Thomas Pilz, Wolfram Luther and Ulrich Ammon

Retrieval of Spelling Variants in Nonstandard Texts – Automated Support and Visualization

Abstract

This article describes ongoing research in the RSNSR[1] (Regelbasierte Suche in Textdatenbanken mit nichtstandardisierter Rechtschreibung, "Rule-based search in text databases with nonstandard orthography") project. The focus of this project is making historical text documents digitally available; consequently, it examines the challenges for digitization procedures and subsequent retrieval operations, like fuzzy full-text search. Difficulties are posed by scans of low quality facsimiles, old font types, inconsistent transcriptions and especially typical optical character recognition (OCR) errors and spelling variation. This article discusses recent solutions to such problems, concentrating on stochastic string edit distance measures, so-called evidences and the avoidance of static dictionaries. By presenting visualization approaches for retrieval in and browsing of historical databases and nonstandard text documents, as well as a prototype for visual evaluation of distance measures, it proposes a progression of information visualization in linguistics.

1. Introduction

In 2001 the Institute of Computer Science and the Institute of German Language and Literary Studies at the University of Duisburg-Essen began work on a joint project, Projekt Nietzsche-CD, which is aimed to create a digital literature archive with the reception of the German philosopher Friedrich Nietzsche. It is embedded in the scope of various literature research projects within the bachelor's/master's program Applied Communication and Media Science.

The realization of such a digital literature archive includes several working fields: a collection of literature assets, a web-based communication interface, digitization software supporting German black letter fonts, database design and implementation, a user-friendly system

[1] We would like to thank the Deutsche Forschungsgemeinschaft for supporting this research.

interface, a search engine for text documents in nonstandard spelling, administrative tools and a digital rights management system (Biella 2005). Furthermore, the literature archive should utilize library-oriented data standards for archival storage. Since the project's beginning numerous students from a variety of disciplines have participated in digitizing historical material dating from 1865 to 1945.

2. Digitization of historical documents

2.1 Optical character recognition

Even though the digitization of text documents is a standard procedure nowadays, it is still problematic. Since most of the photocopies of the documents were received by interlibrary loan, their quality is often less than moderate: shades, overexposure, skew and warping decrease optical character recognition (OCR) accuracy significantly. Even today the most reliable way to counter recognition errors is to manually revise the data.

Not only in the Project Nietzsche-CD but also in many other international projects, manual correction has to be limited due to restricted resources. Many retrodigitization projects focus on the constructional steps of the digitization process, which involve digitizing as well as tagging and aligning the text. For example, Compact Memory (www.compactmemory.de), a project working on the digitization of historical Jewish periodicals, combines an attractive interface with a respectable archive and is well used. But, as it is a publicly funded project, the operator cannot devote its resources to manually revising optical character recognition (OCR) errors in the digitized texts or to offering advanced search capabilities. A reliable search engine, however, is the means that makes the data fully accessible.

Users searching for the word *Fruchtbarkeit* 'fertility', for instance, will not be able to find a certain periodical from 1904 even though it clearly contains the word. Worse, they will not even realize that this text was missed. Because the full text aligned with the graphical representation of the text contains recognition errors, only the search for the misspelled word *Piuchtbaikeit* instead of *Fruchtbarkeit* finds the correct page (cf. Figure 1). Misinterpretation of the graph <r> as <i> is very common because of the graphical similarities of the two characters. Even though there are many possible recognition errors, only about 75 occur regularly.

Auch der *schone* indische *Blutenstrauch* Hibiscus rosa *smensis* sowie der als Stolz Indiens (Pride of *Jndia*) allbekannte *Zieibaum* Melia *azedaiach tiagen* neben schlingenden *üopischen* Winden, *gelbblutigen Buddleiastiauchein* etc dazu bei dass man glauben *mochte*, man sei *m* dem üppigen *Paike* eines indischen *Glossen* und nicht *m* einem Hotelgarten des „*^ustenahnlichen"* Palästina Aber auch die wenigen Reisenden die von Jaffa zu Wagen nach Haifa *fahien*, *meiken*, obgleich sie eine der *zukunftieichsten* Ebenen Palastinas *dmchieisen*, kaum *et^as* von *dei Piuchtbaikeit*, da die wenigen *judischen Kolomen* meist abseits der grossen Route liegen [..]

auffielen. Auch der schöne indische Blütenstrauch Hibiscus rosa sinensis, sowie der als Stolz Indiens (Pride of India) allbekannte Zierbaum Melia azedarach tragen neben schlingenden tropischen Winden, gelbblütigen Buddleiasträuchern etc. dazu bei, dass man glauben möchte, man sei in dem üppigen Parke eines indischen Grossen und nicht in einem Hotelgarten des „wüstenähnlichen" Palästina. Aber auch die wenigen Reisenden, die von Jaffa zu Wagen nach Haifa fahren, merken, obgleich sie eine der zukunftreichsten Ebenen Palästinas durchreisen, kaum etwas von der Fruchtbarkeit, da die wenigen jüdischen Kolonien meist abseits der grossen Route liegen, die zerstreut liegenden arabischen Ansiedlungen

Figure 1. Example of recognition errors (in italics) in the text (upper box) aligned with the graphical representation (lower box) taken from the Compact Memory database.

To make matters worse, many historical German documents were printed using German black letter fonts (Fraktur). These typefaces feature certain characteristics that are uncommon for modern fonts and pose a problem for standard OCR software. As shown in Table 1 typical recognition errors are likely to differ between different typefaces. While, for example, <ei> in Antiqua will hardly be misinterpreted as <ü>, such an error is probable in Fraktur or Textur where <ei> and <ü> are designed with similar characteristics.

Table 1. The various typeface designs have differing probabilities for recognition errors.

Antiqua	Rotunda	Fraktur	Textur
u-n	u-n	u=n	u—n
ei-ü	ei-ű	ei=ü	ei=ü
z-g	ȝ-g	ȝ=g	ȝ—g

There are partial solutions for recognition errors in general and Fraktur in particular. A preprocessing module for binarization, component analysis, skew correction and de-warping of digital text documents has been developed (Mischke & Luther 2005). Analysis and preclassification of words and letters, localization with vertical bar patterns and the combination of different recognition approaches provide the high quality retrieval of keywords selected by literary scholars on Fraktur documents (Mischke 2007). Full text search operations are still highly problematic, even with elaborate algorithms, especially if the sources are of poor quality. The commercial product ABBYY FineReader XIX (Abbyy 2004) certainly yields good results but only with a costly license.

2.2 Spelling variation

While spellings caused by faulty character recognition are errors per se and OCR programs attempt to avoid them, spelling variation – whether intentional or unintentional – cannot be categorized so easily. It is worth mentioning that there seems to be no general definition of spelling variants yet, even though everybody seems to have an intuitive apprehension of its meaning. Many spelling variants we encounter today are the result of dialects or language varieties. Since dialects are mainly practiced orally, they are generally of minor importance in standard document retrieval. Comparison, classification and retrieval is done mostly on the basis of phonetic transcriptions (Nerbonne & Siedle 2005). Nevertheless, dialectal text production has always existed. Famous fictional examples are Lerner and Loewe's *My Fair Lady* (cf. "Wouldn't It Be Loverly?") or Gerhart Hauptmann's *Der Biberpelz*. Standard varieties feature not only spelling variants but whole new words. A dictionary of standard varieties of

German in Austria, Switzerland, Germany and other countries is available (Ammon et al. 2004).

In contrast to (synchronically) diatopic variation (through space), diachronic variation (over time) is often encountered when dealing with text production. For the greatest part of any language's development, written resources represent the only source of linguistic information because spoken evidence simply does not survive. Thus, it is all the more astonishing that until the last century many linguists regarded the written form of language as secondary in the meaning of less relevant (cf. Fleischer 1966: 8). Luckily nowadays historical spelling variation is a well researched topic (cf. Elmentaler 2003).

Historical German spelling variants existed officially as long as German orthography was not standardized. The Second Orthographical Conference in Berlin announced formally binding regulations in 1901. But even today we have competing spellings as a result of resistance to the spelling reform of 1996, for example, *Gesichtscreme, Gesichtskreme* and *Gesichtskrem* 'face cream' and *Potential* and *Potenzial* 'potential'. Such spellings may of course have different status. Even though all five spellings are indeed official (cf. Duden 2004), *Gesichtskreme* and *Gesichtskrem* are rarely used. But phenomena of historical and regional spelling variation are by no means an exclusively German problem. Similar problems are documented for numerous other European languages as well, including Dutch, English, French and Slovenian. Consequently, when performing search operations on nonstandardized texts, one needs to have profound knowledge of historical spelling variation for successful retrieval.

While variation in German was already limited in the 19th century, the frequency of variant spellings increases significantly with the age of the text documents[2]. Texts on the outer limits of High German, for instance, may contain up to 60 percent nonstandard spelling tokens (Kempken et al. 2006, see below).

We define a spelling variant as an alternating signifier of a signified word variable – in de Saussure's understanding – where both belong to the same word family. Therefore, both are identical in inflection and

[2] Unless otherwise noted, the following statistics are based on calculations from our manually collected database of spelling variation, which contains 12,697 entries. A thorough statistical analysis is given in section 8.

derivation. Morphology-based variation or variation in vocabulary can be understood as "variation in a broader sense".

It is important to note that a spelling variant alternates only on the level of encoding, as an additional identifier. Thus, the standard spelling related to, for example, the singular accusative masculine *bankerotten* is not the lemma *bankrott* 'bankrupt' but *bankrotten* in identical declension. In older texts, an increasing number of obsolete words occur that might have a translation but no related standard spelling of the same word family; for instance, a 15th-century German text featured the word *bemelcht*, which was used in the sense of 'referred to as'.

Even more important than the percentage of spelling variants in a text document is the form of their variation. In the 19th century only a few major letter replacements occur, including

<k> - <c>, *Punktation – Punctation* 'punctuation'

<t> - <th>: *teilen – theilen* '(to) separate'

<ä> - <ae>: *Änderung – Aenderung* 'change'

<ie> - <i>: *ignorieren – ignoriren*. '(to) ignore'

Even though the average number of letter replacement operations per word increases only slightly from ~1.3 in the 19th century to ~1.8 in the 14th century, the possible replacements are multiplied. Koller, for example, identified nine different substitutions for <i> in Early High German Texts (cf. Table 2). Comparing the most frequent letter replacements in historical texts, it can be seen that between 1800 and 1900 about 80 different replacements were commonly applied. Between 1700 and 1800, there were 145; between 1600 and 1700, 167; between 1500 and 1600, 214; and, between 1200 and 1500, 295. This shows that the degree of variation – the possible spellings a historical writer could choose from – increases significantly with the age of the text.

Additionally, the maximally occurring number of replacements per word also increases considerably. In 19th-century texts, the variation maxima, that is, the words with the most replacements, vary between two and five operations per word (for example, *räsonierendes – raisonnirendes* 'arguing') with an average of ~3.41. In the 18th century this average value climbs to four, and in the 17th century words occur with eight or more replaced letters (*domprobst – thuembbröbst* '*cathedral provost*').

Table 2. Examples of letter replacements in Early New High German (Koller, 1989, Source: Munske, 1997).

Graphemes	Letter replacements									
<i>		<i>	<ie>	<ieh>	<ih>	<j>	<jh>	<y>	<Ÿ>	<Ÿe>
	%	64.7	3.9	0.1	0.2	16.5	0.1	6.3	8.3	0.1
	Examples: ir, ihr, jr, jhr, Ÿr (rounded values)									
<f>		<u>	<v>	<f>	<ff>	<ph>				
	%	0.3	22.6	55.5	21.4	0.2				
	Examples: fux, vux, pulver, pulfer, brif, briff									
<u>		<u>	<uh>	<ue>	<ů>	<v>	<w>			
	%	48.9	0.3	0.2	0.2	37.7	12.7			
	Examples: und, vnd, wnd, guet, gůt, fuhr, fůr									

To determine where this progressivity in variation comes from one has to take a closer look at text production in bygone times. The following example is taken from the work *Gründtlicher Bericht Von einem vngewohnlichen Newen Stern* (*De Stella Nova*, 1604) by the German astronomer Johannes Kepler (1561–1630).

