

# Information Access to Historical Documents from the Early New High German Period

Andreas Hauser<sup>1</sup>, Markus Heller<sup>1</sup>, Elisabeth Leiss<sup>2</sup>, Klaus U. Schulz<sup>1</sup> and Christiane Wanzeck<sup>2</sup>

<sup>1</sup> CIS, University of Munich, Oettingenstr 67, 80538 München, Germany

E-mail: andy@splashground.de, heller@cis.uni-muenchen.de, schulz@cis.uni-muenchen.de

<sup>2</sup> Institut für Deutsche Philologie, University of Munich, Schellingstr 3/RG, 80799 München, Germany

E-mail: e.leiss.@germanistik.uni-munechen.de, ch.wanzeck@germanistik.uni-munechen.de

## Abstract

With the new interest in historical documents insight grew that electronic access to these texts causes many specific problems. In the first part of the paper we survey the present role of digital historical documents. After collecting central facts and observations on historical language change we comment on the difficulties that result for retrieval and data mining on historical texts. In the second part of the paper we report on our own work in the area with a focus on special matching strategies that help to relate modern language keywords with old variants. The basis of our studies is a collection of documents from the Early New High German period. These texts come with a very rich spectrum on word variants and spelling variations.

*Keywords:* historical documents, information access, Early New High German, historical language, information retrieval, word similarity, approximate matching.

## 1 Introduction

Until today, a huge part of the world-wide cultural heritage is hidden in historical books and documents. For various reasons, the problem of how to make this information accessible and public has recently gained much attention. An immense number of historical books and text repositories are threatened with physical ruin. In order to preserve these documents for future generations they have to be digitized. The digitization in symbolic form opens the door for using modern techniques of information access such as Information Retrieval (IR), text mining, hyperlinking, flexible rendering and presentation of documents. In the humanities, new forms of E-science and collaborative scientific work are simplified by enabling shared access to distributed and heterogeneous document resources. While these possibilities mainly improve the working conditions of historians, paleologists, linguists,

and philosophers, the contents of many historical books are also interesting for non-experts. The idea to make the contents of historical books publicly accessible gains more and more popularity. A number of projects and initiatives recently followed these lines. Examples are Open Content Alliance, Google Print, Gutenberg project, Early English Books Online, European digital library project.

Unfortunately, a serious problem is immediately found when trying to access historical documents in symbolic digital form. For most periods, language does not have normalized spelling. And even today many languages still do not. The large amount of spelling variants of the same word makes it impossible to directly use standard indexing techniques for IR and text mining. Only recent papers [Ernst-Gerlach and Fuhr, 2006; Pilz *et al.*, 2006; Archer *et al.*, 2006] have started to analyze this problem seriously. The following questions and research issues represent the kernel of a new research area.

1. Which kind of historical mutations and variants can be observed in the orthography of distinct languages? How can we describe these variations in a formal way?
2. What are the consequences for distinct fields/techniques such as IR and text mining?
3. How can existing techniques be adapted to better cope with historical texts?

The aim of this paper is twofold. As to items 1 and 2 we give a coarse survey with a focus on German language. As to 3, we present some of our own work in the field, concentrating on IR and matching strategies that help to establish correspondences between modern and old spelling variants.

The paper is structured as follows. Section 2 briefly outlines the current role of digitized historical documents. Section 3 collects the most important facts and observations on historical language change, focussing on German language and spelling. Section 4 discusses the resulting difficulties for various forms of information access, focussing on informa-

tion retrieval, and sketches solution variants. Section 5 gives a brief survey on related work, projects and resources that have been developed to overcome these problems. Sections 6 outlines our program, Section 7 describes recent own work on matching and approximate search in historical documents. The Conclusion sums up and comments on further relevant work in our group.

## 2 Digital historical documents

Historical texts come into existence as documents in the public executive, judicial and church administration, in companies, as erudite and aesthetic literature, but also as private notes. They are created intentionally in order to inform future readers, but also occasionally while addressing contemporary recipients.

In this form, they are deposited on a regulated basis at a filing department, or they arrive after some meandering over time at some historical archive, where they may be retro-digitized. The lowest grade of digitization is a representation as image, scanned from the original. While this gives first access to the document, the access is rather limited. The next grade is a textual representation from transcription by hand usually using a Unicode representation like UTF-8. The third grade is a structured and possibly annotated version of the textual representation, often in XML, e.g. TEI<sup>1</sup>.

