Hanna Lüscho*^{*}

Quantifying graphemic variation via large text corpora

<https://doi.org/10.1515/zfs-2021-2038>

Abstract: The use of some basic computer science concepts could expand the possibilities of (manual) graphematic text corpus analysis. With these it can be shown that graphematic variation decreases constantly in printed German texts from 1600 to 1900. While the variability is continuously lesser on a text-internal level, it decreases faster for the whole available writing system of individual decades. But which changes took place exactly? Which types of variation went away more quickly, which ones persisted? How do we deal with large amounts of data which cannot be processed manually anymore? Which aspects are of special importance or go missing while working with a large textual base?

The use of a measurement called *entropy* quantifies the variability of the spellings of a given word form, lemma, text or subcorpus, with few restrictions but also less details in the results. The difference between two spellings can be measured via *Damerau-Levenshtein distance*. To a certain degree, automated data handling can also determine the exact changes that took place. Afterwards, these differences can be counted and ranked.

As data source the *German Text Archive* of the Berlin-Brandenburg Academy of Sciences and Humanities is used. It offers for example orthographic normalization – which is extremely useful –, preprocessing of parts of speech and lemmatization. As opposed to many other approaches the establishment of today's normed spellings is not seen as the aim of the developments and is therefore not the focus of the research. Instead, the differences between individual spellings are of interest. Afterwards intra- and extralinguistic factors which caused these developments should be determined.

These methodological findings could subsequently be used for improving research methods in other graphematic fields of interest, e. g. computer-mediated communication.

Keywords: historical graphematics, spelling variation, German Text Archive, methodology, entropy

***Corresponding author: Hanna Lüscho,** Institut für Germanistik, Carl von Ossietzky Universität Oldenburg, Ammerländer Heerstr. 114–118, 26111 Oldenburg, Germany, e-mail: hanna.lueschow@uol.de

1 Introduction

Research in historical graphematics has come a long way over the last decades, even more so given the developments in data digitization. Important research has already been conducted for specific phenomena like capitalization (e. g. Bergmann and Nerius 1997; Dücker et al. 2020) or the establishment of morphological spellings and stem constancy (e. g. Ruge 2004; Voeste 2008a). There are also more comprehensive analyses which, however, often only investigate a very narrow time span and/or are restricted to very specific authors, areas, or genres (e. g. Moser 1977; Glaser 1985; Koller 1989 – just to name a few).

The research at hand instead sets out to collect as much data as possible to allow for reliable conclusions concerning each possible variation phenomenon, independently of whether or not it belongs to a specific research area. It is a quantitative rather than a qualitative approach and does not consider phonological aspects, which distinguishes it from methods such as those described in Elmentaler (2018).

The definition of graphemic variation used here is similar to the definition of spelling variation in Barteld (2017: 13): The realization of a word form using different spellings. In the corpus-linguistic terminology adopted here, a word form differs from a morphological word in that it does not include part-of-speech information, other morphological information such as case or number, or word sense disambiguation. The normalized surface form and the lemma are the main factors in determining different word forms; see the generated data structure in Section 2.

It is well known that spelling variation decreases constantly until the beginning of the twentieth century, even before the orthographic conferences – whose influence on the German orthography is still debated – took place in 1876 and 1901 (Nerius 2002). The research at hand aims to build on these findings and provides a methodological proposal. While the specific research questions and long-term goals concern the determination of the exact changes that took place and which types of variation went away more quickly while others persisted, the present paper can be understood as a pilot study showing a few preliminary and exemplary results.

For this, the German Text Archive (Deutsches Textarchiv, DTA) of the Berlin-Brandenburg Academy of Sciences (BBAW 2019) is used. It contains more than 150 million tokens of written historical German data from the 15th to the 20th century. Most of it has the original spellings preserved while also being annotated on various levels, which is why it is very well-suited for this topic and for the approach suggested in the present paper.

This article presents a methodology which

1. quantifies how evenly distributed variants are,

2. quantifies their frequency, but
3. is independent of sample size, and
4. makes it possible to easily automatically determine the concrete differences between two spellings of one word form.

The remainder of this article is organized as follows: Section 2 gives more information about the DTA in general and the derived data structure in particular. Section 3 explains the proposed methodology; some resulting data is shown in Section 4. Section 5 offers a conclusion and an outlook.

2 Data

The version of the DTA corpus used in the present paper is the Text Corpus Format (TCF) version published on 18th October 2018, with the core corpus as well as the supplementary texts (3527 texts in total).

The data needs some clean-up before being used for the analysis, both on the text level and on the token level. This helps to ensure that the database is as homogeneous as possible.

On the text level, handwritten texts and those written in Roman type are excluded, because printed texts make up the vast majority, and Fraktur was the most common font at the time. The data from before 1600 and after 1899 is also excluded because it is too sparse, which is a result of the fact that data collection for the DTA corpus mainly focused on the time between 1600 and 1900. Tokens tagged as punctuation or as foreign material are also excluded. Lastly, some sub-corpora which have proven to be less true to the original than others, as well as some words containing very rare characters, have been excluded as well. In the final step, capitalization has been removed; in the concept of variation used here, it is not considered relevant, primarily because it would entail that all sentence-initial words would be recognized as variants as well.¹ After cleaning the data in this way, around 100 million tokens are left.