> *Demnach nunmehr zwey vnd dreyssig* (*zweiunddreißig* 'thirty-two') *Jahr/ das die Astronomi etwas newes* (*Neues* 'new')*/ zuvor in allen Büchern/ so viel deren auff vns* (*auf uns* 'on us') *gelanget* (*gelangt* 'arrived at')*/ vnvermeldetes wunderwerckh* (*unvermeldetes wunderwerk* 'unreported marvel') *am Himmel befunden/ das nemlich* (*nämlich* 'namely') *ein newer* (*neuer* 'new') *sehr grosser* (*großer* 'large') *heller gläntzender Sterne* (*glänzender Stern* 'brilliant star') *vnder* (*unter* 'under') *die höchste Sphaeram vnd vnbewegliche* (*und unbewegliche* 'and fixed') *sterne in sydere Cassiopeae vnd* (*und* 'and') *der Jacobsstrassen* (*Jacobsstraßen* 'Jacob's Street' [as the Milky Way was also known]) *oder via lactea einkhommen* (*eingekommen* 'came in')*/ alda* (*all da* ,there') *in die 16. Monat lang an einem ort still gestanden/ vnd entlich widerumb* (*und endlich wiederum* 'and finally again') *verschwunden ist (...)*[3]

The simplest forms of spelling variation in Kepler's text occur because of phonetic similarity of graphemes (*nämlich – nemlich* 'namely', *endlich – entlich* 'finally') and are a logical result of a lack of standardization. The older the texts are, the more frequent are the representations of slightly different pronunciations (*wiederum – widerumb* 'again'). While some forms of variation are still quite common for German native speakers

[3] Nonstandard spellings are underlined; standard spellings and translations are in brackets.

because they still appear in family names (*zwei – zwey* ‘zwei’ as in the name Meyer) or poetry (*gelangt – gelanget* ‘arrived at’), other forms are completely obsolete in the modern standard. Good examples are variants featuring grapheme-phoneme correspondences that are invalid today. For instance, the <ew> in *newes* ‘new’ corresponds to /oi/; today, this phoneme is represented by the grapheme <eu>. Similarly, <v> in *vnd* ‘and’ corresponds to /u/, the modern <u>.

Another example of obsolete spellings is Barocke Letternhäufelung (Baroque letter accumulation). The aesthetic principle of orthography (Maas 2000: 48) aims to embellish the type face. The word *Hoheit* ‘highness’ is a compound of *hohe* ‘high’ and *heit* ‘being’ and should therefore be spelled *Hohheit*, but the aesthetic principle perceives the accumulation of <h> as unpleasant. Contradictory perceptions of this principle in different times are not overly surprising. In the 17th century Barocke Letternhäufelung was a method of decorating words as Kepler does in *wunderwerckh* (instead of the standard *Wunderwerk* ‘marvel’).

As mentioned above, spelling variation can be found in other European languages as well. Koolen et al. (2006: 409) state that spelling in *Middelnederlands*, a form of historical Dutch spoken during the Middle Ages, was based on pronunciation, which again varied in different regions of the Netherlands. Dutch became more uniform in the 17th century but was still a “collection of dialects” (Vandenbussche 2002), spelling variants like *heyligh* (standard: *heilig* ‘holy’) prevailed. Various systems of orthography continued to change spellings throughout the 19th and 20th centuries (cf. Table 3). In 1996, for example, rules for the composition of words were changed, and *pannekoek* became *pannenkoek* ‘pancake’.

Table 3. Spelling norms of three Dutch phonemes in five spelling systems (cf. Vandenbussch 2002: 31, excerpt).

Phonemes	Des Roches 1761	Siegenbeek 1804	Behaegel 1817	Commission 1844	de Vries & te Winkel 1864
[i:]	<ie> <y>	<ie> <i> <ij>		<ie> <y> <i>	<ie> <i>
[ɛi]	<ey>	<ei> <eij>	<ey>	<ei> <eij>	<ei>
[œy]	<uy>	<ui> <uij>	<uy>	<ui> <uij>	<ui>

Medieval French texts pose similar problems. O'Rourke et al. (1997) give the example of the name of a chief villain spelled variously *Hoiaus*, *Hoiax*, *Hoiel* and *Oiaus* in the poems they edited. Rayson, Archer and Smith collected a list of 45,805 English spelling variants from 17th-century newspapers, the Oxford English Dictionary and 18th- and 19th-century fiction (Rayson et al. 2005). As in French, Dutch and German, there often is a considerable amount of variation (*maintenance – mayntaynaunce*).

A case that does not occur in Kepler's text is obsolete graphs, that is, letters not within the modern German alphabet, like the digraph[4] <ů>. Early New High German texts regularly use <ů> in the period of passage between the Middle High German diphthong <uo> and the New High German monophthong <u>.

2.3 Manual transcription

This leads directly to the third kind of variation we will focus on, after OCR errors and spelling variation. Because the Latin alphabet was used for the spelling of German words, specific digraphs had to be employed for the identification of non-Latin sounds. When those words are transcribed in the process of digitization, diacritics in particular pose problems. At least from

[4] Following Elmentaler (2003), graphs consisting of a single letter are labeled monographs, and two letters (such as <eu>) or a letter and a diacritical mark (like <ů>) are labeled digraphs.

a historical linguist's point of view, the worst thing to do is to simply omit the diacritic (for example, transcribing *zů* as *zu* 'to') and thus lose a historical variant. Changing *zů* to *zuo* improves the situation only slightly because the digraph <uo> also exists in historical texts. To transcribe it as *zu^o*, as programmers often paraphrase the square of a number (n^2 = n^2), is quite common and preserves the information of the diacritical mark. It involves a logographical form, however, that is independent of the German language. Furthermore, the circumflex <^> is not uncommon in recognition errors as a misinterpretation of <v> or <w> (for example, *worden* - *^oiden* 'was', *von* - *^on* 'of'). The best solution would be to use the current Unicode Standard, Version 5.0 (http://www.unicode.org). The digraph <ů> is defined in the chart Latin Extended A as 016F; it can also be built using the Combining Diacritical Marks in range 0300–036F with the codes 0075 (u) + 0366 (°). Those codes can – and often have to – be used in HTML texts as well; while there is the entity definition *å* for <å>, *&uring* is not interpreted. But even Unicode poses problems because the codes – especially combined codes – are often interpreted incorrectly. The MS Internet Explorer 7.0 omits many diacritics, and Mozilla Firefox 1.5 displays graph and diacritical marks consecutively (cf. Figure 2).

Internet Explorer			Firefox		
o□	O□	773	o̅	O̅	773
o̍	O̍	774	o̍	O̍	774
o̲	O̲	775	o̲	O̲	775
o̎	O̎	776	o̎	O̎	776
o̊	O̊	777	o̊	O̊	777
o□	O□	778	o̊	O̊	778
o̎	O̎	779	o̎	O̎	779
o̍	O̍	780	o̍	O̍	780
o□	O□	781	o̍	O̍	781
o□	O□	782	o̎	O̎	782

Figure 2. MS Internet Explorer 7.0 (left side) fails to display several diacritics, while Mozilla Firefox 1.5 (right side) cannot combine codes.[5]

[5] This test was performed using the Test for Unicode support in Web browsers (http://www.alanwood.net/unicode/combining_diacritical_marks.html).

To summarize the perceptions of sections 2.1–2.3, the words of a nonstandard text document are separated into

a) words without a related standard spelling in the understanding of our definition (cf. Section 2.2),
b) variant spellings (which include all types of variation, even recognition errors) and
c) standard spellings.

There are cases in which it is difficult to assign words to one of these classes. The Middle High German word *knicht* seems to be a spelling variant of *Knecht* 'servant', and the two words are indeed etymologically related. However, the correct translation of *knicht* is *Ritter* 'Knight' and, thus, it belongs in class (a).

All variant spellings have one important issue in common: They are related to a standard spelling by more than just their meaning. Their concrete characteristics can be manifold regarding their type (for instance graphical or phonological) or cause of variation (such as dialect or historical development): they may even cover deliberate variation, like Leetspeak. While words without related standard spellings are, of course, interesting, the processing of variant spellings is the most challenging issue algorithmically.

To summarize our insights regarding the problems of recognition errors, spelling variants and varying transcriptions the older the text, the more frequently the following issues occur:

1) The total number of letter replacements increases because of the original's older font types and poor states of preservation, the lack of standardization and the involvement of obsolete letters.
2) The maximum number of replacement operations per word increases (that is, variants become increasingly different from standard spellings).
3) Therefore, the number of possible variants relating to a single standard spelling increases.
4) As a result, search tasks on nonstandard texts become increasingly difficult and require specific handling.

3. The RSNSR project

The RSNSR (Rule-based search in text databases with nonstandard orthography) project, which was funded by the German Research Foundation (DFG), was initiated in 2005 to provide a reliable and flexible full-text search engine for the documents of a prior project, the Projekt Nietzsche-CD (cf. Figure 3), and similar material. It was our intention not to rely on dictionaries – an approach that is different from most capacious glossary projects, such as the digitization of the famous Deutsches Wörterbuch (DWB) by Jacob and Wilhelm Grimm, which is maintained by the University of Trier in Germany (Christmann & Schares 2003).

Making use of extensive wordlists surely has its advantages, especially in processing speed. But even though corpora and dictionaries of many millions of words in standard spelling exist, they will never be complete because German is an inflecting language making extensive use of composition and is, therefore, by definition infinite. Dictionaries of historical words are much rarer and much smaller – even though the possibilities for variation are enormous. Through this avoidance of wordlists, we expect an increased recall ratio, especially with documents of highly varied spelling. Furthermore, the additional expenditure of manually adding word-relations is eliminated.

While at first it focused on data from 1865 to 1945, the RSNSR project soon started to broaden its perspective, reaching further back in time. In order to have a basis to work on, we manually collected pairs of standard and variant spellings from historical texts. Provided with metadata about their origin (time, location) and type (caused by OCR, not caused by OCR), we called the pairs *evidences* because they bear evidence of variation. In the same way, we built a collection of synchronic spelling variants. The texts from which we extracted the evidences came to us courtesy of the Bibliotheca Augustana, Compact Memory, Digitales Archiv Hessen-Darmstadt and documentArchiv.de.

Our constantly growing database of evidences currently features 12,697 entries from 107 different texts. These originate from all over the German-speaking area and date from 1293 to 1919. The spelling variants therefore cover diachronic language development, diatopic variation, differences in transcription and evidences of OCR errors. Among the latter are variants from antiqua as well as black letter sources.

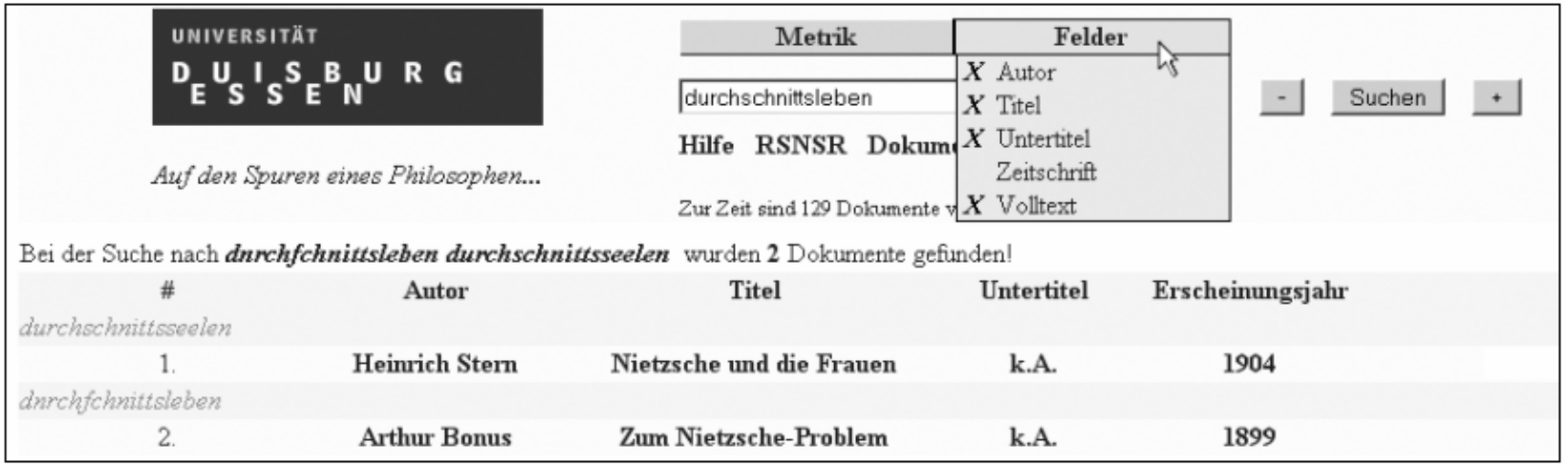

Figure 3. The improved interface of the second edition of the online Nietzsche search engine.

With the information gathered from this database and our algorithms in development, a search engine is no longer our only goal; new ways of displaying the results of a search query allow for additional information and overview. We used the renowned Java package for information visualization called Prefuse (http://prefuse.org). Information Visualization is a fairly new field of research and is rapidly evolving. A well established definition of information visualization is "the use of computer-supported, interactive, visual representations of abstract data to amplify cognition" (Card et al. 1999).