The digitized form is necessarily an intentional historical source: In order to provide for the informational demand of future users, their presumed requirements are to be considered. There will be mainly three possible approaches:

- The linguist will be interested in the document's language and may wish to generate analyses as common in corpus linguistics, such as concordances, statistical distribution schemes etc.
- The paleographer will be interested in annotated information about the external and non-textual properties of a historical source.
- The historian will be mainly interested to work on the sources' contents, either in the original or in the edited and annotated form, such as information on the historical context. The general public interest in historical sources may fall into this category, too, even though the queries may not be grounded on a methodical approach.

The user groups' query requirements demand varying effort in the annotation process: Queries into the original text may be answered straight away, but queries on the contextual, as well as towards the paleographic information require

<sup>1</sup><http://www.tei-c.org/>

the manual recording and classification thereof. This effort also has a strong influence on the number of available documents. Sizeable digitization and digital library projects, such as the Digital Library Foundation<sup>2</sup> generally employ a more shallow and more automatic degree of annotation compared to very specialized libraries such as the collection efforts of the Charters Encoding Initiative<sup>3</sup>.

Direct access to historical texts however may be hindered through language change: Traditional information retrieval techniques rely on the identity of the search term and the occurrence in the corpus, which is less the case, the older the texts are. In order to reach good recall values, access therefore must cope with phonetic, derivational and semantic variation.

## 3 Historical language change

As German texts are the basis of the project, the present abstract firstly presents a short introduction to the chronological structure of the German language. In total, there are four important time stages:

*Old High German (OHG)* (8th century until approx. 1100) is the oldest German language of which evidence is given. The linguistic material consists of names and individual words contained in documents and narrative texts. There are approximately 70 literary texts from this time stage (e.g. the *Hildebrandslied*). The prose texts are mainly translations partially closely following the Latin text.

*Middle High German (MHG)* (1100-1350) is a linguistic period during which German as written language is gaining increasing importance. Whereas clergymen were responsible for German as written language as far as Old High German is concerned. They were joined during the Middle High German era by noble laymen. The poetical language can be found, for example, in the *Nibelungenlied* and in the works of WALTER VON DER VOGELWEIDE. The said texts reflect the first intentions to standardize the German language.

*Early New High German (ENHG)* (1350-1600) is the era in which German as a written language comes into being to an increasing extent. Written German was no longer limited to particular types of text, but was extended to almost any kind of text thanks to the invention of printing. The main characteristic of the texts written in Early New High German is the large dialectal variance. Martin Luther's translation of the Bible is decisive for the evolution of the language, and a large part of linguistic evolutions was based on this translation.

*New High German (NHG)* (since 1600) is an epoch subdivided into three periods: The first period (until the end of the

<sup>2</sup><http://www.diglib.org/about/dfcharter.htm>

<sup>3</sup><http://www.cei.lmu.de/>

18th century) includes literary figures such as J. W. Goethe and F. Schiller, the second period (19<sup>th</sup>-20<sup>th</sup> century) includes the scientists Jacob and Wilhelm Grimm and contemporaneous German. The number of linguistic variants at the end of the second period significantly decreased due to the introduction of Konrad Duden's reform of orthography. A distinction has to be made between High German and Low German, which has also to be considered as German, but which did not undergo sound shift. The preliminary stage of today's Dutch belongs to the same speech area as Low German.

The linguistic variations are related to the following language levels:

1. *Phonological/graphical*. The rate of graphical variants is very high in historical texts. The said variants are often based on dialectal and stylistic elements. As a result, the following variants can be found for one letter vowel graphemes:

Grapheme	Variants
<a>	< á, â, ah, aa, ai, ae, â >
<e>	< eh, ee, ei, ey, â, ê, ä >
<i>	< j, y, ÿ, ie, iee, i <sup>e</sup> , ij, ye, ih, jh, ieh, yh >
<o>	< oh, ô, oe, oi, oy, oo >
<u>	< ú, û, û, v, w, uh, wh, ûh, uy >
<ä>	< â, e, a, æ, ae, äh >
<ü>	< û, u, û, v, û, ÿ, y, w, ue, üe, üh, uy >
<ö>	< ô, ó, o, ôh, oe, ôe, ôe, œ >