Note that the data for each decade is not evenly distributed on the text level or on the token level. Figure 1 shows the tokens for each decade, as it is the most useful unit for size comparison.

¹ In the process of removing the capitalization, every capital <S> was converted to <f> because of the two variants of the <s>-spelling and their distribution – both <S> and <f> tend to be used word-initially. The occurrences of <s> are not changed.

Figure 1: Token counts of the (filtered) DTA from 1600–1899.

Figure 2: The generated data structure.

Based on the annotation available in the DTA, a special data structure has been generated, as shown in Figure 2. There are two different perspectives: (1) A **text** exists of only the tokens in this specific text and is enriched with meta-data such as the genre, the year of publication or the publication location(s). This will be called the *local perspective*. (2) A **collection** is a set of texts which share a specific feature, e. g. were published in the same decade. This will be called the *global perspective*. These two perspectives allow for a comparison of the writing conventions of specific authors – or other people involved, such as printers or typesetters – and the writing conventions of whole decades, regions, genres etc.

In the following, the preparatory work for the analysis both from a local and from a global perspective is described.

The different annotation layers of the DTA were merged to generate the data structure used here and pictured in Figure 2. Each text or collection consists of a number of lemmas. The lemmas then consist of word forms, as defined in Section 1. Wherever a normalized word form of the individual spellings is available in the annotation, the normalized word form is used.² The details behind the normalization and its evaluation are described in multiple papers, e. g. Jurish (2011).³

The word forms then consist of different spellings. Each spelling has its own part-of-speech tag, a frequency count, and so-called position counters. These are generated using successive numeration and reset after each text. They are stored for possible later use in analyses of the distance between two spellings in a text (more in Section 5).

As an example,⁴ consider the lemma *geben* ‘give’ consisting of the word forms *gegeben* (past participle) and *gibt* (3rd person singular). While *gegeben* has only one spelling, <gegeben>, *gibt* is attested in multiple spellings: <giebet>, <gibt> and <giebt>.⁵

One of the peculiarities resulting from the digitization of texts with letters that are unusual today is the fact that so-called combining characters occur. These are special characters that are combined with the preceding character to modify it. The most frequent combining marks are diacritics such as tremas, tildes or macrons. For example, <â> is a combination of a regular <a> and the “combining small letter e” (cf. Ebert et al. 1993: 35).

Combining characters are counted as separate characters, which will be relevant in Section 3.3. For some of the characters like the “combining small letter e” this is more obvious because there is no other character with this diacritic. However, whether a letter like <ö> is encoded as one or two characters depends on whether the OCR software categorizes it as a character in its own right (unicode ð – U+00F5) or as a combination of <o> with a combining tilde (ö, U+0303), and also on the transcription decisions. Therefore Unicode normalization has been applied to each token.

² A similar approach is used by Dipper and Waldenberger (2017: 38): “All word forms with identical normalizations are considered equivalent.” The difference here is that only those tokens that differ from their normalized form have an explicitly specified normalization.

³ A manual evaluation of 5000 randomly selected sentences – with only one sentence per also randomly selected file – shows a correct normalization for about 97 % of the tokens. A smaller sample for the lemmatization shows a very similar quality. These numbers are very good for historical data.

⁴ Extracted from the following text: http://deutschestextarchiv.de/book/show/hc_1532409_1751

⁵ Each spelling occurs once and has the POS tag ‘VVFİN’. The position counters are 1748, 4086 and 4743.

3 Method

Before the method proposed in this paper can be discussed, it is necessary to introduce a concept crucial to it: a *difference*. From now on, the individual operations used to get from one spelling to another will be referred to as differences. These differences are the main unit of analysis.

Those differences are determined via pairwise comparisons of spellings. All the spelling pairs affected by a certain difference are considered collectively. This means that certain variations (e. g. of two graphemes) are observed over the whole data and allows for a generalization across specific word forms or spellings.

The term *difference* is preferred to *rules* as used by Dipper and Waldenberger (2017) because it has a less normative sound to it. While they pursue a very similar goal, their analysis also includes information about the non-varying parts of the spellings, called “identity rules”.⁶ Beyond that, their exact method of quantification is not fully transparent, and therefore cannot be reconstructed.

As already stated in Section 1, we need a measurement that takes the distribution of the individual spellings as well as their frequency into account. Therefore, the methodological approach combines three different aspects: First, we want to quantify how much variation there is, e. g. in a decade.⁷ For this, a measurement called *entropy*, described in Section 3.1, is useful.

The shapes the differences can take will be shown in Section 3.2, which discusses a second aspect vital to this methodology, the Damerau-Levenshtein distance. This also helps in achieving an automatic analysis of the concrete differences between two spellings of a word form.