When performing fuzzy search operations, the classic ranking of results we know from our daily Web searching via Google may no longer be the best visualization of results. When searching for "imprisoned", which variant spelling is the "better" result, *imprison'd* or *imprisonde*? Both occur in historical English documents of the same era. Even though computers can be employed to ease retrieval tasks, should it be for a machine to decide what the user is looking for? Figure 5 shows an interface for retrieval on historical documents. It focuses on the different kinds of spelling variation rather than on the documents themselves. Users can explore the trees to the right of the spelling variants to see who used those spellings when and where.

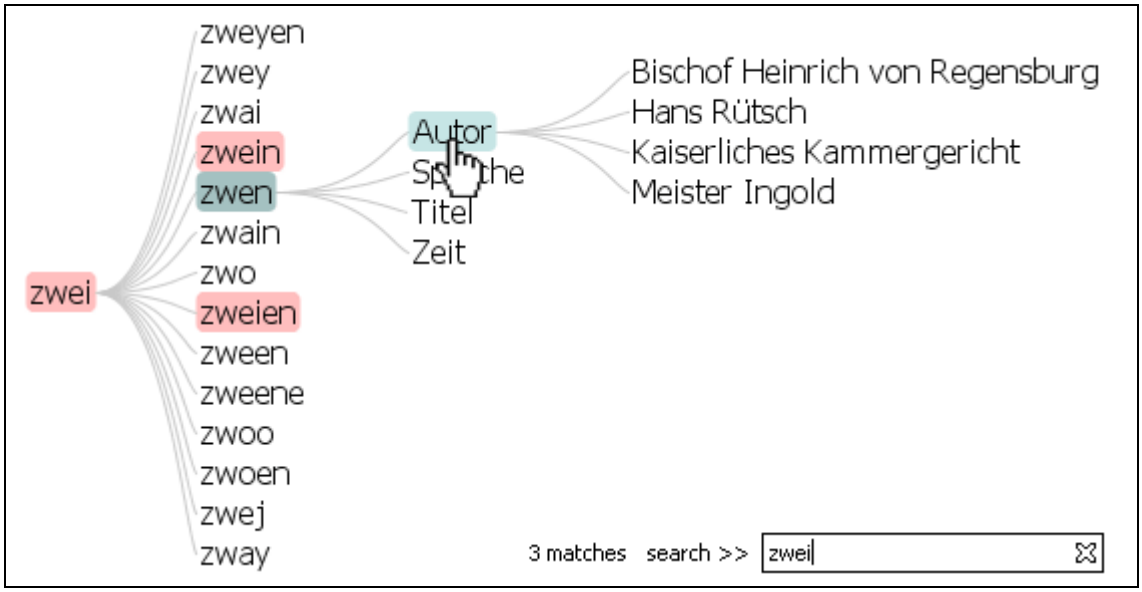

Figure 4. An experimental search interface for tasks involving variant spellings.

For browsing databases of nonstandard spellings, like historical dictionaries, even more overview is needed. Since all spellings are already in a database, their relations can be preprocessed, in contrast to browsing

arbitrary texts. Figure 6 shows the browse of a portion of our database. It is fully zoomable and dragable and features a lens function (seen on the five enlarged spellings). The forces pushing spellings apart or pulling them together are fully adjustable. With this configuration, the browser shows all entries in the database, whether standard or variant spelling, aligned by a simple Levenshtein distance measure (see below). The user can explore the vicinity of interesting words (here, for example, spelling variants of *tausend* 'thousand', which are similar to variants of *tugend* 'virtue').

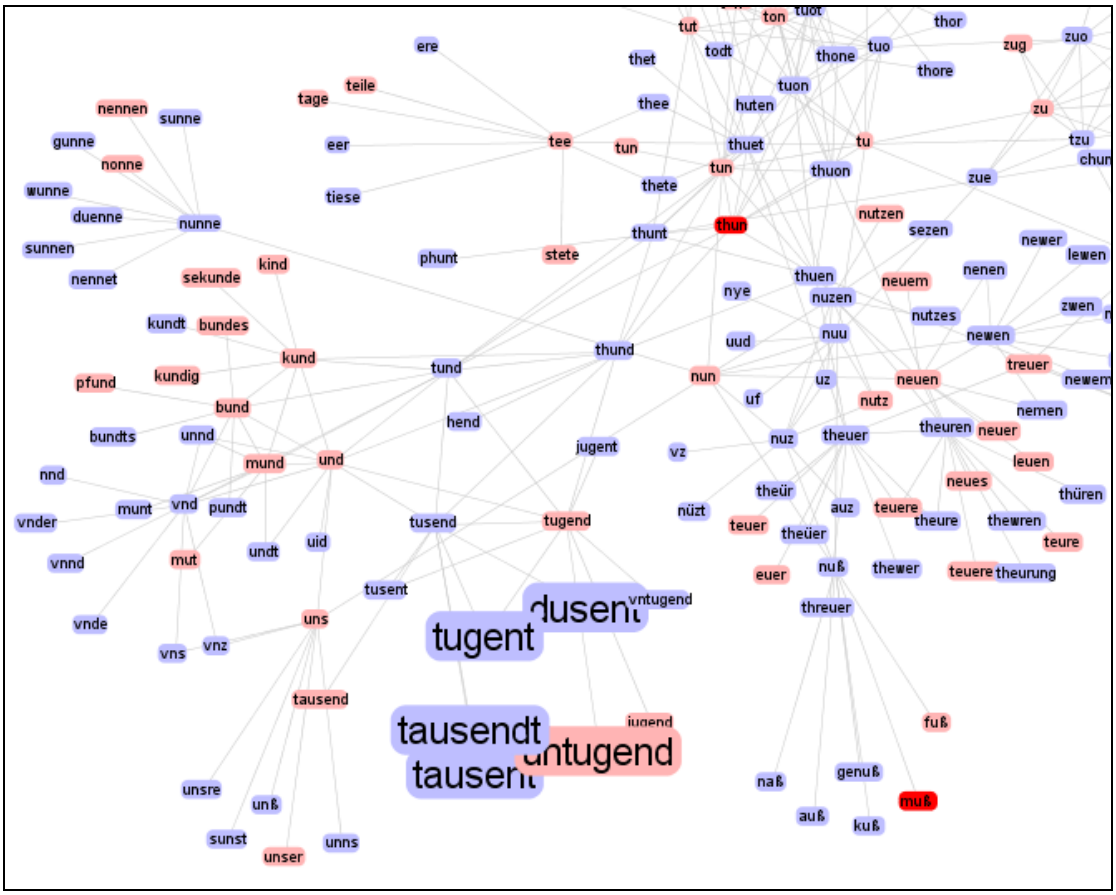

Figure 5. A simple browser for historical databases.

Similar in origin to the interface in Figure 5, the Word Explorer prototype (cf. Figure 7) allows the examination of spellings with high variance and multiple connections. It distinguishes between a standard spelling (in the center of the "stars") and the spelling variants (surrounding the standard spellings). Even though string edit distances are not represented, the numeric values are displayed when spellings are selected. In this example

the variants of the infinitive *wollen* '(to) want', its simple past form *wollte* 'wanted' and the second person plural *wollt* '(you) want' are displayed. Here, users of this interface will see that the spelling variant *wölle* can be both a variant of *wollen* and of *wollte*.

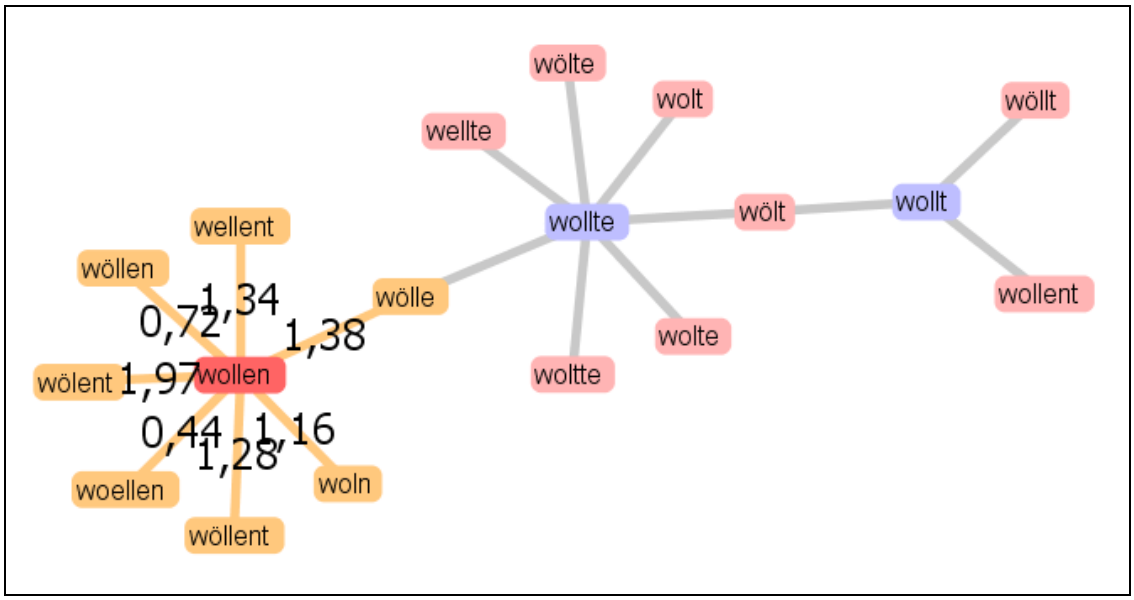

Figure 6. Interface of the Word Explorer prototype for examination of spellings with high variance and multiple connections.

Visualizations like the ones presented in Figures 5–7 can be very useful in literature information systems (LIS). Furthermore, we are certain that our algorithms can also be employed for automatic text categorization alongside authorship attribution methods, like stylometrics, the analysis of a text's internal statistics (Holmes, 1998) and entropy coding (Benedetto et al. 2003). This topic is currently being researched. (Semi-)automatic evidence retrieval in combination with automatic correction of recognition errors has been investigated (Wedershoven 2007). The detection of nonstandard spellings in a text is a rather simple matter of comparison with large dictionaries and inflection tables (such as Deutscher Wortschatz or Canoo). All spellings not found in those databases are potential spelling variants. It is much more complicated to find the correct standard spelling corresponding to a spelling variant or recognition error. Even though related to retrieval on nonstandard texts (input: standard – output: spelling variant), the methods cannot be transferred without adaptation. In some cases, it is even harder to decide whether a spelling variant was caused by historical/regional variation or misrecognition. A spelling **ungcrn* (*ungern* 'reluctantly') is most certainly a recognition error caused by the graphical

similarity of <e> and <c>, but *vngern* can be both, because <u> is often replaced by <v> in old texts.

Knowledge derived from analyses of large databases of recognition errors can help with the decision. Pollock and Zamora, for example, reported that in only 3.3 percent of the 50,000 words they examined was the first letter misrecognized (Pollock & Zamora 1983). For historical spellings, however, this finding does not apply; when we examined our database, we found that 13.7 percent of misrecognitions occurred in the first letter.

4. Generation of spelling variants using manual rules

In our research we examined two contrary approaches:

- The *generation* of possible spelling variants. A fraction of the spellings generated correspond to known historical spelling variants. These variants are called "established spellings".
- The *measurement* of word distance using string edit distances.

In the first stage of the project, we started with the manual composition of rules. Linguistic replacement rules are successfully used in a variety of programs, such as VARD (VARiant Detector), an existing English system (Rayson et al. 2005).

Using Sun's regular expressions formalism[6] (*java.util.regex*) with minor extensions to ease the input of linguistic data, we built 68 replacement rules. These consist of 62 different sequences and, in parts, historical *n*-graphs (like <a>, <äu> and <eau>). In contrast to the first edition of the online Nietzsche Archive mentioned above, these rules are fully able to support context sensitivity. The rule *%K% #ö|eu# [tb]*, for example, can be interpreted as "*If a consonant sound (%K%) on the left and <t> or <b> on the right ([tb]) surrounds an o-umlaut (ö), then replace the <ö> with <eu> (ö|eu)*".

Figure 8 shows the derivation tree of a typical variant generation algorithm. The gray nodes are spellings not found in our database. Of course, this tree is a simplified example, even though the nodes with dates in the brackets are existing spelling variants taken from our database. In

[6] http://java.sun.com/docs/books/tutorial/essential/regex/

reality there are 19 different documents containing the spelling *zwey*, not just one. There also are other variants of *zwei* 'two', like *zwoo*, not listed here. We even discovered the interesting fact that the spelling *zweyen* is not only a variant of *zwei* but also a variant of the inflected standard form *zweien*, which itself is a variant spelling of *zwei*.