2. *Morphological*. Inflectional morphology is characterized by a high degree of variability. The old case endings show complex patterns which are subsequently levelled to an increasing extent. The formation of the plural is developed in the course of the language evolution, resulting in doublets such as Germ. *Licht-e* vs. *Licht-er* for Engl. *light*. As far as verbal inflection is concerned, up to 4 inflectional possibilities are existing simultaneously for the plural number during the Early New High German period: 1. *-(e)nt* (1.-3. pers.); 2. *-(e)n* (1. pers.), *-(e)nt* (2.-3. pers.); 3. *-(e)n* (1./3.pers.), *-(e)t* (2. pers.); 4. *-(e)n* (1.-3. pers.). Within the field of word formation, the limit between base and suffix is temporarily shifted, e.g. MHG *truic-heit* is replaced by ENHG *trui-cheit* for Engl. *sadness*. Compounds such as *Rechts Sachen* (Engl. *legal matters*) which cannot always be clearly distinguished from a syntactic unit or mere word succession represent a major problem. [Wegera and Solms, 2000; Wegera and Prell, 2000]

3. *Lexical*. The lexical changes of words are of great relevance as it may occur that not only the current meanings of a word appear in a text, but also the old meanings. The meanings of words differ according to the different periods of time, to take an example, *urlaub* (Engl. *vacation*) means in OHG and MHD "permission" and subsequently shifts to ENHG "farewell" and then in NHG gains the meaning of "leisure time for recovery". Due to the increasing use of for-

eign words, there is a large number of doublets in ENHG, e.g. Lat. *amant* vs. Germ. *Liebhaber* (Engl. *lover*). Differentiation of terms has the consequence e.g. in the area of legal terms. A large number of older terms such as ENHG *dingtag*, *tagsatzung* meaning 'appointment' continue to exist simultaneously for some time. The result is an ENHG vocabulary which is extraordinarily comprehensive as compared to other historical eras, thanks to the great variety of variants [Wolf, 2000].

4. *Syntactical*. In particular the development of the first regularities in the development of syntactical structures is very interesting. The phrases in historical texts may reach a considerable extension due to the newly gained possibilities of extension (e.g. the NP *von Gott (dem jr feind seid) gaben* 'from God (towards whom you have a hostile attitude) gifts').

Attributive phrases are partially placed in front of the nominal head and partially behind it (e.g. *des vergossen bluts Christi fur unser sunde* 'the blood Christ shed for our sin'). As far as syntax is concerned, the variations of the position primarily go back to pragmatic reasons and to a low extent only to dialectal reasons. The syntactical principle of punctuation comes into being in the era of Early New High German only, and the syntactical structure existing before that epoch therefore is not very pronounced [Erben, 2000]. The validity of the statements about historical changes of language depends on the representative character and the size of the text corpus. Digital text corpora enable us to evaluate the processes of language change on the basis of a comprehensive data base and to collect all relevant linguistic variations.

#### 4 Resulting problems and solution alternatives

For the applications mentioned in Section 2, the variability of historical language represents a serious problem. Hyperlinking of concepts and frequency based data mining methods are distorted. Obvious difficulties arise for corpus linguistic techniques such as annotation, concordancing, and n-gram clustering. Standardized indexing techniques in IR fail to produce satisfactory results since distinct occurrences of the same word come with various orthographic variants.

For sketching solution alternatives we concentrate on IR on historical texts. We look at the following simplified problem: given an input word of modern language, how can we compute all those hits in a document repository that represent an (old or modern) variant of the same word? Three solution alternatives are the following:

1. *Special dictionaries*. The dictionary stores with each modern word entry a list of observed historical variants. Further information (time, place, source) may be added.

Each input word of the user is (interactively or automatically) replaced by the corresponding variants stored in the dictionary.

An advantage is that stored correspondences are manually checked. No assumptions on word similarity are needed. However, the creation of suitable dictionaries is time consuming, and with a static dictionary the coverage of historical spelling variants reached in arbitrary texts will remain modest. In order to improve recall, techniques for approximate matching are a natural choice. Two classes can be distinguished.

2. *Rule-based generative matching.* The differences between new and corresponding historical spelling variants are described by a set of rules. In the online-variant, rules are applied to a given input word, thus generating possible old variants for search. In the offline-variant, we try to normalize historical variants at indexing time by applying inverse rules.
3. *Matching based on word similarity.* The correspondence between new and old variants is modelled by a special form of word similarity. Given an input word  $W$ , all words (types) of the collection are presented that are sufficiently similar to  $W$ .

Since distinct historical spelling variants of the same word often have a similar pronunciation we may also try to compute a kind of phonetic normal form for all words and then use a special similarity measure on normalized words. Rule-based approaches may be used for phonetic normalization. In practice, all approaches can be combined.