The third important aspect are the frequencies of the individual spellings affected by a certain difference. As the data for each decade differ in size (as shown in Figure 1), a unifying measurement is needed. Therefore, a measure of *scaled frequency*, discussed in Section 3.3, is used.

It should also be noted that these three sections only serve as introduction to the concepts which are important for this topic and differ in some details from the eventual procedures/variants used in the proposed method. Therefore, the data shown in the next few sections is not necessarily based on the same metrics as those shown in the preliminary results in Section 4. These discrepancies will each be addressed, where deemed necessary, in the respective sections.

⁶ Another aspect that distinguishes these rules from the results obtained here is the separate treatment of each context around the application of a rule and the missing abstraction that results from that.

⁷ Of course the use of other time spans would have been possible. Decades are used here mainly because of their use being relatively conventionalized; future research might even use sliding-window approaches to avoid the arbitrariness that lies in most time spans.

How these three aspects interact is discussed in Section 3.4. This provides us with the methodology that underpins the rest of the paper.

3.1 Entropy

Entropy is a versatile measurement used in many different contexts. Shannon (1948) was the first to use it for linguistic data. The formula for this is

$$H = - \sum p_i \log p_i$$

where p_i denotes the probability of encountering a certain linguistic unit.⁸

Only determining the proportion of words that have more than one spelling variant – as it is done by Baron et al. (2009) – does not account for cases in which words have multiple spellings, but one of these spellings significantly outweighs the other spellings. Entropy is much more suitable for quantifying the distribution of variants because such aspects are taken into account. If multiple spellings have very frequent occurrences, this is reflected in the entropy value.

Another advantage of this measure is its use in other areas of historical linguistics such as morphology and syntax (Moscoso del Prado Martín 2014) and the comparisons and cross references this enables.⁹

The linguistic unit used by Shannon is the character; other units, e. g. morphemes, are also possible. In the approach described here, a different, more unusual unit is used: the occurrences of a certain spelling variant in relation to all spelling variants of the word form.

Consider the following example: The word form *und* ‘and’ is used seven times in a text, five times as <vnd>, once as <vnnd>, and once as <vñ>.

The probabilities of the variants are therefore $p_{<vnd>} = \frac{5}{7} \approx 0.714$ and $p_{<vnnd>} = p_{<vñ>} = \frac{1}{7} \approx 0.143$. The entropy value H_{und} is then calculated in the following way:

$$\begin{aligned} H_{und} &= -(p_{<und>} \times \log_2 p_{<und>} + p_{<vnnd>} \times \log_2 p_{<vnnd>} + p_{<vñ>} \times \log_2 p_{<vñ>}) \\ &= -(0.714 \times \log_2 0.714 + 0.143 \times \log_2 0.143 + 0.143 \times \log_2 0.143) \approx 1.15^{10} \end{aligned}$$

⁸ Modifications have been made where the probability of a character is calculated depending on the context (Shannon 1951); e. g. if a <c> is the last letter encountered, then the probability of the next character being a <h> is much higher than the probability of it being a <x>.

⁹ Although not directly relevant for the research at hand, it is also worth noting that the use of entropy-based measures has proven to be informative in cognitive linguistics (Baayen and Moscoso del Prado Martín 2005; Moscoso del Prado Martín 2016).

¹⁰ Here a logarithm to the base two is used; one might consider other bases, for example the size of the present alphabet. More important than the base itself is the consistency of the base used.

Word forms that occur only once (hapax legomena) are excluded because of the fact that they cannot show any variation and therefore give no information for the present research question.

With these values, we can quantify the distribution of different spellings of a word form – the higher the value, the more evenly distributed are the variants. Thus, the degree of variation of a word form with a high entropy value can be considered to be higher compared to word forms with lower entropy values. From a diachronic perspective, higher degrees of variation can be understood as an indicator that the spellings in question have not yet reached a final, unified state.

However, this does not have to stop at the word form level. To quantify the entropy of a higher-level unit such as the lemma, the text, or the decade, the weighted mean is calculated. This means that the more frequent a word form is, the more impact it has on the mean, because it is seen more often and therefore accounts for a higher proportion of the variation.

An example: Assume we have the lemma *sagen* ‘say’ with the word forms *sagte* (past tense) and *gesagt* (past participle). *Sagte* occurs six times in total, three times spelled as <fagete> and three times as <fagte>, therefore with an entropy of 1. *Gesagt* is only attested in the spelling variant <gefagt>, which occurs twice, with an entropy of 0. The weighted entropy would thus be:

$$\frac{6 * 1 + 2 * 0}{8} = 0.75$$

The same logic applies at the level of the individual text (local perspective, see Section 2) and at the level of text collections (global perspective). Thus, the weighted means of the token-level entropies both at the text level and at the level of text collections are calculated, giving rise to the measures of local entropy and global entropy that are key to the method proposed here. Figure 3 shows the progression of entropy in the DTA between 1600 and 1899, depicted by the blue lines. The left panel shows the local entropy, the right one the global entropy.