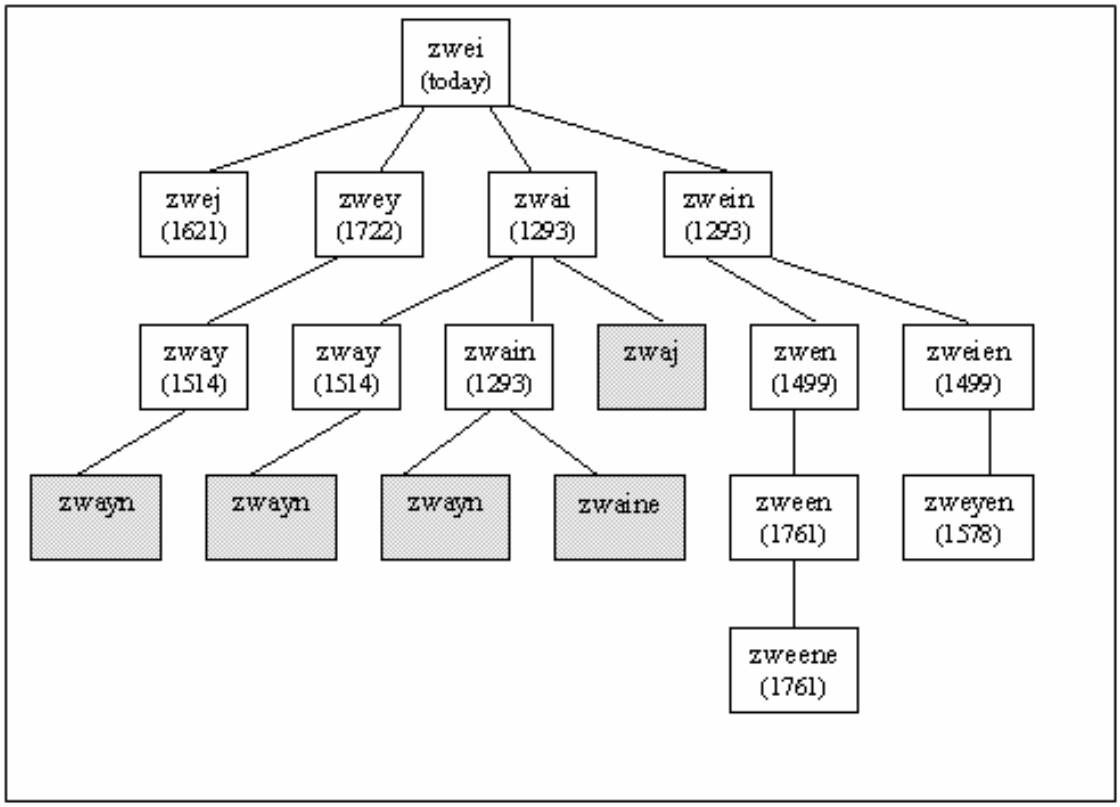

Figure 7. Example of a derivation tree for the standard spelling *zwei* 'two'. The numbers in brackets depict selected dates of documents using the variant spellings shown. Gray nodes are hypothetical variants not yet found in historical documents.

Looking at the example, we can see the main cases we encounter in variant generation:

- Not all spellings generated by the rules are found in our database. Even though this is exactly what we want, because – as mentioned above – a database will never contain all possible spelling variants, even simple rules build an enormous number of new variants. It is possible that most of these do not occur in any existing text.
- A large number of redundant spellings are produced on different paths.

5. Displaying generation rules with treemaps

In Kempken et al. (2007) we presented a treemap approach to displaying details of such single word derivations. The treemap visualization serves five purposes:

- It allows the detection of relevant rule sequences. A sequence of rules is considered relevant if it leads to an actual historical spelling (established spelling). Irrelevant sequences should be pointed out in parallel.
- It makes it easy to find permutations of rules that produce the same spellings.
- It discerns patterns to describe characteristics of nonstandard orthography (depending on location and period).
- It enables the derivation of upper bounds for the length of relevant rule sequences.
- It provides a means of accessing extensive amounts of information about one spelling.

Johnson and Shneiderman (1991) developed the treemap algorithm in 1991 for visualizing hierarchical data structures. Their original slice-and-dice approach defines a 2D-space–filling technique for mapping a hierarchical structure into nested rectangles: A rectangular area is recursively subdivided into a set of smaller rectangles alternating between vertical and horizontal subdivision. Each rectangle represents a node of the tree and the enclosed subrectangles correspond to all descendants of this node. The subdivided areas can be given specific size, color or texture. In this way, it is possible to display additional properties of the corresponding tree node. Since his original algorithm was introduced, many have tried to make the treemap approach more effective in visualizing an information hierarchy through such methods as using other space-filling techniques or extra navigation help on the tree structure. Shneiderman (2006) gives an overview of different implementations and applications of the treemap visualization approach. That treemaps are not limited to a few thousand items was proven by Fekete and Plaisant (2002).

For the construction of a treemap of spelling variants, we derive candidates for historical spellings from a current standard spelling by

recursive application of rules. In each step, one or more new spellings for the next step are produced, as shown in Figure 8.

Each derivation node is therefore described by three key properties: the original spellings, the applied rule and the newly produced spellings. Due to the recursive nature of the process, the original spellings are always the ones produced in the previous step. In order to optimize the rule set, we analyzed the rules involved in the derivation process, taking into account the following key aspects:

- *Applicability*. The application of a given rule is restricted to a specific context. The less restrictive this constraint is, the more spellings a rule can be applied to. Hence, the applicability of a rule depends on its context.
- *Productivity*. One rule may produce more than one derived spelling. As rules are always applied to all variants contained in a node, the number of spellings produced also relies on the rule's applicability. Thus, both account for its productivity. A certain rule set may produce established spellings, that is, spellings found in historical texts. Minimal subsets with this property should be identified.
- *Commutativity*. Another interesting aspect is commutativity. In some cases, two or more rules may be applied independently. For example, consider a rule A that is applied to an original spelling. Another rule B may afterwards be used to transform all of the results of A and yield new spellings. If this process can be reversed in such a way that rule B is applied first, rule A is applicable to all the results and the results of both are constant, the order of rule application is no longer important, and the rules are considered commutative. If this property can be proven for a set of rules, the derivation process can be sped up significantly. After the results of the application order A-B are determined, the results of B-A no longer need to be derived but can be looked up. Of course, this feature of a rule set has to be proven by using the formal rule definition, but a firm visualization may provide important clues as to which rules may be commutative.
- *Redundancy*. One rule may foil the results produced by another. For instance, one rule may insert an additional <e> whereas another rule removes it. Thus, the application of either leads to no new variants. It is also possible that for the same spelling to be produced on different paths (for example, **zwayn* via **zwey* or **zwai*, as in the example above).

Analogous to the considerations above, the derivation process can be curtailed in such cases. Thus, one goal of the optimization process is to identify redundant rules and prevent useless work, by such means as restricting rules to a more specific context.

- *Dependency*. A rule may not be applicable to original standard spellings but require the previous use of another rule. Subsequently, it can be applied only to the results of the previous rule. As a result, spelling variants are produced in different levels of the tree (for instance, **zwej* in level 1 and **zweene* in level 4). Additionally, inner nodes as well as leaf nodes can contain relevant variants, but it is also thinkable that some inner nodes are just transitions

We implemented a Java application that uses the treemap approach to show the key aspects of rules involved in the treelike derivation process in an interactive presentation. The productivity of a rule is indicated by the size of the corresponding shape. The squarifying algorithm (Bruls 2000) arranges the rectangles according to their hierarchical order.

We have designed several views to point out different aspects of the derivation process. The color assignment for the views without special coloring (see below) was defined corresponding to Table 4. Since selection presupposes derivation, all nodal states can be represented by this color scheme. Light green and orange apply only to redundancy visualization. The color is assigned according to three attributes:

- *Established*. If any of the spellings associated with a certain rectangle has actually been found in a historical text, we consider this spelling established. The corresponding form is highlighted.
- *Selected*. In most of our visualization approaches, the user is able to define constraints on the derivation process. Hence, only a subset of all rectangles is selected. The selected subset is expressed by a different color.
- *Redundant*. If any of the spellings associated with a rectangle can be otherwise derived, that is, if it is already contained in the selected subset, it is considered redundant.

Table 4. Color scheme for treemap visualization.

Color / Meaning	Established	Selected	Redundant
Gray	No	No	No
White	Yes	No	No
Yellow	No	Yes	No
Light green	Yes	Yes	No
Orange	No	No	Yes
Dark green	Yes	No	Yes

The potential of our treemap visualization approach can be seen in the following two examples. A typical screenshot of the implemented tool is shown in Figure 9. Here, the user is able to interactively select a subset of the rules. The nodes that can be derived using this subset are highlighted in yellow or green if the respective spelling is established. Additionally, all the spellings that can be derived with this subset – whether established or not – are highlighted in orange or dark green respectively. The main advantage of this approach is that the user may interactively select a rule subset and redundant rule applications are immediately highlighted according to the selected scheme. Hence, a typical rule set optimization task is to find a minimal rule subset such that all established spellings are accentuated either in light or in dark green, meaning the spellings (not necessarily the nodes) can be derived using just this subset.

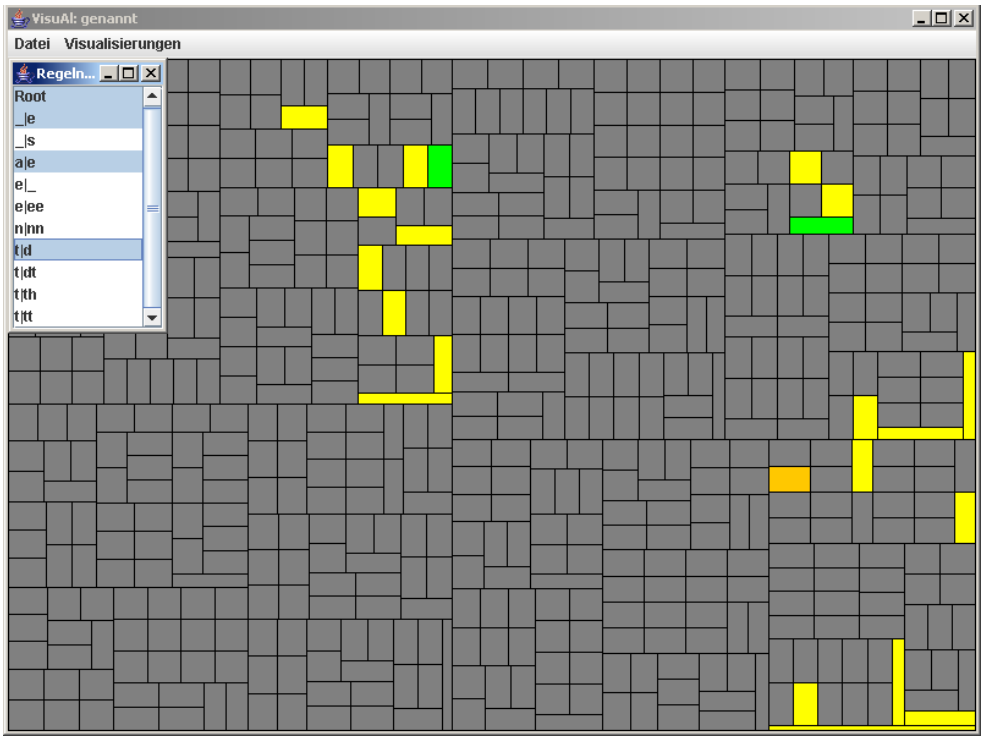

Figure 8. Redundancy view with some rules selected.

The mixed rainbow view is another of eight available views and is depicted in Figure 10. Each rule is assigned a color, and the color of a rectangle is then determined by the mean value of the colors of the affected rules. Hence, the influence of particular rules in the overall derivation process can be displayed in parallel. Of course, mapping the rule combination into the RGB color space can only provide an impression of the rule set's structure. Even color spaces with higher degrees of freedom can represent the information only marginally better.