## 5 Related work and resources

Search in digitized images of historical documents is described in e.g. [Rath *et al.*, 2004] and [Gatos *et al.*, 2005]. Images are generated from search terms and compared to images of the documents. Recall could probably be improved with techniques similar to the ones described in 7.

*Information Retrieval on historical text collections.* [Ernst-Gerlach and Fuhr, 2006] describe an approach where probabilistic rules are applied to search terms in order to generate possible historical spelling variants. Rule sets are produced in a two-step procedure. In the first step, the tokens of the historical text collection are matched against a dictionary of contemporary words. Tokens in the dictionary are excluded. The remaining tokens are manually inspected. To each proper historical word the present-day spelling variant is assigned. The list of all pairs of the form (historical word; modern spelling) represents the input for the second step. An algorithm produces a list of transformation rules that

may be applied to arbitrary modern words and yield possible historical spellings. In terms of the classification presented in Section 4, [Ernst-Gerlach and Fuhr, 2006] represents a rule-based matching approach (Type 2).

In [Pilz *et al.*, 2006] the authors describe the project “Rule-based search in text databases with nonstandard orthography (RSNSR)”. A rule-based fuzzy search engine is introduced that allows users to retrieve text data independently of its orthographical realization. Rules are derived manually using expert knowledge and statistically through a machine learning approach using n-grams. In addition, a weighted Levenshtein algorithm was employed, the weights for which were computed with the Ristad-Yanilos [Ristad and Yanilos, 1997] algorithm. Special rules for OCR errors may be added on demand. The project has a focus on the German reception of Nietzsche, thus addressing the period 1865-1945.

Being a collaborative effort together with the German project [Pilz *et al.*, 2006], the authors of [Archer *et al.*, 2006] first present some more details of the above. In a second part, they portray the VARD (‘variant detector’) tool developed by Archer and Rayson, which has been designed to automatically normalize variants and thus aims to determine the correct modern equivalent - in contrast to the German project, which intends to find and highlight the historical spellings.

Both approaches use a manually crafted set of letter replacement heuristics. VARD also uses a manually collected list of spelling variants and a small set of contextual lexical rules in order to find spelling variants, such as ‘than’ in contrast to ‘then’.

The approaches are similar and the authors hope to develop general procedures for Indo-European languages.

*Information Retrieval on text collections for languages without fixed orthography.* [Strunk, 2003] considers IR and matching techniques for Low Saxon texts. The Levenshtein distance is refined to a special “Low Saxon distance”, introducing classes of substrings that are “equivalent” from a graphemic or phonetic point of view. Edit operations (insertion, deletion, substitution) receive costs 1, 0.5, or 0.25, depending on the classes used. In the tests, the Low Saxon measure behaves slightly better than standard Levenshtein.

*Matching variants, approximate name matching.* In [Zobel and Dart, 1995] the authors consider distinct methods for selecting approximate matches for input tokens in large lexicons. A toolbox for measuring the similarity of names using various distance measures is introduced in [Schnell and Bachteler, 2004]. A comparison of matching techniques for historical variants of words can be found in [Rayson *et al.*, 2005]. Efficient methods for selecting approximate matches in large dictionaries based on the Levenshtein distance are

described in [Schulz and Mihov, 2002; Mihov and Schulz, 2004].

*Workshops, conferences.* Recently, the problem of how to access historical documents was discussed in special workshops and conference sessions. A “Workshop on Historical Text Mining”<sup>4</sup> was organized by Paul Rayson and Dawn Archer in July 2006 at Lancaster University, UK. At ICDAR<sup>5</sup> 2005, two special sessions addressed the analysis of historical documents. In December 2006, a meeting on the same subject is organized in Dagstuhl (Germany) by Norbert Fuhr.

*Dictionaries for historical language.* Electronically available are: DWB<sup>6</sup>, four of Middle High German<sup>7</sup>, the *Goethe Wörterbuch*, *Deutsche Rechtswörterbuch* and four dialectal<sup>8</sup>. Links to these can be found on the homepage *Das Wörterbuchnetz*<sup>9</sup>. All of them are very comprehensive with examples of the use of the words and standardized lemmata and therefore can only be used to a limited extent for automatic text identification.

*Electronic corpora for other languages* There are similar problems as far as the digitisation of historical texts of other European languages is concerned. At present, there is the *Helsinki-Corpus*<sup>10</sup> for the English language and *Fran-text*<sup>11</sup> for French. These projects are facing the same difficulty, which is the preparation of historical texts for digital, complex research enquiries.