The entropy for each decade in Figure 3 is determined using texts from the respective decade (e. g. from 1600–1609); therefore, the global entropy in principle shares the same textual base with the local entropy. The resulting values, however, are very different, which might not be evident intuitively. This can be explained as follows: When considering all texts separately, there are fewer different spelling variants of a word form than when putting all of these texts in a common “data pool” first. Also, some word forms might not be hapax legomena anymore and subsequently included in the calculation for the global perspective. All of this leads to possibly higher entropy values for each word form and this in turn affects the value for the whole decade. In other words: If two texts have a

Figure 3: Left: local entropy (blue) and token counts (red); right: global entropy (blue) and token counts (red).

very consistent spelling, their individual entropy will be close to 0. But if the respective spelling systems are very different (and they have a similar vocabulary), a collection of these two texts will have a much higher entropy.

In both graphs the entropy continually decreases, with the local entropy – the entropy of the texts – starting at a lower value. This shows that while the writers, typesetters and other people included in the text manufacturing process at each point in time seem to be closer to a state of less graphemic variation, the writing conventions of the decades as a whole are converging at a higher rate, but have a higher degree of graphemic variation at all times.

There are some spikes and dips in the global figure which would need to be further investigated, but the overall trend is not changed drastically.

The token counts are shown as well because a higher token count can have an influence as it leads to more variation possibilities and usually also to higher entropy values. This makes the progression even more conclusive because of the token count being lower in the first century as compared to the other two.

On the other hand, a lower token count in combination with less representative or otherwise diverging data can lead to data which is more misleading or should at least be interpreted with more caution. This might also account for the spikes and dips on the right-hand side in Figure 3 because they occur in some (but not all) of the decades with token counts differing from the surrounding decades.

The much smaller size of the texts could lead to two opposing results: Either the smaller text size leads to a lower level of entropy because of less possibility of variation, or it leads to a higher level of entropy because of marginal variants having a bigger influence than in the much bigger collections. The results of Figure 3 could be interpreted as an indicator for both of these things; this is a topic that should be further investigated in the future.

While entropy itself is not restricted to a certain range of values, due to the use of pairwise comparisons (further explained in Section 3.2) in this approach, when exactly two *p*-values are inserted into the entropy equation above, the result will always lie between 0 and 1. The weighted mean then also only allows for values between 0 and 1. This has not yet been the case for Figure 3 because working with spelling pairs was not yet necessary and all spellings of a word form were instead considered simultaneously which allowed for values higher than 1 – as seen in the example above. But in Section 3.4 this value limit will be applicable.

3.2 Damerau-Levenshtein distance

The Damerau-Levenshtein (DL) distance (Damerau 1964) and its basis, the Levenshtein distance (Levenshtein 1966), were originally used in error correction and word detection. In this approach, different spellings of the same word form are compared with each other in pairwise comparisons. This is necessary because a DL distance can only be determined for two spellings at a time.

Both Levenshtein distance and DL distance quantify the smallest number of operations needed to get from one text string (here: spelling) to another. Compared to the original measure of Levenshtein distance, DL distance allows for one additional operation, transposition. As such, four possible operations are taken into account:

- insertion (<stehn> → <stehen>),
- deletion (<todt> → <tot>),
- replacement (<bei> → <bey>),
- transposition (<nham> → <nahm>).

An example for multiple operations to get from one spelling to another would be the pair <fürgenommen> – <vorgenommen>. The following steps are necessary:

1. replacing <f> with <v>,
2. replacing <u> with <o>,
3. deleting the combining character <°>,
4. replacing the second <m> with a combining tilde.

The distance in this example therefore is 4. The relevance of this value for the research question pursued here lies in the conclusions it allows about the differences between two spellings. But in some cases, even in the example above, there is a problem of ambiguity; it would also be possible to assume other operations, e. g. inserting the <v>, replacing the <u> with the <v>, and replacing the <o> with the combining character, then again replacing the second <m> with a combining

Figure 4: Weight of the different Levenshtein distances form 1600–1699 (global perspective).

tilde. That this entails the same value for the DL distance is not of importance. It might not always lead to a different DL distance value but differing operations are very relevant for the concept of differences applied here.

These problems can be avoided by only looking at spelling pairs with a DL distance of 1. A better coverage of differences is achieved by including merge and split operations.¹¹ Another advantage of the merge and split operations is the additional possibility of handling the combining characters mentioned in Section 2 in differences such as <â> vs. <e>. Aspects such as the switch between <ph> and <f> or <ff> and <ß> can also be evaluated.

This is not meant to exclude data but to achieve a better coverage with automatic analyses. It is also not much of a problem because the majority of cases have a Levenshtein distance of 1, as shown in Figure 4. In this plot, we see a weighted calculation¹² of the proportion of different distances when looking at each pairwise comparison, the theoretical maximum possible value being 1.0. This makes it very clear that – when doing pairwise comparisons with all spellings of a whole

¹¹ Bollmann et al. (2011: 37) call these operations sequence-based rules. Merging and splitting operations are only applied if two spellings differ in the replacement of two adjoining letters by an individual third letter (as in <laffen> vs. <laßen> or <wäre> vs. <were>). Bigger splits (e. g. three adjoining letters) could theoretically be implemented, but this is not planned at the moment as they are not expected to improve the results.