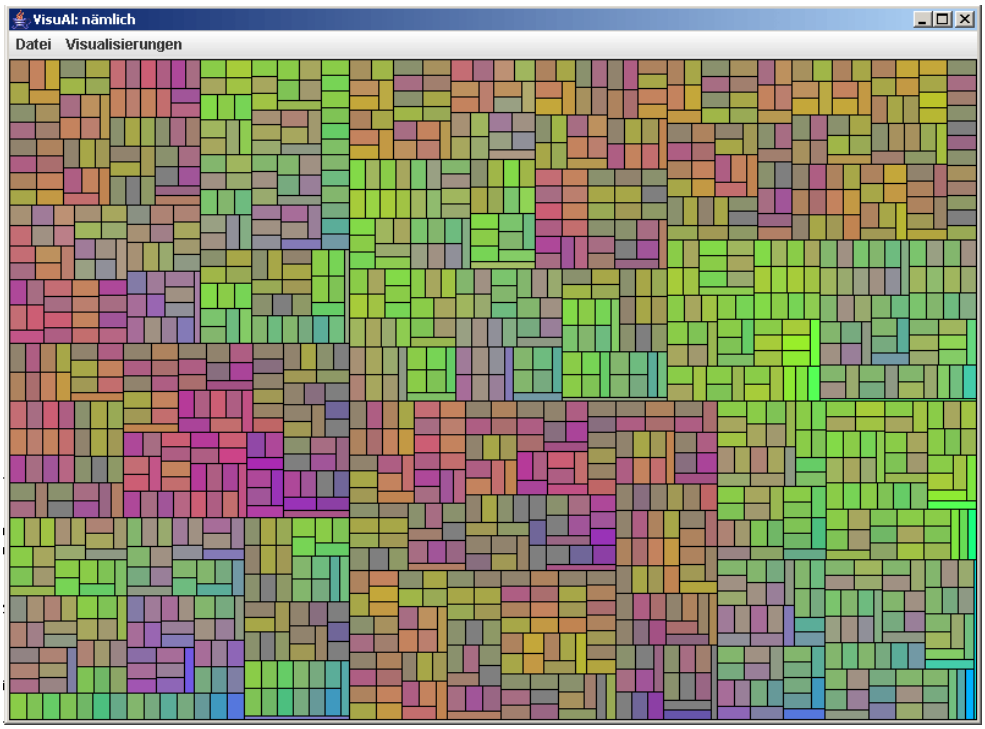

Figure 9. Mixed rainbow view showing predominant influences of the "red" and the "green" rule.

However, the design of a rule set for the period from 1803 to 1806, which was based on only 338 pairs of evidences, took about three days to create. Dawn Archer spent more than a year creating the letter replacements for VARD. Koolen et al. (2006: 409) recount similar experiences for historical Dutch. If an approach is to be applicable in inhomogeneous scenarios, the manual construction of replacement rules is simply not affordable. At the same time, manual rule derivation is prone to human error. This is especially true once the rule set exceeds certain limits, where unexpected side effects become more and more likely. As a result, automatic approaches became of interest.

6. Distance measures

Comparing different spellings of the same word often gives rise to the question which spellings are more similar than others. Similarity and difference can both be expressed as a function of distance. However the distance between words is not fixed. Is *aufwändig* more similar to

aufwendig 'elaborate'[7] than *Jngenieur* is to *Ingenieur* 'engineer'? While most today's native German speakers would agree that it is, a time traveler from 1750 quite certainly would not, because the perception of grapheme-phoneme correspondences in the 18th century was different than it is today (cf. Section 2.2). Distance measures help to answer such questions by calculating the distance between two words. String edit distance is defined as the minimum number of character replacements, insertions and deletions required to transform the one string into the other. In 1965 Vladimir Levenshtein presented a recursive algorithm for calculating edit distance. A more efficient way is to use a dynamic programming approach, as described by Wagner and Fischer (1974). String edit distance is widely used in a variety of applications as it can be determined efficiently and delivers good results. Another type of string distance measure relies on the comparison of the n-grams derived from each of the strings. The term n-gram denotes a continuing sequence of n characters. Using padding tokens, (L + n − 1) subsequences can be extracted from a particular string, where L denotes the length of the actual string. Usually, sets of bigrams or trigrams are compared. There are several possible ways of deriving a nonnegative number that represents the distance (Erikson 1997). In our experiments, we used the following formula. In contrast to the other algorithms, it does not denote a distance but a similarity measure for the two strings x and y, where B_x denotes the set of bigrams derived from string x and B_y those derived from string y, respectively:

$$sim(x, y) = 2\frac{|B_x| \cap |B_y|}{|B_x| + |B_y|}$$

Zobel and Dart (1996) presented the Editex algorithm as a new phonetic matching technique. This algorithm combines the properties of string edit distances with letter-grouping strategies used in well known phonetic indexing algorithms like Soundex (Knuth 1973) or Phonix (Gatt 1990). By doing so, they achieved superior results for tasks of phonetic matching. Ristad and Yianilos (1998) suggest a stochastic interpretation of string distances. They model them according to the probability of individual operations needed to transform one string into the other. These operations

[7] Both *aufwändig* and *aufwendig* are standard spellings in modern German.

are equivalent to the character replacements, insertions and deletions used to define the string edit distance. Additionally, the probability of identity operations (such as <a> to <a>) is taken into account.

Distance measures such as stochastic distance are commonly used in dialectrometry to calculate the distance or similarity between different dialect variants (Heeringa et al. 2006: 51). That is especially so because distance measures are fuzzy by definition. Most standard information retrieval systems build up an index of occurring terms, allowing the user to quickly find all documents containing the words he queried for. As mentioned above, an exact search may not yield good results for historical texts. An adequate distance measure operating on spelling variants provides arbitrary degrees of search fuzziness within a reasonable retrieval time. Standard fuzzy search, though, is of limited use as it does not take linguistic features into account. For example, if the user queries for the German term *urteil* 'judgment', the Levenshtein algorithm does not differentiate between the existing variant *urtheil* and, for instance, **ubrteil* with respect to the string distance. A measure that takes heed of linguistic connections will be able to determine the actual variant from a list of candidates.

We developed a framework for arbitrary distance measures, i.e. all concepts that define a distance between two objects. The measure we normally use in the FlexMetric framework is a measure that was derived from stochastic distance by scaling the probability distribution to a cost table. It combines the simplicity of a dynamic programming algorithm with the flexibility of defining arbitrary costs for each possible character transformation. The basic idea is very similar to the concept behind the string edit distance. The only difference is that, rather than the number of transformations, the costs for the individual operations are taken into account. The costs for the least expensive sequence of operations required to transform the one string into the other define the distance between the two strings. The cheapest sequence can be calculated using a dynamic programming algorithm resembling the one used for evaluating the string edit distance.

Distance measures can be used in other stages of a query as well and, therefore, in more than one module of the engine:

– *Ranking of Boolean results*. Retrieval in historical text documents is possible starting from a given query term, using automatically or

manually constructed rules that generate spelling variants. The variants produced are used for Boolean retrieval, returning unclassified results. Afterwards, a distance measure is required to rank the results according to their distance from the term queried.

– *Transformation*. Historical spelling variants can be automatically transformed into their modern counterparts. The distance measure is used to identify the correct spelling in a modern dictionary.
– *Reflection*. The differences between a historical or regional spelling variant and its modern equivalent are often hard to evaluate, even for native speakers. An adequate distance measure is a means of mapping linguistic distinctions on a single number. The visualization of word distances supports the reflection that language is in a state of constant change.

6.1 Training of distance measures

As mentioned above, we implemented a stochastic distance measure for trainability. In the course of three months, we collected nearly 13,000 string pairs of spelling variants and their standard spellings. Within those pairs is hidden the extent to which spelling variants differ from spellings in modern orthography. All single letter replacements in our database can be modeled by $\theta = 39 \times 39$ operations with replacement costs (German alphabet, umlauts, ß and some historical combined diacritical marks). To train a distance measure, we use our database as a sample set $X = x_1, \ldots, x_n$ and maximize the estimator $\hat{\theta}$ until we find an optimal set of operations to model the sample: that is, we calculate the maximum likelihood function

$$L(\hat{\theta}) = \max_{\theta} f(x_1, \ldots, x_n | \theta)$$

Of course, even 13,000 samples contain not nearly enough information to represent all the forms of variation that might occur. For this reason, we postulate a set of missing data, Y, which – added to the known sample – creates the complete data set $X \cup Y$. Furthermore, we can assume a joint relationship between X and Y (Bilmes 1998). The so-called expectation-maximization algorithm (Dempster 1977) alternates between the estimation of Y given constant X and θ^i and the maximization of θ^{i+1} given constant Y and θ^i. After numerous iterations, the algorithm reaches a (local) maximum and an optimal set of letter replacement operations.

The amount of support such distance measures can provide depends on their practicability in the particular context of historical spelling variants. Given not only trained measures but the abundance of different metrics and edit distances available, a thorough evaluation is needed.

7. Evaluation of distance measures

The main problem in judging the quality of string distance measures lies in comparing their applicability for different tasks. It is obvious that a distance measure that has been specifically trained to detect certain linguistic deviations can no longer yield objective results when used to quantify a relation between spellings as it necessarily evaluates the familiar deviation with lower costs, leading to a shorter distance. Thus, if, for instance, the measure is used to build up a genealogical tree of spelling variants of the same term, it inherently prefers relations it was specifically trained for. This effect leads to unusable results. In order to avoid this conflict, we have to concentrate on evaluating the potential of the various algorithms for the following text retrieval task: the user queries for the modern spelling, and all documents containing the query term or a historical variant are returned as results. Hence, a synthetic information retrieval system (IRS) has to be constructed consisting of a document collection, a retrieval function, and a set of queries along with relevance judgments.

The structure of the data itself can also significantly influence the outcome of an evaluation. One important factor is word length. If the dataset consists of many small words, the average distance will increase, because even a single letter replacement changes a high percentage of the word's recognizability. Also, if a distance measure is sensitive to word length, differences in length between the standard and the variant spelling can yield diverse results. In the 17th and 18th centuries, for example, extensive use was made of derivational suffixes. Whereas nowadays the adjective *streng* 'strict' is used, in 1650 Hans Michael Moscherosch wrote *zu geben strängiglichen gebotten (zu geben streng geboten* 'strictly commanded to give'). Figure 11, based on our collection of historical evidences, clearly shows the increased word length of the spelling variants in those centuries. Normalization by length appears to be a solution to differences in word length, but, as Heeringa et al. (2006) show, it only perverts the measures. Normalization optimizes for minimum normalized

length of the replacement path rather than minimum replacement costs (Heeringa et al. 2006: 54).

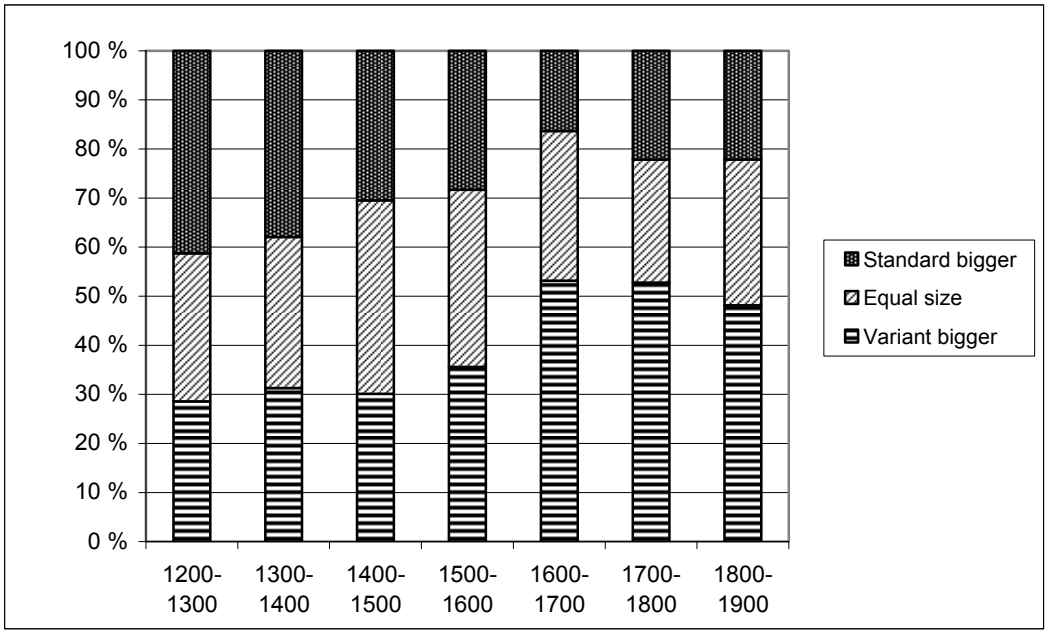

Figure 10. Comparison of the word lengths of standard spellings and spelling variants from 1200 to 1900.

The standard information retrieval methods for measuring performance are *precision* (proportion of retrieved and relevant documents to all documents retrieved) and recall (proportion of retrieved and relevant documents to all relevant documents). In our case, it is certain that a relevant counterpart exists for every query; that is, for every historical spelling there is a matching standard spelling. Also, using distance measures, every entry in the database is retrieved, and its distance to the query calculated. Therefore, retrieved and relevant documents are equal and so are precision and recall. As a result, we use precision at *n* (*P@n*). This measure is often used in cases where, instead of Boolean retrieval, a ranking of documents is returned, for example, in Web retrieval. Precision at 10 means that relevant documents are retrieved within the ten documents with the highest ranking.