## 6 Own work: general goals

The focus of our interest is set on the acquisition of the early prints (14<sup>th</sup>-17<sup>th</sup> century). In contrast to manuscripts they are numerous, so that a comprehensive data base can be gained. The diversity of the text types, which comes into being with the beginning of printing only. Chronicles, sermons, documents and legal texts are available in the German language from that time on only. Previously the said texts were prepared in Latin only. And, last but not least, the historical prints offer - unlike manuscripts - the possibility of automatic digitization.

Automatic preparation of digital historical texts is a great challenge due to the original documents being worn out and in Gothic print of various styles, types vary from one printing press to other and change over time. Some of the characters are only available since the introduction of Unicode, and

<sup>4</sup><http://ucrel.lancs.ac.uk/events/htm06/>

<sup>5</sup>Internat. Conference on Document Analysis and Recognition

<sup>6</sup>*Deutsches Wörterbuch von J. und W. Grimm*

<sup>7</sup>G. F. Benecke/W. Müller/F. Zarncke; Lexer

<sup>8</sup>Alsation, Palatine, Rhineland, Lotharingian

<sup>9</sup><http://www.woerterbuchnetz.de/>

<sup>10</sup><http://www.ling.upenn.edu/hist-corpora/>

<sup>11</sup><http://www.lib.uchicago.edu/efts/ARTFL/databases/TLF/>

some only as Combing Character which today's software still has problems with.

Thus even OCR software designed to handle “standard” 19<sup>th</sup> century Gothic print exhibit poor results, same for search tools. The lack of usable electronic dictionaries and the high variation of the same words deepen the problems.

We thus determined the following steps in our research:

- Manually create a small corpus
- Handle spelling and compound variations
- Create an usable electronic dictionary
- Incorporate morphology and syntax
- Incorporate document structure and meta-information
- Use all this to improve OCR and digitize more texts

This leads to an iterative process, as the above points are interwoven and improving one will directly help the others.

*Corpus.* From a selection of 23 texts from the Early New High German time, eleven have been digitized. Four<sup>12</sup> of these have been tagged to include information about category, New High German translation, underlying Early New High German lemma, corresponding New High German lemma. The 11 texts represent a total of about 18,000 lines and 130,000 words (tokens).

## 7 Own work: matching of concepts and IR

*Classifying matching problems.* As a starting point of our own work we manually collected correspondences between old and new variants of words in the text *Dyll Vlnspiegel*, thus creating a small dictionary of the form described in Section 4. As by-product we collected a list of phenomena that explain distinct types of correspondences between modern and historical word forms. We found variation rates of about 50% in relation to contemporary German on the token level, which are much higher than in the corpora from more recent ages focused in [Ernst-Gerlach and Fuhr, 2006; Pilz *et al.*, 2006], which only dealt with up to 15%. We identified the following problem classes:

1. *New word form.* The input word corresponds to an old word of completely distinct form. Today, the old word form is no longer in use. Example: *handeln*  $\mapsto$  *marcken* (‘to trade’). Here and in what follows triples  $x \mapsto y(z)$  denote a modern word  $x$ , and old equivalent  $y$  and the English translation  $z$ .

<sup>12</sup>Alexander Weissenhorn, *Dyll Vlnspiegel*, Augsburg 1540; Gervasius Stürmer, *Eyn sehr hoch nö<sup>e</sup>tige Erinnerung*, Erfurt 1548; Johann Scharfenberg, *Christliche Bekaentnis*, Breslau 1586; Pamphilus Gengenbach, *Dz lob der pfarrer*, Basel 1521

2. *Latin words.* Even in non-scientific and non-religious texts, Latin words were often used to demonstrate education. Spelling of Latin words was not normalized. E.g. *appellacion, appellacionn, appellation, appellatioun* ('appellation').
3. *Variations in word splitting.* Compounds that are now written as a single word were often separated into two words *Winters zeiten*  $\mapsto$  *Winterzeit* ('wintertime').
4. *Partial new word form.* In the old variant, a morpheme or subword is replaced. Examples are *Mönchswesen*  $\mapsto$  *Moencherey* ('monasticism'), *Großteil*  $\mapsto$  *Mehrteil* ('bigger part'), *hinauslaufen*  $\mapsto$  *außlaufen* ('to amount to'), *feindselig*  $\mapsto$  *feindlistig* ('hostile').
5. *Variation of prefixes/suffixes.* A given prefix/suffix is found to be often replaced by another prefix/suffix in a more or less systematic way. Example: *-chen*  $\mapsto$  *-lein* as in *Kindchen*  $\mapsto$  *Kindlein* ('little child').
6. *Typesetting variations.* For example, when running out of printing letters *i*, ancient typesetters used letters *j* instead.
7. *Graphemic-phonetic variations* Example: *Abertheur*  $\mapsto$  *Abenteuer* ('adventure')
8. *New character* that is not used in modern language. Example: *für*  $\mapsto$  *für* ('for')

