348	14_1-omd-PU-G	14.1	Jena-Weidaer Urkunden	1,507
M350	14_1-mfrk-PUV-G	14.1	Kölner Urkunden	5,934
M352	14_1-rhfrhess-PU-G	14.1	Mainzer Urkunden	10,432
M353	14_1-ofrk-PU-G	14.1	Nürnberger Urkunden	10,842

References

Bergmann, Rolf. 1999. Zur Herausbildung der deutschen Substantivgroßschreibung. Ergebnisse des Bamberg-Rostocker Projekts. In Walter Hoffmann, Jürgen Macha, Klaus J. Mattheier, Hans-Joachim Solms & Klaus-Peter Wegera (eds.), *Das Frühneuhochdeutsche als sprachgeschichtliche Epoche*, 59–79. Frankfurt a. M.: Peter Lang.

Bollmann, Marcel. 2018. Normalization of historical texts with neural network models. *Bochumer Linguistische Arbeitsberichte* 22.

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

- Brockhaus, Wiebke. 1995. *Final devoicing in the phonology of German*. Tübingen: Niemeyer.
- Cappelli, Adriano. 2006. *Lexicon abbreviaturarum*. 6th edition. Milano: Hoepli.
- Dipper, Stefanie & Sandra Waldenberger. 2017. Investigating diatopic variation in a historical corpus. In *Proceedings of the EACL-Workshop on NLP for Similar Languages, Varieties and Dialects* (VarDial), 36–45. Valencia, Spain: Association for Computational Linguistics.
- Dülfer, Kurt & Hans-Enno Korn. 2006. *Gebräuchliche Abkürzungen des 16.-20. Jahrhunderts*. Reprint of the 9th, revised edition. Edited by Karsten Uhde. Marburg: Archivschule Marburg.
- Elmentaler, Michael. 1993. Probleme der Rekonstruktion stadtsprachlicher Schreibsysteme am Beispiel Duisburgs. *Zeitschrift für Dialektologie und Linguistik* 60. 1–20.
- Elmentaler, Michael. 2018. *Historische Graphematik des Deutschen. Eine Einführung*. Tübingen: Narr.
- Jurish, Bryan (2011). *Finite-state canonicalization techniques for Historical German*. Potsdam: University of Potsdam dissertation.
- Gärtner, Kurt. 1991. Die Williram-Überlieferung als Quellengrundlage für eine neue Grammatik des Mittelhochdeutschen. *Zeitschrift für deutsche Philologie* 110. 23–55.
- Glaser, Elvira. 1985. *Graphische Studien zum Schreibsprachwandel vom 13. bis 16. Jahrhundert. Vergleich verschiedener Handschriften des Augsburger Stadtbuches*. Heidelberg: Winter.
- Goblirsch, Kurt G. 1994. *Consonant strength in Upper German dialects*. Odense: Odense University Press.
- Klein, Thomas, Klaus-Peter Wegera, Stefanie Dipper & Claudia Wich-Reif. 2016. Referenzkorpus Mittelhochdeutsch (1050–1350), Version 1.0. [https://www.linguistics.ruhr-uni-bochum.de/rem/\[ISLRN 332-536-136-099-5\]](https://www.linguistics.ruhr-uni-bochum.de/rem/[ISLRN%20332-536-136-099-5]).
- Klein, Thomas & Stefanie Dipper. 2016. Handbuch zum Referenzkorpus Mittelhochdeutsch. *Bochumer Linguistische Arbeitsberichte* 19.
- Lemke, Ilka. 2020. *Das Komma. Zur syntaktisch-graphematischen Klassifikation des Zeichens im Sprach- und Schriftsystem des Deutschen und zur historischen Entwicklung aus formaler und funktionaler Perspektive*. Frankfurt a. M.: Peter Lang.
- Levenshtein, Vladimir I. 1966. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady* 10(8). 707–710.
- Lindgren, Kaj B. 1953. *Die Apokope des mhd. -e in seinen verschiedenen Funktionen*. Doctoral dissertation. Helsinki: Suomalainen Tiedekatemi.
- Michel, Wolf-Dieter. 1959. Die graphische Entwicklung der s-Laute im Deutschen. *Beiträge zur deutschen Sprache und Literatur (H)* 81. 456–480.
- Mihm, Arend. 2000. Zur Deutung der graphematischen Variation in historischen Texten. In Annelies Häcki Buhofer (ed.), *Vom Umgang mit sprachlicher Variation. Soziolinguistik, Dialektologie, Methoden und Wissenschaftsgeschichte*, 367–390. Tübingen: Francke.
- Mihm, Arend. 2004. Zur Geschichte der Auslautverhärtung und ihrer Erforschung. *Sprachwissenschaft* 29. 133–206.
- Mittmann, Roland. 2020. *Zur althochdeutschen Zeit- und Dialektgliederung. Eine computergestützte Untersuchung auf Grundlage der textlichen Überlieferung*. Hamburg: Baar.
- Paul, Mhd.Gr. = *Mittelhochdeutsche Grammatik von Hermann Paul*. 2007. Newly edited by Thomas Klein, Hans-Joachim Solms, and Klaus-Peter Wegera. Syntax by Ingeborg Schröbler, new edited by Heinz-Peter Prell. 25th edition. Tübingen: Niemeyer.
- Pettersson, Eva (2016). *Spelling normalisation and linguistic analysis of historical text for information extraction*. Uppsala: Uppsala University dissertation.

- Russ, Charles V. J. 1978. *Historical German phonology and morphology*. Oxford: Clarendon Press.
- Simmler, Franz. 2003. Phonetik und Phonologie, Graphetik und Graphemik des Mittelhochdeutschen. In Werner Besch, Anne Betten, Oskar Reichmann & Stefan Sonderegger (eds.), *Sprachgeschichte. Ein Handbuch zur Geschichte der deutschen Sprache und ihrer Erforschung*. Vol. 2, 1320–1331. 2nd comp. newly revised and corr. edition. Berlin & New York: de Gruyter.
- Wegera, Klaus-Peter, Sandra Waldenberger & Ilka Lemke. 2018. *Deutsch diachron. Eine Einführung in den Sprachwandel des Deutschen*. 2nd revised edition. Berlin: ESV.
- Wickham, Hadley. 2016. *ggplot2: Elegant graphics for data analysis*. New York: Springer. <https://ggplot2.tidyverse.org>.