An evaluation required us to strike a balance on what we hoped to achieve. We could either build a strictly controlled setup with a few hundred items or a much larger setup with less control. The advantage of the explicit results of the first version is greatly reduced by their narrow area of application. Since we are dealing with natural language data and

unknown types of variation, we suspect that too small an evaluation will yield results with limited value to practical applications.

To build a collection of 3,156 searchable terms and spelling variants, we used our evidence database and a manually maintained dictionary of 217,000 contemporary German words derived from the free spelling-correction tool Excalibur. The historical word forms found by the Information Retrieval System (IRS) are added to the dictionary, whereas the corresponding modern terms are removed. In this way, we try to raise the probability that no other relevant documents (that is, spelling variants) are collected. With an annotated corpus there is no problem at all, but without such a thoroughly tagged collection or manual inspection (of more than half a billion results!), it is impossible to be completely sure about the relevance of its entries. Looking back at the example of Kepler's text given above, we can see the spelling variant *Sterne* related to the first person singular standard spelling *Stern* 'star'. Unfortunately, *Sterne* is also the first person plural standard spelling 'stars' of the same word paradigm. Therefore, even if a distance measure is functioning perfectly and attests very low costs to the insertion of <e> (*Stern* → *Sterne*), the string identity (*Sterne* → *Sterne*) will always be cheaper, because the collection has no information about the word's grammatical number. As a result, the outcome of our evaluation heavily depends on the size and structure of the collection. Rather than the total numbers themselves, it is their relation that is of interest. Using a dictionary of 217,000 words is a balance between the 80,000-word OpenOffice dictionary and a combined dictionary of more than five million words we could also have used.

Table 5. Results of a comparison of distance measures.

Measure	P@1	P@2	P@3	P@4	P@5
Bigram evaluation	24.5 %	35.6 %	42.6 %	48.2 %	54.4 %
Editex	43.3 %	55.2 %	63.4 %	69.2 %	72.6 %
Levenshtein	22.9 %	36.6 %	47.1 %	53.4 %	58.9 %
Scaled stochastic measure	38.6 %	58.2 %	65.7 %	70.8 %	75.0 %
Stochastic measure	46.7 %	65.3 %	74.7 %	79.6 %	83.1 %

The results of the evaluation (cf. Kempken et al. 2006) show that the Levenshtein distance and the n-gram algorithm yield comparable results. This was to be expected as both of them evaluate a deviation regardless of its context or the affected characters. The Editex algorithm, the stochastic measure and its logarithmically scaled version deliver superior results. While Editex takes into account linguistic aspects due to its letter-grouping

strategy, the stochastic measures are trained on real linguistic data. This is definitely an advantage when dealing with historical data or recognition errors, where letter-groups can change. If one recalls the example at the beginning of Section 6 (*Jngenieur* vs. *Ingenieur* 'engineer'), for an 18th century document, the graphemes <i> and <j> should both belong to the same letter group; however, in Editex <i> belongs to group 1 and <j> to group 6 (Zobel and Dart 1996). The results of the stochastic measure are better than those of the scaled version, even though both rely on the same algorithm.

```
Ährenkranz
Ältestenrat
Ämter
Ämterverteilung
Änderns
Änderung
Änderungsantrag
Änderungsgesetz
Änderungsindex
```

Figure 11. Measures using dynamic programming can use previously calculated prefixes (underlined) to increase processing speed.

The main difference lies in their conceptual complexity; the scaled stochastic measure uses a cost measure that was derived from the stochastic measure. Whereas the stochastic distance measure needs an evaluation of the probability distribution for each term pair, the scaled version uses a derived cost measure in a simple dynamic programming algorithm. Hence, it allows intuitive optimizations like re-using previously calculated values (cf. Figure 11) for 1:*n* comparisons, which alone increases processing speed by more than 50 percent. For single queries such an enhancement is of minor importance, but increased speed allows for calculations that were previously out of reach. The evaluation described in Section 9 requires more than 9 billion word-by-word comparisons and still takes about half an hour. Furthermore, the derived cost measure is more likely to be understood and optimized by a human user for such purposes as linguistic analysis. Since it uses a table of replacement costs, the user can simply lower or raise costs for selected operations, while, in a probability distribution, any change influences all other values because the probabilities have to add up to 1.

We can draw the following conclusions:

- The better adapted an algorithm is to specific phenomena in the domain of historical spellings, the better the retrieval results that can be expected from it.
- The paramount results of a trained distance measure can be transferred to a simpler evaluation algorithm with a ~12 percent loss in quality but more than 50 percent of gain in speed.

8. Improvement of the stochastic measure using clustered training data

As we have seen, spelling variation increases with the age of the text. But the more inhomogeneous the training data becomes, the harder it is to train reliable measures with it. The characteristics of a certain period (such as the Barocke Letternhäufelung mentioned above) are diluted by the variation of others. However, clustering the evidences using the document's metadata allows more homogeneous training sets to be built. Yet the question remains: What is the size of an optimal training set? Too small a set might not reflect enough features, whereas too large a set can subdue the details. Our tests suggested training sets of about 4,500 evidences.

We defined two classes, *timeframe* and *location*, to deduce a semantic clustering. Their subcategories are based on commonly accepted stages and regions. As we learned through personal communication during a recent seminar on digital historical corpora, the DDTA project, an initiative of numerous renowned German language experts, proposed similar categories. *Timeframe* depicts four significant stages in the development of the German language:

- Late Middle High German (1250–1350)
- Older Early New High German (1350–1450)
- Later Early New High German (1450–1650)
- New High German (1650–1900)

Location is divided according to the region:

- Upper German (south of the Speyer line),

- Central German (south of the Benrath line but north of the Speyer line) and
- Low German (north of the Benrath line)

At the same time, *category* indicates OCR/Non-OCR errors.

Since, at the moment, we do not have enough evidences to fill all 12 clusters with 4,500 training entries, we have to reduce the clusters to the most significant ones. But the information of timeframe and location is immanent in all evidences and cannot be "extracted" separately. We examined the influence of the parameters time and location on the variability of spellings, or – to be more precise – the influence of time in contrast to all other parameters (except OCR and transcription). The 54 text documents used to create these data were selected randomly given the limited choice of available texts. They include chronicles, judicial documents, fiction, cookbooks and newspaper articles.

- We manually examined 54 historical documents containing 74,781 words, including 13,135 variant tokens. Due to the length of some documents, we had to use excerpts.
- Every occurrence of a spelling variant (cf. definition in Section 2.2, no OCR errors) was counted as a variant token.
- Proper nouns and non German segments (esp. Latin) were removed prior to calculation.

Table 6. The manually collected list of variant token amounts in historical German text documents.

Document	Year	# Words	# Var. tokens	Words : tokens
Bayrischer Landfrieden	1293	1182	573	48%
Mainauer Naturlehre	1300	871	568	65%
Das Buch von guter Speise (Auszug)	1350	841	514	61%
Wilhelm Durandus: Rationale	1384	1296	526	41%
Johannes von Tepl - Der Ackermann	1401	886	535	60%
Meister Ingold - Das püchlein vom guldin spiel	1432	1006	462	46%
Die Auslegung vber den pater noster	1441	992	583	59%
Das Helmaspergersche Notariatsinstrument	1455	1526	598	39%
PillenreuthMystik	1463	1428	679	48%

Übergabe der Stadt an die Schweizer	1499	659	238	36%
König Maximilian an die Bünde	1499	665	312	47%
Heinrich Hug - über den Schwabenkrieg	1499	464	205	44%
Tübinger Vertrag	1514	534	205	38%
Rede Bischof Friedrich Nausea	1527	163	71	44%
Gründungsurkundes des Hospitals Hofheim	1535	357	126	35%
Ach liebe fromme Kuhmäuler	1548	175	77	44%
Reichskammergerichtsordnung	1555	1306	400	31%
Sigismund von Herberstein - Moscovia, Hauptstadt der Reissen	1557	879	406	46%
Anekdote der Zimmerischen Chronik	1560	178	52	29%
Landgraf Philipp an seine Getreuen	1560	129	51	40%
Chronik des Grafen von Zimmern	1564	551	175	32%
Der Krieg in der Geschlechterchronik Eisenberger	1568	691	232	34%
Beauftragung des Superintendenten Johannes Angelus	1578	197	65	33%
Mängelrügen des Johannes Angelus	1579	230	96	42%
Vom Hasen Wildpret	1581	1201	455	38%
Gründtlicher Bericht von einem vngewohnlichen newen Stern	1604	1539	451	29%
Kleine Salzburgische Chronik	1624	342	114	33%
Berliner Zeitung 1626	1626	701	204	29%
Hans Michael Moscherosch Gesichte	1650	1415	353	25%
Christoph Schorer Chronik Memmingen	1660	2438	724	30%
Leibniz: Societät und Wirtschaft	1671	1081	232	21%
Christian Thomasius: 3 . Monat oder Martius	1688	1019	141	14%
Lehrzeugnis eines Apothekergehilfen	1691	209	63	30%
Brief von Landgraf Ernst Ludwig	1715	384	61	16%
Briefwechsel zwischen Landgraf Ernst	1715	254	55	22%
Beschluss des Landtags vom 16. Mai 1722	1722	485	101	21%
Berlinische Privilegirte Zeitung	1748	1884	211	11%
Neuer Lehrbegriff der Bewegung und Ruhe	1758	3935	122	3%
Berlinische Privilegirte Zeitung	1761	2039	186	9%
Karschin - Brief an Michaelis	1763	693	96	14%
Reglement der Berliner Kunstakademie	1776	804	75	9%
Zum ewigen Frieden	1795	4297	182	4%
Kaiserliche Ratifikation des Reichsgutachtens	1803	989	55	6%
Reichsdeputationshautpschluss	1803	3807	315	8%
Bedingungen, unter welchen die in der Rheinbundsakte angewiesenen	1806	675	48	7%

Besitzungen				
Vertrag zwischen dem Bevollmächtigten Sr. Majestät des Kaisers der Franzosen	1806	3482	161	5%
Hessenverfassung	1820	1260	88	7%
Sachsenverfassung	1831	1147	81	7%
Der Luftschiffer Blanchard	1850	502	10	2%
Die Aehnlichkeit der Locomotive mit einem Thiere	1858	634	30	5%
Philosopie und Erfahrung - Eine Antrittsrede	1861	3799	171	5%
Welt als Vorstellung	1870	4336	174	4%
Die Grenzen der sinnlichen Wahrnehmung	1876	7822	272	3%
Ueber den Einfluss des Gefühls auf die Thätigkeit der Phantasie.	1900	4402	155	4%
subtotal		74,781	13,135	

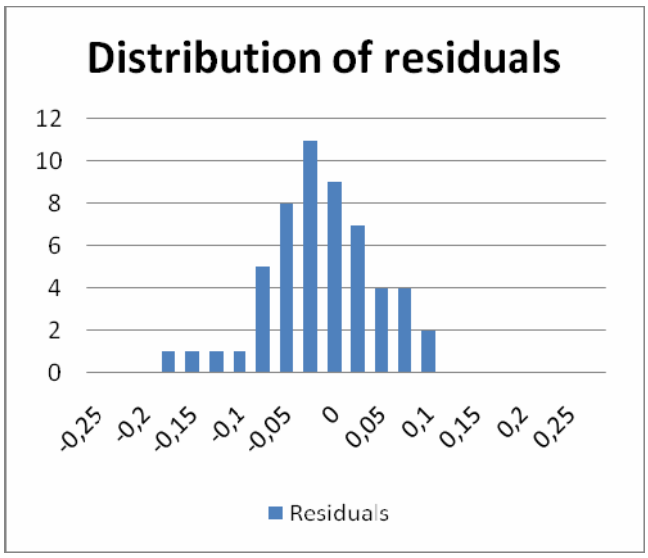

Figure 12. The residuals are normally distributed.