*An optimal matching strategy.* Given the phenomena described above, we are currently designing a matching strategy that optimizes precision and recall, combining all components described in Section 4. The dictionary component is meant to cover all variations of type 1, 2, 3, as well as all irregular patterns of the form 4. In addition, associations of either type that have been manually checked are stored in the dictionary. In this way, the dictionary offers a solid basis for evaluating the reliability of specific matching strategies. When storing an association of type 1, 2, 3, or 4 in the dictionary, the historical variants may be garbled in a second step using matching rules (s.b.). In this sense, these variants define a *first layer* of expansion for a given input term, characterized by very high precision. To each variant of the first level we may apply further expansion steps as described below.

Graphemic-phonetic variations 7, regular variations of type 3, 4, 8 prefix/suffix variations 5 and typesetting variations 6 are conveniently described using a special set of matching rules. In earlier experiments with rule-based generation of candidates for orthographic errors [Ringlsetter *et al.*, 2006] we found that applying rules eagerly tends to produce an immense number of useless variants. Hence we intend to use only "safe" rules that produce possible historical spellings with sufficient confidence. The selection of rules

should be dependent on the vocabulary of the actual document basis.<sup>13</sup> "Safe" rules are applied to a given input word and its first-level expansion (s.a.), producing a set of possible historical variants that represent a *second layer* of expansion. Even if the candidates of the second layer are not manually validated, the emphasis is rather on precision than on recall.

In order to improve recall, a fine-grained special word similarity measure is used to produce an additional set of historical variants of a given input term. Details are given below. For a given input term *W* we compute all words occurring in the document basis that are sufficiently similar to *W*. In this way we obtain a *third layer* of expansions for the input term.

In a final step, candidates of all three layers are then ranked, using word similarity, frequency information and a suitable heuristics for giving high (small) additional preference to candidates of layers 1 (2). In what follows we describe the practical work that has been done for realizing the three-layer architecture.

*Dictionary construction and linguistic workbench.* The dictionary construction process is embedded in a larger working context where historical texts are linguistically analyzed and annotated. To support and facilitate dictionary construction, text analysis and annotation, a linguistic workbench with underlying SQL database has been realized. Central features are the following:

- Concordancing tool visualizing occurrences of words in their contexts.
- Search for spelling variants in a preliminary way, based on regular expressions, a simple form of fuzzy search and soundex.
- Enhanced support for searching compounds which use constraints on part-of-speech when linguistic annotation is available.
- Statistics are provided about documents, number of matches and occurrences

*Design of matching rules and special word distance.* After initial attempts with standard Levenshtein distance it became obvious that refinements are preferable where weights both depend on the kind of operation (insertion, deletion, substitution) and on the particular symbols to be acted on. The use of these fine-grained distances then leads to a natural interplay between rule-based matching and (training of) edit weights: each natural transformation rule can be used to reduce the costs of the corresponding edit operation. Conversely, if we find via training edit operations with low costs, good candidates for the rule set arise in a natural way. Looking at the

<sup>13</sup>We can learn from the methods for automated computing of rules sketched in [Ernst-Gerlach and Fuhr, 2006; Pilz *et al.*, 2006].

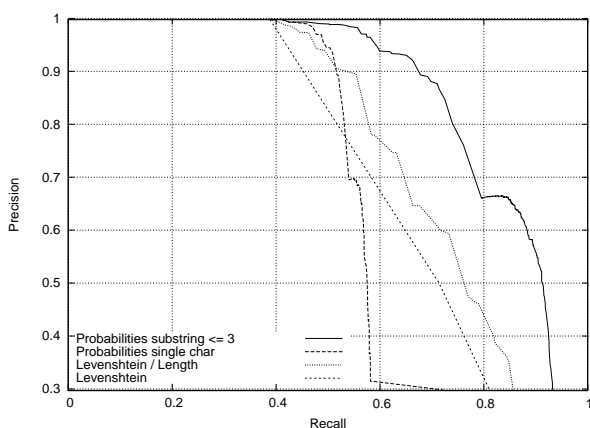


Figure 1: Fuzzy Matching

first side, we derived some rules from linguistic literature like [Stockmann-Hovekamp, 1991] and added further important rules that became obvious from the dictionary construction. On the word distance side we first used variants of the method for learning edit weights described in [Ristad and Yianilos, 1997]. However, a closer look at the variation patterns in the above list shows that standard edit operations, fail to capture the needed context. First many transformations (e.g.  $i \mapsto y$ ) are modelled much more naturally when adding relevant context ( $lein \mapsto leyn$ ). Second there are also substring to substring transformations ( $\delta h \mapsto oe$ ). Hence we implemented a version of the approach described in [Brill and Moore, 2000], where edit operations are based on sequences of symbols instead of single symbols.