¹² The weighted calculation for each distance consists of the scaled frequency (see Section 3.3) for the sum of each spelling pair of this distance, multiplied with the weighted entropy of these spelling pairs.

decade – the distance value of 1 is by far the most frequent and the most relevant.¹³ Regarding the pilot study character of this paper it is therefore seen as sufficient to pose a limit at this point.

It can also be assumed that higher distances are either less relevant on a graphemic level¹⁴ (i. e. morphological differences such as in <festiglich> vs. <feft>; see also Barteld et al. 2019: 701–702) or that they are combinations of lower-distance differences. If the latter is the case, there is no reason to assume a different distribution.

It also has to be kept in mind that many ambiguities might be implicitly resolved when manually determining differences because of an unintentional/unconscious bias. This would make the results harder to reconstruct. The above-mentioned spelling pair <fürgenommen> – <vorgenommen> serves as a good example for this.

In conclusion, a closer look at the DL distances has shown that the majority of spellings only differ in small aspects and the aforementioned limitations to the differences considered in the rest of the paper should not impair the results too much.

3.3 Scaled frequency

The concept of scaling frequencies can not only be applied to tokens but also to differences. It can thus be used to quantify the prevalence of individual spelling variants. There exist different ways of calculating a scaled value, the one used here is as follows:¹⁵

$$freq_{scale} = \frac{freq - min}{max - min}$$

¹³ This graph only shows the 17th century; the other two centuries are omitted due to lack of space. The 18th century shows a very similar distribution and there is only one decade in the 19th century which shows a higher relevance for distance 2 than for distance 1; that is the decade where a big shift for umlaut spellings (e. g. <û> to <ü>) takes place.

¹⁴ Barteld et al. (2019: 689): “Regarding the generation method, the best F-scores are obtained using a Levenshtein distance of 1.” This is a different perspective, but it also shows that such restrictions do not have to be a hindrance.

¹⁵ The formula would not be applicable if *min* were equal to *max*, but this cannot realistically be the case for the research at hand because that would mean that each of the many differences has exactly the same frequency.

Consider the following examples with the four values 2, 7, 50 and 78:

$$\begin{aligned} \text{freq}_{\text{scale}}(2) &= \frac{2-2}{78-2} = 0 \\ \text{freq}_{\text{scale}}(7) &= \frac{7-2}{78-2} \approx 0.066 \\ \text{freq}_{\text{scale}}(50) &= \frac{50-2}{78-2} \approx 0.632 \\ \text{freq}_{\text{scale}}(78) &= \frac{78-2}{78-2} = 1 \end{aligned}$$

This example is by no means realistic in terms of size, but it should suffice to demonstrate the general principle. The least frequent item is assigned a value of 0, the most frequent item is assigned the value 1. Everything in-between is scaled with regard to the distance between the maximum and the minimum value.

If there were only a few frequencies upon which the normalization is based, this value would be less informative. But the use here is based on enough values to mitigate these concerns.

The scaled frequency is more independent of the size of a (sub)corpus than the original frequency in that it still does not guarantee the representativity of the data, but it reduces the effect of sample size on our entropy measures.

A further advantage of this value in comparison to the relative frequency is that we always get a value that ranges from 0 to 1 which allows for better comparison between different absolute frequencies.

A difference that appears only with a few but very frequent spellings gets a higher value than a difference that appears with many but infrequent spellings. This alone could be problematic, but it is also precisely the reason why the proposed metric includes entropy as well. This combination will now be discussed in the following section.

3.4 The proposed metric

Keeping the concept of a difference as described in Section 3 in mind, the frequency of each pair of spelling variants relevant for this difference is determined as follows:

For each decade, while going through the (several hundred) individual differences, the frequency of both spellings of each spelling pair is added to the full frequency for the difference in question:¹⁶

$$\text{freq}(S_{\text{diff}}) = \sum_{k=1}^{|S_{\text{diff}}|} (\text{freq}(i) + \text{freq}(j))_k$$

¹⁶ The excluded spelling pairs are also excluded from this calculation.

i. e., for each spelling pair k in S_{diff} (the set of spelling pairs of type *diff*) the frequencies of the spelling i and j are added.

Again, a small example: Let us assume the difference <i> vs. <y> only occurs in two spelling pairs, <sein> (2 occurrences) vs. <seyn> (3 occurrences) and <bei> (1 occurrence) vs. <bey> (5 occurrences). The summation of each of these pairs would lead to $freq(<i> \text{ vs. } <y>)$ being 11.