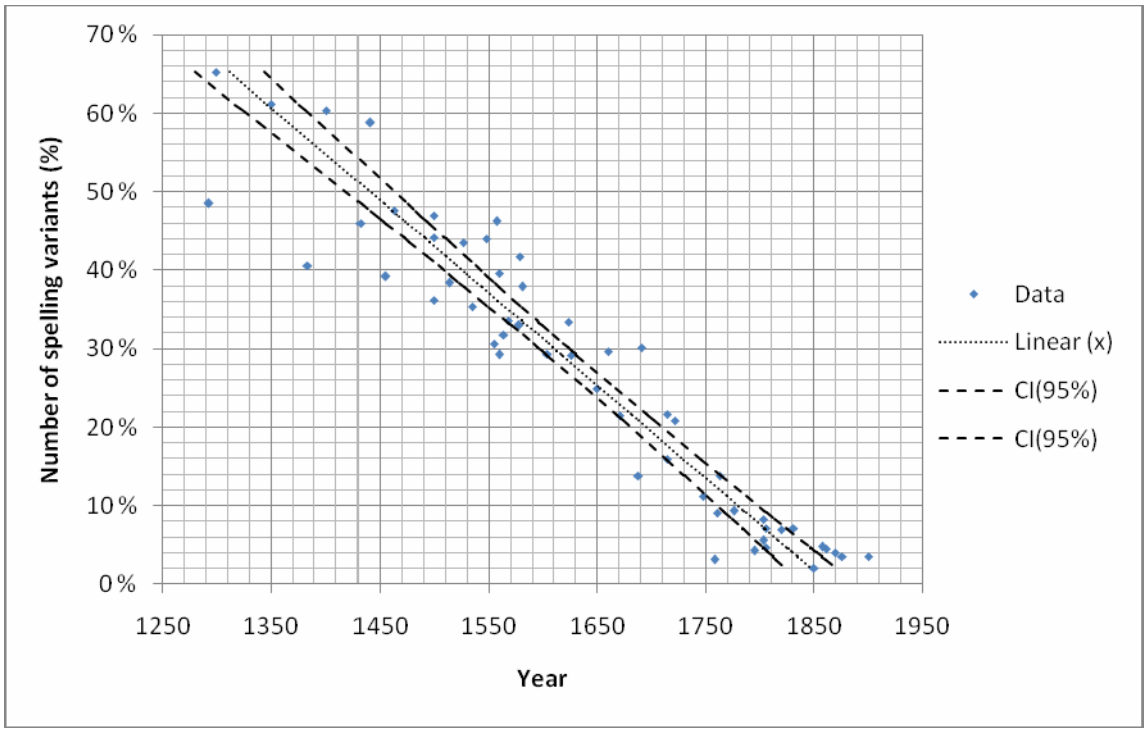

Figure 13. Number of spelling variant tokens found in 54 historical German texts between 1293 and 1900.

Given this variety, the homogeneity of the result is astounding. As can be seen in Table 6 and Figure 13, the number of spelling variant tokens (every occurrence of identical variants is counted separately) increases steadily from 2–4 percent in 1850–1900 to 65 percent in 1300. The figure already suggests a (negative) correlation between date of origin and number of spelling variants in a document. Calculating the Pearson correlation coefficient with $X = (x_1, x_2, \dots, x_n)$ being the dates of origin, $Y = (y_1, y_2, \dots, y_n)$ the percentage of spelling variants and $(\overline{x}, \overline{y})$ the centroid of the data, we get a very strong decreasing linear relationship of $r \approx -0.95145$.

Since the distribution of the residuals is normal (cf. Figure 12), it is feasible to suppose a linear data regression $\hat{y}_i = \alpha + \beta x_i$. Figure 13 shows the calculated y-regression model with $\alpha = 2.0182$ and $\beta = -0.00106$. Regarding the coefficient of determination, it is possible to explain 90.52 percent of the sample's variance, while the F-test with $q_{dist} = 9{,}5539 > q_F(1,\ 50, 95\%) = 4.03$ yields a relation between sample and model of greater than 95 percent significance. Minimizing to X, that is, the dates of origin, instead of Y, we can calculate $\hat{x}_i$ accordingly. The

regression $\hat{x}_i = -847.02 + 1864.38\, x_i$ allows for the prediction of a document's date where the number of variant tokens is known. Its standard error of estimate $\hat{\sigma}_{xy} = 49.85$ accounts for ~50 years of error between the data and our estimation. The upper and lower bounds of the 95 percent confidence interval for $\hat{x}$ are calculated by

$$\hat{x} \pm t_{(1-\alpha/2)} \cdot 49.85 \cdot \left(\frac{1}{54} \cdot \frac{(x_i - 1635)}{54 \cdot 0.0319}\right)$$

and range from ± 32.14 years in 1300 to ± 13.71 years in 1626 and ± 23.50 years in 1850. To compare these findings to synchronous variation, we need a definition of temporal equality since we do not have enough documents from identical years. If we define a difference in the temporal origin of documents of less than one generation (that is, 25 years) as equality, it is possible to calculate empirical variance $\tilde{s}^2$ as well as standard deviation $\tilde{s}$ for the occurring groups (cf. Table 7 and Figure 14). Using the given data and requiring a minimal group size of four items, we get seventeen groups of equal documents with four to eight members. The maximal standard deviation of 6.966 percent in the 16th century is noticeable but still surprisingly low. By the 19th century, synchronic factors ($\tilde{s} < 1.6$ %) become negligible.

Our findings suggest that time indeed has a bigger influence on variation than synchronic factors. Therefore, metrics trained on diachronically clustered data should be superior to synchronic metrics.

Table 7. Empirical variance and standard deviation of synchronic document groups.

Groups of document		$\tilde{s}^2$	$\tilde{s}$
Years	#	%	%
1499-1514	4	25.055	5.006
1535-1560	6	48.528	6.966
1548-1568	7	46.313	6.805
1555-1579	8	36.621	6.051
1557-1581	8	32.679	5.717
1560-1581	7	20.654	4.545
1560-1581	6	16.301	4.037
1564-1581	5	17.135	4.139
1568-1581	4	16.734	4.091
1748-1763	4	20.934	4.575
1758-1776	4	19.463	4.411
1795-1820	6	2.494	1.579

1803-1820	5	2.039	1.428
1803-1820	4	2.343	1.531
1806-1831	4	1.148	1.122
1850-1870	4	1.558	1.248
1858-1876	4	0.310	0.557

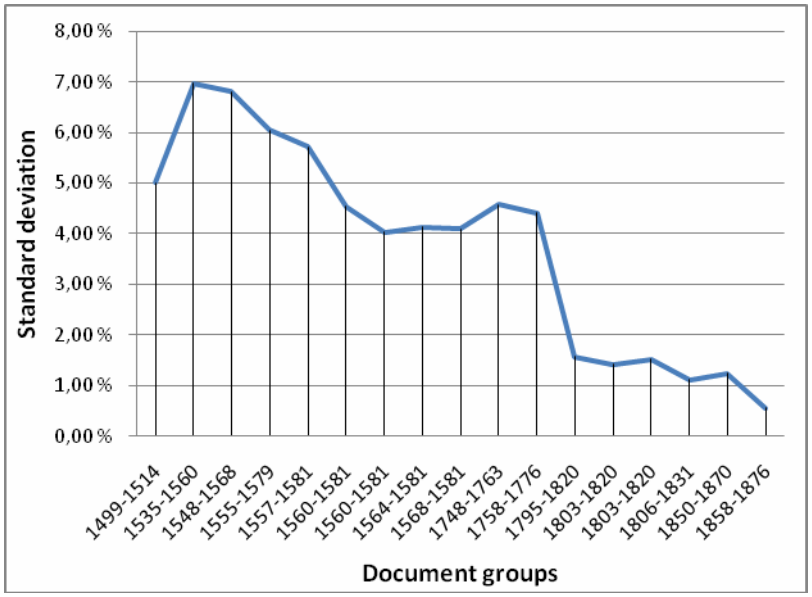

Figure 14. Standard deviation of synchronic document groups.

It is interesting to notice that the diachronic standard deviation corresponds to general linguistic expectations. With the exception of the first group of synchronic documents, $\tilde{s}$ is (strictly) monotonically decreasing until 1564. Only in 1795 does it again reach the level that the prior slope suggested. This period (1564–1795) correlates noticeably with the Baroque, from ~1575–1770, a period well-known for its extensive spelling variation (‘Barocke Letternhäufelung’, cf. 2.2).

9. Visualization as a means to ease evaluation

Clustering data and using the clusters for the training of stochastic distance measures produces many different measures. Their relevance to the required retrieval tasks has to be evaluated separately. To speed up and ease the evaluation process, we propose options for visual support. The prototype we have developed is but one example of these options and is

meant to encourage scientists to benefit from visual information representation.

While planning the prototype, we also kept Shneiderman's paradigm in mind: "Overview first, zoom and filter details on demand" (Shneiderman 1996). We employed multidimensional scaling (MDS) to display abstract distance in 2D space (see below). Interactivity is gained with the ability to select and remove spellings from the calculations, lower or raise cutoff frequencies and filters and even change replacement costs with instantaneous effect (see below). This led to a user interface separated into three main views:

- The *Histogram* allows an overview of thousands of data items. The selection of a certain portion of data triggers MDS and table views (cf. Figure 13).
- *Multidimensional Scaling* (MDS) functions as a detail view. Such visualization is used to display sets of several dozen to a few hundred items (cf. Figure 14).
- The *Table View* can display different levels of detail (cf. Figure 15).
- *Treemaps* (cf. Section 6) are another way to display details of single word derivations as an add-on for table views. We have not yet embedded them in our prototype for metric evaluation.

To acquire a first impression of how a spelling distance performs on a set of evidences, we calculate the distance between a spelling variant and the entries in a dictionary as described above, with one difference:

$$p@n' = p@n - \sum_{i=1}^{n-1}(p@i)$$

The histogram provides a good representation of the overall performance of a spelling distance given for a set of test data. If a large number of spellings are found in the acceptable ranking range, if there are noticeable isolated outliers or if the values are spread widely over the whole interval, the user will quickly notice. In addition, histograms can be useful as tools for comparing different spelling distances. Usually, multiple histograms are viewed one after another or arranged next to each other. While this might be enough to perceive considerable differences in distributions, small-scale variations may pass unnoticed. An easy solution to this problem is to

arrange the different histograms in a combined display area where the relevant subinterval bars are lined up next to one another and made distinguishable by color or texture.

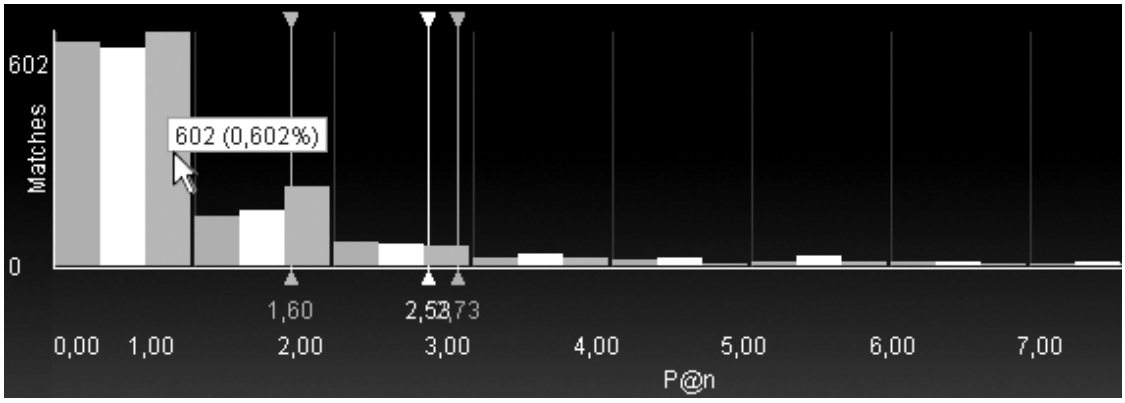

Figure 15. Histogram comparison of three different distance measures.

The *MDS view* displays smaller subsets, thus allowing further refinement while providing additional information detail. MDS is a class of statistical methods that has its roots in psychological research. The main application of such techniques is to assign the elements of an item set to a spatial configuration in such a way that it represents the elements' relationships with as little distortion as possible. In this context, MDS can be used to arrange spellings in a two-dimensional space according to their spelling distances from one another. Every available dimension reduces the need for distortion but increases the difficulty of interpretation. Two or three dimensions are a good trade-off. This allows for an intuitive display of distances and clusters of spelling variants. It also makes it possible to discover distance anomalies. If this representation is provided with filtering features, it can be used to select subsets of elements quickly and comfortably. These subsets can then be displayed in detailed information views that would be too cluttered with greater numbers of items.