Figure 1 depicts a recall-precision diagram obtained from preliminary evaluation. From 3600 lexemes pairs, old and modern spelling, not depending on the dictionary method, one half was used as training, the other as test data. In the learning phase the frequencies of the operations needed to transform the modern spelling to the old spelling are obtained and then converted to weights. Thus when like in our case the transformation  $i \mapsto j$  was frequent, the cost in the distance function was low. The training considered edit operations on single chars, [Ristad and Yianilos, 1997], and substrings  $\leq 3$ , [Brill and Moore, 2000]. No EM algorithm was used, unlike described in the mentioned papers. The original Levenshtein measure and a length dependent version is given as baseline. Sources <sup>14</sup> are available.

## 8 Conclusion

We surveyed the role of digital historical documents and the problems caused by historical language change for access-

<sup>14</sup>[http://www.splashground.de/andy/programs/weighted\\_distance/](http://www.splashground.de/andy/programs/weighted_distance/)

ing the documents with methods from IR and data mining. We then outlined our project centered about the digitization of texts from the early new high German period and presented preliminary results for approximately matching modern words against old vocabulary.

Looking forward we are confident that the applied strategy and techniques are also helpful to other applications where mappings are needed from one language to a related one with a high rate of variations, like Modern English to Early Modern English, where e.g. it is realized as *it* and *yt* <sup>15</sup>.

When digitizing historical texts, many problems arise that are not touched here. Further work in our group is centered around the following problems.

1. The conversion of historical documents via optical character recognition (OCR) often leads to poor results, due to special fonts and print styles such as Gothic print. Despite of many approaches to post-correcting OCR results ([Kukich, 1992; Taghva and Stofsky, 2001; Strohmaier *et al.*, 2003]), the special problems resulting from historical texts, language change and historical printing styles have been widely ignored so far.
2. Collections of electronic historical texts are often annotated using special XML dialects. This gives raise to the question of how to combine matching and approximate search with XML retrieval.

As to 1, a serious problem is the *recognition* of wrongly recognized tokens. We currently try to design a confidence measure for estimating the plausibility that an OCR token represents a correct recognition result.

As to 2, we are developing a cross-platform library framework to support fast structure and content XML query processing, while allowing for the previously defined degree of variance. The technology is conceptionally based on [Weigel *et al.*, 2004] in terms of XML indexing. Matching techniques go back to [Mihov and Schulz, 2004].

*Acknowledgements.* This work was supported by the German Research Foundation DFG and by VolkswagenStiftung.

## References

- [Archer *et al.*, 2006] Dawn Archer, Andrea Ernst-Gerlach, Sebastian Kempken, Thomas Pilz, and Paul Rayson. The identification of spelling variants in english and german historical texts: manual or automatic? In *Conference Abstracts: Digital Humanities 2006. Paris - Sorbonne, July 5th - 9th 2006.*, pages 3–5. Association for Digital Humanities Organisation, 2006.

<sup>15</sup>E.g. in Parke tr. Mendoza's, *The History of the Great and Mighty Kingdom of China and the Situation Thereof*, 1588