Afterwards, the difference with the lowest frequency and the difference with the highest frequency are determined. Based on these two values and the formula discussed in Section 3.3, the scaled frequency for each difference is calculated.¹⁷

For the entropy of a difference we will again have a look at the example above: The entropy of the first spelling pair (<sein> vs. <seyn>) is about 0.97, and for the second spelling pair (<bei> vs. <bey>) it is about 0.65. We multiply the first value by 5 and the second value by 6 and afterwards divide the result by 11 – the weighted entropy therefore is almost 0.8.

For every difference taken into account, the weighted entropy is multiplied with the scaled frequency:

$$weight_{diff} = freq_{scale}(S_{diff}) \times H_{diff}$$

This leads to a value which still lies between 0 and 1 – the higher the value, the more relevant the difference. The combination of these values therefore quantifies the weight of specific differences. The values then serve as an orientation to make a comparison or ranking of the differences possible.

4 Example results

Because of the vast amount of data that is generated by the analysis of around 100 million tokens, only an exemplary analysis of the decade from 1700 to 1709 will be shown to demonstrate the possibilities of this methodology.

If we use the methodology as described in Section 3, limited to the local perspective, we get the results depicted in Table 1.

¹⁷ To add to the example, further differences would be needed to get a more useful value for the scaled frequency, which we will refrain from here due to lack of space.

¹⁸ The values in this and the following tables are rounded to three decimal places for readability.

¹⁹ These two spellings could refer either to the possessive pronoun or to the infinitive verb form. The word forms are separated because the lemmatization helps to avoid a mix-up. In a subsequent step these different cases are (separately) included in the calculation for the whole respec-

Table 1: The top 10 local differences from 1700–1709.

difference	example	weight _{diff} ¹⁸
e vs. ∅	<andere> vs. <andre>	0.165
h vs. ∅	<wohl> vs. <wol>	0.052
i vs. j	<ich> vs. <jch>	0.051
f vs. ∅	<auff> vs. <auf>	0.041
r vs. ʀ	<herr> vs. <her>	0.035
s vs. ß	<des> vs. <deß>	0.031
i vs. y	<fein> vs. <feyn> ¹⁹	0.023
t vs. ∅	<jetzt> vs. <jetz>	0.021
r vs. ∅	<darzu> vs. <dazu>	0.018
° vs. ∅	<kómmt> vs. <kommt>	0.018

Although, theoretically, values of 1 would be possible, the weighted difference rarely exceeds 0.2 (and most of these exceptions are found in earlier decades). This shows that the combination of the different measurements discussed in Section 3.4 helps to avoid a high ranking of differences that are only frequent or only have a high level of entropy, and therefore achieves a better integration of the different factors which influence variation than would be possible by just focusing on one measurement at a time. On the downside, this makes it harder to accurately distinguish the exact influences of the entropy and the scaled frequency.

If we widen our scope to the global perspective, the most relevant 10 differences are the ones shown in Table 2.

As the main goal of this article is to introduce the methodology, the individual results will not be discussed in too much detail. It should only be mentioned that the results in Table 1 and Table 2 are very similar. Thus, the analysis should focus on the differences such as the insertion of <t>.

These results largely correspond to the phenomena discussed in the literature (cf. Nerius 2003: 2467–2468; Ebert et al. 1993: 22–25, 32–35, 79–83).²⁰ This shows that the methodology discussed in Section 3 yields credible results. But the data, while being quantitatively meaningful, needs some further refinement to allow for qualitative conclusions as well.

tive differences and lead to a final value for the difference like the one shown in Table 1. Other homographic words that also share homographic lemmas unfortunately cannot be divided as easily.

20 Although they investigate a different, earlier timespan, Dipper and Waldenberger (2017) also have similar top (non-identity) rules, e. g. the mutual replacement of <i> and <j>.

Table 2: The top 10 global differences from 1700–1709.

difference	example	weight_{diff}
e vs. ∅	<andere> vs. <andre>	0.230
i vs. j	<ich> vs. <jch>	0.120
r vs. ʀ	<herr> vs. <herʀ>	0.074
h vs. ∅	<wohl> vs. <wol>	0.073
f vs. ∅	<auff> vs. <auf>	0.059
i vs. y	<fein> vs. <feyn>	0.050
s vs. ß	<des> vs. <deß>	0.044
r vs. ∅	<darzu> vs. <dazu>	0.032
ʳ vs. ∅	<kômmt> vs. <kommt>	0.027
en vs. ě	<einen> vs. <eině>	0.024

For some of the differences it would be useful to separate between different uses, e. g. whether an <e> is inserted after a vowel (<giebt> vs. <gibt>) or after a consonant (<andere> vs. <andre>) – the latter would result in a new syllable in most cases, especially if it occurs word-finally or if another consonant follows. This could be an interesting distinction as the factors leading to the insertion may differ from other, non-syllable-generating contexts. Another example for different usages is the doubling of a letter (<kann> vs. <kan>) as opposed to the insertion of a different letter (<nun> vs. <nu>).