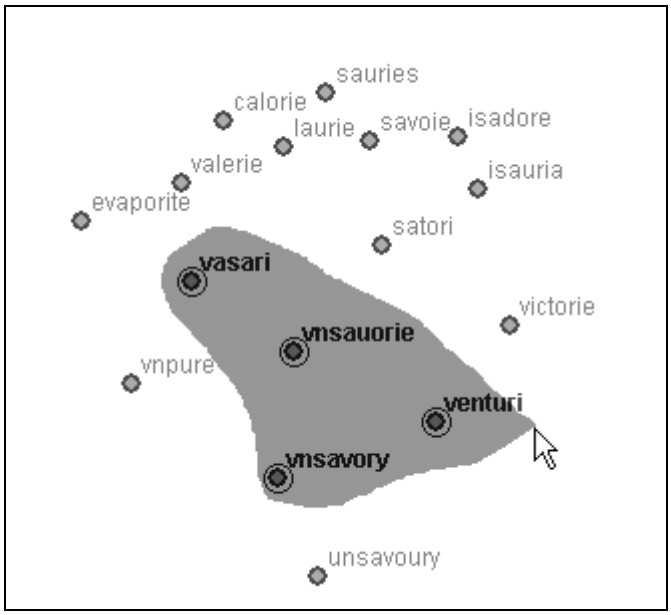

Figure 16. MDS view of the spelling variant *vnsauorie* and its standard spelling, *unsavoury*.

The task of this view is not to reconstruct the calculated distance perfectly but to uncover characteristics of the spelling distances and spelling sets used. These characteristics, such as clusters and outliers, usually outweigh the distortion that results from the conflict between the exact calculated distances between the items and their 2D spatial arrangement. This visualization approach is applicable to a wide variety of spelling distances as long as they provide a quantitative measurement of two spellings. There are no assumptions made about the distance value except that small values represent a high degree of similarity.

Tabular views display detailed results or interactively modify the replacement costs of the distance measure in use.

	kundt>kind	kundt>kund	kundt>kunde	kundt>kunz
Distance sum	1.572	0.572	0.903	1.326
del(t) : 0.572	0.572	0.572	0.572	0.572
ins(e) : 0.331			0.331	
repl(d, z) : 0.754				0.754
repl(u, i) : 1.0	1.0			

Figure 17. Table view of replacement costs mirroring deletion, insertion and replacement costs. These costs can be manually adjusted to trigger an MDS view update.

In Pilz et al. (2007a), we describe a cross-language comparison of English and German spelling variation. We noticed that – presumably because of their kinship – distance measures trained on German data can be successfully used to search historical English databases. Since a stochastic distance measure represents the variation of the data it was trained on, it can be employed to measure the degree of correlation between two data sets. The German distance measure that delivers the best results on English data should yield the most similar data to the English text. Since a manual comparison of multiple measures varying in number of training data and their origin (in our case, the time period) can cost a lot of time and work, it is an ideal setting for use of the Metric Evaluation Tool. We determined that German training data from the 13th to the 15th century is best suited to represent spelling variation in Shakespearean English. For more and more thorough examples, please see Pilz et al. (2007b).

10. Conclusion

In this paper we described the challenges one faces when digitizing printed text material. We especially examined the problems caused by OCR, transcription and the spelling variation involved with historical documents. The RSNSR project has been researching this topic for two and a half years now. Working with an archive for the reception of Friedrich Nietzsche as well as with fellow researchers from Great Britain and the Netherlands, we have developed a Java framework for fuzzy full-text retrieval on nonstandard texts based on letter replacement rules as well as string edit distances. Its two main purposes are to grant professional researchers and interested amateurs easier access to the store of knowledge residing in historical text documents and to support the deployment of such texts by means of computer science. Our particular goals are an – as far as possible

– automatic process chain of evidence collection, training of stochastic distance measures and successful retrieval. The framework was applied to two search engines and also used in various prototypes of information visualization interfaces for retrieval, browsing and detailed examination of historical data. A by-product of our research was the Metric Evaluation Tool, another example of how information visualization can significantly ease the daily work of a researcher. We are therefore proposing increased usage of automation and visualization in linguistic research.

References

ABBYY FineReader XIX Brochure. http://www.frakturschrift.de/PDF/Leaflet_FRXIX_D_lo.pdf (online, v. 16.12.08).

Ammon, Ulrich; Bickel, Hans & Ebner, Hans (2004) *Variantenwörterbuch des Deutschen. Die Standardsprache in Österreich, der Schweiz und Deutschland sowie in Liechtenstein, Luxemburg, Ostbelgien und Südtirol*. Berlin: de Gruyter.

Benedetto, Dario; Caglioti, Emanuele & Loreto, Vittorio (2003) Zipping Out Relevant Information. *Computing in Science and Engineering* 5(1):80–85.

Bibliotheca Augustana. http://www.fh-augsburg.de/~harsch/augustana.html (online, v. 16.12.08).

Bilmes, Jeff A. (1998) *A Gentle Tutorial of the EM Algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models.* Berkeley, California: U.C. Berkeley, International Computer Science Institute.

Bick, Andre (2002) *Die Realisierung der Nietzsche-CD mit Unterstützung einer MySQL-Datenbank: Dokumentation & Konzept*. Duisburg: Gerhard-Mercator-Universität Duisburg.

Biella, Daniel; Dyllong, Eva; Luther, Wolfram & Pilz, Thomas (2005) An On-line Literature Research System with Rule-Based Search. *Proceedings of the 4th European Conference on e-Learning (ECEL2005)*: 67–76.

Bruls, Mark; Huizing, Kees & van Wijk, Jarke J. (2000) Squarified treemaps. *Proceedings of the joint Eurographics and IEEE TCVG Symposium on Visualization*: 33–42.

Canoo. Free online German language resources. Canoo Engineering AG, www.canoo.net (online, v. 16.12.08).

Card, Stuart K., Mackinlay, Jock & Shneiderman, Ben (eds.) (1999) *Readings in Information Visualization: Using Vision to think*. Los Altos: Morgan Kaufman.

Christmann, Ruth & Schares, Thomas (2003) Towards the User: The Digital Edition of the Deutsche Wörterbuch by Jacob and Wilhelm Grimm. *Literary and Linguistic Computing* 18: 11–22.

Compact Memory. http://www.compact-memory.de (online, v. 16.12.08).

Deutscher Wortschatz. Universität Leipzig. http://wortschatz.uni-leipzig.de (online, v. 16.12.08).

Dempster, A.; Laird, N. & Rubin, D. (1977) Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society* 39(1): 1–38.

Digitales Archiv Hessen-Darmstadt http://www.digada.de/ (online, v. 16.12.08).

documentArchiv.de. Dokumenten- und Quellensammlung zur deutschen Geschichte ab 1800. http://www.documentarchiv.de (online, v. 16.12.08).

Dudenverlag (2004) *Duden. Die deutsche Rechtschreibung*. Band 1. Mannheim: F.A. Brockhaus.

Gadd, T.N. (1990) PHONIX: The Algorithm. *Automated Library and Information Systems* 24(4): 363–366.

Elmentaler, Michael (2003) *Struktur und Wandel vormoderner Schreibsprachen*. Berlin: de Gruyter.

Erikson, Klas (1997) Approximate Swedish Name Matching Survey and Test of Different Algorithms. *Nada report TRITA-NA-E9721*.

Fekete, Jean Daniel & Plaisant, Catherine (2002) Interactive information visualization of a million items. *Proceedings of IEEE Symposium on Information Visualization 2002 (InfoVis 2002)*.

Fleischer, Wolfgang (1966) *Strukturelle Untersuchungen zur Geschichte des Neuhochdeutschen,* Berlin: Akademie-Verlag.

Heeringa, Wilbert; Kleiweg, Peter; Gooskens, Charlotte & Nerbonne, John (2006) Evaluation of String Distance Algorithms for Dialectology. In John Nerbonne & Erhard Hinrichs (eds.), *Linguistic Distances Workshop at the joint conference of ICCL & ACL,* pp. 51–62. Sydney: Association for Computational Linguistics.

Holmes, David I. (1998) The evolution of stylometry in humanities computing. *Literary and Linguistic Computing* 13(3): 111–117.

Johnson, Brian & Shneiderman, Ben (1991) Tree-maps: A spacefilling approach to the visualization of hierarchical information structures. *Proceedings of IEEE Visualization Conference*: 284–291.

Kempken, Sebastian; Luther, Wolfram & Pilz, Thomas (2006) Comparison of distance measures for historical spelling variants. In *Artificial Intelligence in Theory and Practice* IFIP Series 217, pp. 295–304. Boston: Springer.

Knuth, Donald (1973) *The Art of Computer Programming, Vol. 3: Searching and Sorting*. Addison-Wesley.

Koolen, Marijn; Adriaans, Frans; Kamps, Jaap & de Rijke, Maarten (2006) A cross-language approach to historic document retrieval. In *Advances in Information Retrieval: 28th European Conference on Information Retrieval (ECIR 2006)*, pp. 407–419. Heidelberg: Springer.

Levenshtein, Vladimir (1966) Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady* 10 (8): 707–710.

Maas, Utz (2000) *Orthographie – Materialien zu einem erklärenden Handbuch zur Rechtschreibung des Deutschen.* Osnabrück.

Mischke, Lothar & Luther, Wolfram (2005) Document Image De-Warping Based on Detection of Distorted Text Lines. In Fabio Roli & Sergio Vitulano (eds.) *Image Analysis and Processing ICIAP 2005*, pp. 1068–1075. LNCS 3617. Cagliari, Italy: Springer.

Mischke, Lothar (2007) *Teilautomatisierte Verschlagwortung von in altdeutschen Schriftfonts gesetzten Texten mit Hilfe lernender Verfahren.* PhD Thesis. Universität Duisburg-Essen. http://www.scg.inf.uni-due.de/Abschlussarbeiten/DissMischke.php.

Munske, Horst-Haider (1999) *Orthographie als Sprachkultur*. Frankfurt a. M.: Lang.

Nerbonne, John & Siedle, Christine (2005) Dialektklassifikation auf der Grundlage aggregierter Ausspracheunterschiede. *Zeitschrift für Dialektologie und Linguistik* 72(2): 129–147.

O'Rourke, Alan J.; Robertson, Alexander M.; Willett, Peter; Eley, Penny & Simons, Penny (1997) Word variant identification in Old French. *Information Research* 2 (4), http://informationr.net/ir/2-4/paper22.html (online, v. 16.12.08).

Rayson, Paul; Archer, Dawn & Smith, Nicholas (2005) VARD versus Word: A comparison of the UCREL variant detector and modern spell checkers on English historical corpora. *Proceedings from the Corpus Linguistics Conference Series on-line e-journal* 1(1).

Ristad, Eric Sven & Yianilos, Peter N. (1998) Learning String Edit Distance. *IEEE Transactions on Pattern Recognition and Machine Intelligence* 20 (5): 522–532.

Shneiderman, Ben (2006) *Treemaps for space-constrained visualization of hierarchies* http://www.cs.umd.edu/hcil/treemap-history/ (online, v. 16.12.08).

Pilz, Thomas (2003) *Unscharfe Suche in Textdatenbanken mit nichtstandardisierter Rechtschreibung am Beispiel von Frakturtexten zur Nietzsche-Rezeption. Thesis. (Civil service examination).* University of Duisburg-Essen.

Pilz, Thomas; Ernst-Gerlach, Andrea; Kempken, Sebastian; Rayson, Paul & Archer, Dawn (2007a) The identification of spelling variants in English and German historical texts: manual or automatic? *Literary and Linguistic Computing* 23(1):65–72.

Pilz, Thomas; Philipsenburg, Axel & Luther, Wolfram (2007b) Visualizing the evaluation of distance measures. *Proceedings SigMorPhon 2007*: 84–92.

Pollock, Joseph J. & Zamora, Antonio (1983) Collection and Characterization of spelling errors in scientific and scholarly texts. *J. American Society for Information Science* 34 (1): 51–58.

Vandenbussche, Wim (2002) Dutch orthography in lower, middle and upper class documents in 19th-century Flanders In Andrew R. Linn & Nicola McLelland (eds.) *Standardization Studies from the Germanic languages*, pp. 27–42. Amsterdam/ New York: John Benjamins.

Wagner, Robert A. & Fischer, Michael J. (1974) The String-to-String Correction Problem. *Journal of the ACM* 21 (1): 168–173.

Wedershoven, Urs (2007) *Konzeption und Erstellung eines Systems zur automatischen Belegerstellung aus korrespondierenden historischen und modernen Schreibungen.* Diploma thesis. University of Duisburg-Essen.

Zobel, Justin & Dart, Philip (1996) Phonetic String Matching: Lessons from Information Retrieval. *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*: 166–172.

Contact information:

Thomas Pilz and Wolfram Luther
Institute of Computer Science and Interactive Systems
University of Duisburg-Essen
Lotharstr. 65
D-47048 Duisburg
Germany

e-mail: pilz(at)inf(dot)uni-due(dot)de and luther(at)inf(dot)uni-due(dot)de

Ulrich Ammon
Institute of German Language and Literature Studies
Lotharstr. 65
University of Duisburg-Essen
D-47048 Duisburg
Germany

e-mail: ulrich.ammon(at)uni-due(dot)de