- [Brill and Moore, 2000] Eric Brill and Robert C. Moore. An improved error model for noisy channel spelling correction. In *ACL '00: Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, pages 286–293, Morristown, NJ, USA, 2000. Association for Computational Linguistics.
- [Erben, 2000] Johannes Erben. *Syntax des Frühneuhochdeutschen*, pages 1584–1593. Stefan Sonderegger, Berlin, New York, second edition, 2000.
- [Ernst-Gerlach and Fuhr, 2006] Andrea Ernst-Gerlach and Norbert Fuhr. Generating search term variants for text collections with historic spellings. In *28th European Conference on Information Retrieval Research (ECIR 2006)*, 2006.
- [Gatos *et al.*, 2005] Basilios Gatos, Thomas Konidakis, Kostas Ntzios, Ioannis Pratikakis, and Stavros J. Perantonis. A segmentation-free approach for keyword search in historical typewritten documents. In *ICDAR*, pages 54–58. IEEE Computer Society, 2005.
- [Kukich, 1992] Karen Kukich. Techniques for automatically correcting words in texts. *ACM Computing Surveys*, pages 377–439, 1992.
- [Mihov and Schulz, 2004] Stoyan Mihov and Klaus U. Schulz. Fast approximate search in large dictionaries. *Computational Linguistics*, 30(4):451–477, December 2004.
- [Pilz *et al.*, 2006] Thomas Pilz, Wolfram Luther, Norbert Fuhr, and Ulrich Ammon. Rule-based search in text databases with nonstandard orthography. *Literary and Linguistic Computing*, 21(2):179–186, 2006.
- [Rath *et al.*, 2004] Toni M. Rath, R. Manmatha, and Victor Lavrenko. A search engine for historical manuscript images. In Mark Sanderson, Kalervo Järvelin, James Allan, and Peter Bruza, editors, *SIGIR*, pages 369–376. ACM, 2004.
- [Rayson *et al.*, 2005] Paul Rayson, Dawn Archer, and Nick Smith. VARD versus Word. A comparison of the UCREL variant detector and modern spell checkers on English historic corpora. In *Proc. of the Corpus Linguistics 2005 Conference, conference series on-line e-journal, vol. 1*, Birmingham, UK, 2005.
- [Ringlstetter *et al.*, 2006] C. Ringlstetter, K. U. Schulz, and S. Mihov. Orthographic errors in web pages - towards cleaner web corpora. *Computational Linguistics*, 2006. to appear.
- [Ristad and Yianilos, 1997] Eric Sven Ristad and Peter N. Yianilos. Learning string edit distance. In *Proc. 14th International Conference on Machine Learning*, pages 287–295. Morgan Kaufmann, 1997.
- [Schnell and Bachteler, 2004] Rainer Schnell and Tobias Bachteler. A toolbox for record linkage. *Austrian Journal of Statistics*, 33(1&2):125–133, 2004.
- [Schulz and Mihov, 2002] Klaus U. Schulz and Stoyan Mihov. Fast String Correction with Levenshtein-Automata. *International Journal of Document Analysis and Recognition*, 5(1):67–85, 2002.
- [Stockmann-Hovekamp, 1991] Christina Stockmann-Hovekamp. *Untersuchungen zur Druckersprache in den Flugschriften Martin Bucers*. Carl Winter, Universitätsverlag, Heidelberg, 1991.
- [Strohmaier *et al.*, 2003] C. Strohmaier, C. Ringlstetter, K. U. Schulz, and S. Mihov. A visual and interactive tool for optimizing lexical postcorrection of OCR results. In *Proceedings of the IEEE Workshop on Document Image Analysis and Recognition, DIAR'03*, 2003.
- [Strunk, 2003] Jan Strunk. Information retrieval for languages that lack a fixed orthography. Technical report, Linguistics Department, Stanford University, 2003.
- [Taghva and Stofsky, 2001] Kazem Taghva and Eric Stofsky. OCRSpell: an interactive spelling correction system for OCR errors in text. *International Journal of Document Analysis and Recognition*, 3:125–137, 2001.
- [Wegera and Prell, 2000] Klaus-Peter Wegera and Hans-Peter Prell. *Wortbildung des Frühneuhochdeutschen*, pages 1594–1605. Stefan Sonderegger, Berlin, New York, second edition, 2000.
- [Wegera and Solms, 2000] Klaus-Peter Wegera and Hans-Joachim Solms. *Morphologie des Frühneuhochdeutschen*, pages 1532–1554. Stefan Sonderegger, Berlin, New York, second edition, 2000.
- [Weigel *et al.*, 2004] Felix Weigel, Holger Meuss, François Bry, and Klaus U. Schulz. Content-Aware DataGuides: Interleaving IR and DB Indexing Techniques for Efficient Retrieval of Textual XML Data. In *Proceedings of the 26th European Conference on Information Retrieval (ECIR)*, pages 378–393, 2004.
- [Wolf, 2000] Dieter Wolf. *Lexikologie und Lexikographie des Frühneuhochdeutschen*, pages 1554–1584. Stefan Sonderegger, Berlin, New York, second edition, 2000.
- [Zobel and Dart, 1995] Justin Zobel and Philip Dart. Finding approximate matches in large lexicons. *Software-Practice and Experience*, 25(3):331–345, 1995.