To arrive at such a more nuanced picture, the context surrounding a difference would be a useful piece of information. However, even today there are some letters which cannot be clearly attributed to the group of vowels or consonants. These uncertainties are amplified when looking at non-standardized historical data. Decisions would have to be documented and well justified.

An easy and transparent way of taking a little more context into account is to distinguish between the initial, medial and final occurrence of the differences. If we remain at the global perspective, it gives us the results shown in Table 3.

It is not surprising that the differences mainly remain the same. But it is interesting to note that while some of the differences show up in different contexts, others are mostly restricted to a certain context. One could argue that these are different types of variation that should therefore be treated separately.

On the other hand, some of the differences might be more meaningful when combined as they belong to the same phenomenon. An example of this would be the alternation of <ö> vs. <õ> and <ü> vs. <û>. The data also has two different variants of apostrophes – probably because of inconsistent OCR – which are not expected to behave differently; keeping these separated could even distort the results.

Table 3: The top 10 global differences from 1700–1709, divided into initial, medial and final occurrences.

difference	example	weight _{diff}
e vs. Ø (medial)	<andere> vs. <andre>	0.212
e vs. Ø (final)	<habe> vs. <hab>	0.117
h vs. Ø (medial)	<wohl> vs. <wol>	0.096
i vs. j (initial)	<ich> vs. <jch>	0.094
s vs. ß (final)	<des> vs. <deß>	0.048
i vs. y (medial)	<fein> vs. <feyn>	0.046
r vs. ʀ (medial)	<herrn> vs. <herʀn>	0.045
f vs. Ø (final)	<auff> vs. <auf>	0.044
f vs. Ø (medial)	<offt> vs. <oft>	0.037
° vs. Ø (medial)	<kõmmt> vs. <kommt>	0.035

Figure 5: Development of the weight of the difference “e vs. Ø” from 1600–1899.

Looking at each decade separately does not allow for direct conclusions regarding the progression of the weight of a difference. Therefore, further analyses of the diachronic development will have to be carried out. One possible way of doing so is visualizing the data as depicted in Figure 5.

We can see in Figure 5 that the difference “e vs. Ø” loses some of its weight. With values like these we can explore the development of specific differences, which offers the possibility of inductively detecting interesting patterns.

5 Conclusions and outlook

This article has introduced a data-driven approach to the investigation of spelling variation. The metric proposed here combines the well-established measure of Damerau-Levenshtein distance with the concept of entropy and a scaled frequency measure. A few proof-of-concept analyses have shown how the method can be applied to investigate spelling variation from a diachronic perspective. The methodology is still open for refinement. One of the next steps could be the alignment of spelling pairs. Although it has been argued and shown that filtering also produces reliable results, the inclusion of more spelling pairs – which means a broader degree of data coverage – should at least be looked at more closely.

It is clear that many of the differences shown and discussed in Section 4 indicate variation on different levels, which is why this is just the first step towards a much needed further examination of the patterns and processes behind the data. The underlying developmental paths as well as intra- and extralinguistic factors which had an influence on these developments need to be identified next.

There are also some open questions, e. g. how to differentiate between differences which might have different causes, like <å> vs. <e> – is there a morphological reason to use the å, as in <glåubig>? Because the concept of relatedness in this data is restricted to the lemma, connections like this are harder to detect.

The attribute for each token which indicates its position in the text as already mentioned in Section 2 could be used to verify a literal interpretation of Voeste's rather interesting thesis, the existence of an exhortation to pursue stylistic diversification (Voeste 2008b). If the rapid reuse of a spelling is seen as being stylistically unpleasant, a higher distance in the text between the same spellings could be a sign of avoiding this. It would be interesting to test this hypothesis in future studies.

In case a different corpus with no preexisting normalization is analyzed, there are multiple approaches which – as preparatory work to use the proposed method – help mapping different spelling variants of the same word form onto each other, e. g. via detection of spelling variants as described in Barteld (2017) and Barteld et al. (2019).

For future research, automatic graphemic annotation tools with purposes other than (or additional to) normalization like the (semi-)automatic analysis of graphemes, graphematic syllables and the like for a subsequent comparison of writing systems are a methodological desideratum which, with the right training data, seem achievable. To avoid starting with uncertain data, these tools could at first be trained and tested with synchronous data; historical data (or other non-normalized data, such as in computer-mediated communication) could be one of the following steps.

A resourceful use of machine learning could significantly advance the field of graphematics, and perhaps bring it more into the spotlight.

Acknowledgment: I would like to thank Kristian Berg, Nanna Fuhrhop, Stefan Hartmann and an anonymous reviewer for their helpful questions and comments.

References

- Baayen, R. Harald & Fermín Moscoso del Prado Martín. 2005. Semantic density and past-tense formation in three Germanic languages. *Language* 81(3). 666–698.
- BBAW. 2019. *Deutsches Textarchiv. Grundlage für ein Referenzkorpus der neuhochdeutschen Sprache*. Herausgegeben von der Berlin-Brandenburgischen Akademie der Wissenschaften. Berlin. <http://www.deutschestextarchiv.de/>.
- Baron, Alistair, Paul Rayson & Dawn Archer. 2009. Word frequency and key word statistics in historical corpus linguistics. *Anglistik: International Journal of English Studies* 20(1). 41–67.
- Barteld, Fabian. 2017. Detecting spelling variants in non-standard texts. In *Proceedings of the Student Research Workshop at the 15th Conference of the European Chapter of the Association for Computational Linguistics*, 11–22. Valencia, Spain: Association for Computational Linguistics.
- Barteld, Fabian, Chris Biemann & Heike Zinsmeister. 2019. Token-based spelling variant detection in Middle Low German texts. *Language Resources and Evaluation* 53(4). 677–706.
- Bergmann, Rolf & Dieter Nerius. 1997. *Die Entwicklung der Großschreibung im Deutschen von 1500 bis 1700*. Heidelberg: Winter.
- Bollmann, Marcel, Florian Petran & Stefanie Dipper. 2011. Rule-based normalization of historical texts. In *Proceedings of the Workshop on Language Technologies for Digital Humanities and Cultural Heritage*, 34–42. Hissar, Bulgaria: Association for Computational Linguistics.
- Damerau, Fred J. 1964. A technique for computer detection and correction of spelling errors. *Communications of the ACM* 7(3). 171–176.
- Dipper, Stefanie & Sandra Waldenberger. 2017. Investigating diatopic variation in a historical corpus. In *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*, 36–45. Valencia, Spain: Association for Computational Linguistics.
- Dücker, Lisa, Stefan Hartmann & Renata Szczepaniak. 2020. Satzinterne Großschreibung von Substantiven und Substantivierungen in Hexenverhörprotokollen. In Renata Szczepaniak, Lisa Dücker & Stefan Hartmann (eds.), *Hexenverhörprotokolle als sprachhistorisches Korpus*. 113–144. Berlin & Boston: De Gruyter.
- Ebert, Robert P., Oskar Reichmann, Hans-Joachim Solms & Klaus-Peter Wegera (eds.). 1993. *Frühneuhochdeutsche Grammatik*. Tübingen: Niemeyer.
- Elmentaler, Michael. 2018. *Historische Graphematik des Deutschen. Eine Einführung*. Tübingen: Narr.
- Glaser, Elvira. 1985. *Graphische Studien zum Schreibsprachwandel vom 13. bis 16. Jahrhundert. Vergleich verschiedener Handschriften des Augsburger Stadtbuches*. Heidelberg: Winter.

- Jurish, Bryan. 2011. *Finite-state canonicalization techniques for historical German*. Potsdam: University of Potsdam dissertation.
- Koller, Gerhard. 1989. *Der Schreiber Albrecht Dürers*. Wiesbaden: F. Steiner.
- Levenshtein, Vladimir I. 1966. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady* 10(8). 707–710.
- Moscoso del Prado Martín, Fermín. 2014. Grammatical change begins within the word: Causal modeling of the co-evolution of Icelandic morphology and syntax. *Proceedings of the Annual Meeting of the Cognitive Science Society* 36. 2657–2662.
- Moscoso del Prado Martín Fermín. 2016. Vocabulary, grammar, sex, and aging. *Cognitive Science* 41. 950–975.
- Moser, Hugo. 1977. *Die Kanzlei Kaiser Maximilians I.: Graphematik eines Schreibus*. Innsbruck: Institut für Deutsche Philologie der Universität.
- Nerius, Dieter (ed.). 2002. *Die Orthographischen Konferenzen von 1876 und 1901*. Hildesheim: Georg Olms.
- Nerius, Dieter. 2003. Graphematische Entwicklungstendenzen in der Geschichte des Deutschen. In Werner Besch, Anne Betten, Oskar Reichmann & Stefan Sonderegger (eds.), *Sprachgeschichte. Ein Handbuch zur Geschichte der deutschen Sprache und ihrer Erforschung* (Handbücher zur Sprach- und Kommunikationswissenschaft 2.3). 2nd edn. 2461–2472. Berlin & New York: De Gruyter.
- Ruge, Nikolaus. 2004. *Aufkommen und Durchsetzung morphembezogener Schreibungen im Deutschen 1500–1770* (Germanistische Bibliothek Band 19). Heidelberg: Winter.
- Shannon, Claude E. 1948. A mathematical theory of communication. *Bell System Technical Journal* 27(3). 379–423.
- Shannon, Claude E. 1951. Prediction and entropy of printed English. *Bell System Technical Journal* 30(1). 50–64.
- Voeste, Anja. 2008a. Drucktechnik und Logographie. Die Etablierung von Schemakonstanz im Typographieum. In Corinna Laude (ed.), *Konzepte von Produktivität im Wandel vom Mittelalter in die Frühe Neuzeit*, 373–392. Berlin & New York: De Gruyter.
- Voeste, Anja. 2008b. *Orthographie und Innovation. Die Segmentierung des Wortes im 16. Jahrhundert* (Germanistische Linguistik Monographien 22). Hildesheim: Georg Olms.