Automated support for evidence retrieval in documents with nonstandard orthography

Thomas Pilz and Wolfram Luther

The goal of the interdisciplinary project on rule-based search in text databases with nonstandard orthography is the development of a fuzzy search engine for orthographically unstandardized electronic documents. Proper results are the consequence of several intermediate stages of processing: the collection of data (evidences of nonstandard spelling and related standard spelling), training of specific search modules and, of course, the actual search task. This paper concentrates on one of the most important stages: the automated support of evidence collection. It describes the process of collection, the underlying methods, their value for linguistic research and how these methods can be implemented in an interface for automatic user support.

In the context of the preservation of cultural heritage, many historical documents have recently been digitized. No small number present difficulties because of nonstandard spellings—spelling variants and optical character recognition (OCR) errors. Both complicate the reading and successful search queries. Low quality facsimiles and scans or simply the use of historical font types significantly decreases the recognition rate of otherwise reliable software (Mischke 2007). Even if the resulting documents are extensively manually revised to remove OCR errors, spelling variants cannot simply be translated into standard spelling because of their linguistic value.

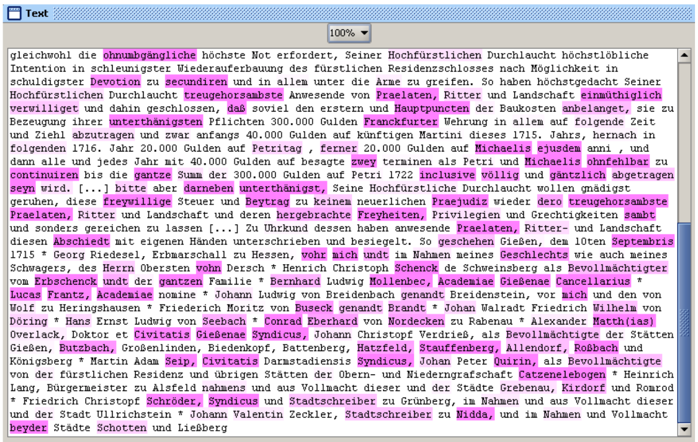
The common solution to the problem of unstandardized spellings, the use of large historical dictionaries, is costly. Instead, we use linguistic evidence transferred to formerly unknown spellings. We manually collected more than 12,800 word pairs of spelling variants or recognition errors and their related standard spellings. Stochastic training on such evidences allows for the development of reliable topic-related search engine modules. Since their quality heavily depends on the amount of available training data, we developed algorithms and interfaces for the automated support of our work.

The first prototype of such an interface was implemented in 2005 and supported the collection of about 9,000 evidences. It is being completely rebuilt and will be presented in the full paper.

When an unknown text is used for evidence retrieval, four main steps are involved:

1. Detection and categorization of nonstandard spellings and their separation from standard spellings
2. Deductions about the text's origin based on inherent information
3. Use of the information deduced to find related standard spellings for all spelling variants
4. Use of those evidences for the continuous enhancement of steps 1–3

We developed a series of filters to separate nonstandard from standard spellings, the simplest being comparison with a modern dictionary. Nonstandard spellings containing letters not included in Unicode Basic Latin and the German umlauts, are a good indication of variation as well. Separating those indicators into sets for historical diacritics and special characters already allows for the early categorization of spellings into historical variants and recognition errors. Special characters, like asterisks, slashes or circumflexes, account for 15 to 20 percent of OCR errors.



```
gleichwohl die ohnumbgängliche höchste Not erfordert, Seiner Hochfürstlichen Durchlaucht höchstlöbliche
Intention in schleunigster Wiederauferbauung des fürstlichen Residenzschlosses nach Möglichkeit in
schuldigster Devotion zu secundiren und in allem unter die Arme zu greifen. So haben höchstgedacht Seiner
Hochfürstlichen Durchlaucht treugehorsambste Anwesende von Praelaten, Ritter und Landschaft einmüthiglich
verwilliget und dahin geschlossen, daß soviel den erstern und Hauptpuncten der Baukosten anbelanget, sie zu
Bezeugung ihrer unterthänigsten Pflichten 300.000 Gulden Franckfurter Wehrung in allem auf folgende Zeit
und Ziehl abzutragen und zwar anfangs 40.000 Gulden auf künftigen Martini dieses 1715. Jahrs, hernach in
folgenden 1716. Jahr 20.000 Gulden auf Petritag , ferner 20.000 Gulden auf Michaelis ejusdem anni , und
dann alle und jedes Jahr mit 40.000 Gulden auf besagte zwey terminen als Petri und Michaelis ohnfehlbar zu
continuiren bis die gantze Summ der 300.000 Gulden auf Petri 1722 inclusive völlig und gäntzlich abgetragen
seyn wird. [...] bitte aber darneben unterthänigst, Seine Hochfürstliche Durchlaucht wollen gnädigst
geruhen, diese freywillige Steuer und Beytrag zu keinem neuerlichen Praejudiz wieder dero treugehorsambste
Praelaten, Ritter und Landschaft und deren hergebrachte Freyheiten, Privilegien und Grechtigkeiten sambt
und sonders gereichen zu lassen [...] Zu Uhrkund dessen haben anwesende Praelaten, Ritter- und Landschaft
diesen Abschiedt mit eigenen Händen unterschrieben und besiegelt. So geschehen Gießen, dem 10ten Septembris
1715 * Georg Riedesel, Erbmarschall zu Hessen, vohr mich undt im Nahmen meines Geschlechts wie auch meines
Schwagers, des Herrn Obersten vohn Dersch * Henrich Christoph Schenck de Schweinsberg als Bevollmächtigter
vom Erbschenck undt der gantzen Familie * Bernhard Ludwig Mollenbec, Academiae Gießenae Cancellarius *
Lucas Frantz, Academiae nomine * Johann Ludwig von Breidenbach genandt Breidenstein, vor mich und den von
Wolf zu Heringshausen * Friederich Moritz von Buseck genandt Brandt * Johan Walradt Friedrich Wilhelm von
Döring * Hans Ernst Ludwig von Seebach * Conrad Eberhard von Nordecken zu Rabenau * Alexander Matth(ias)
Overlack, Doktor et Civitatis Gießenae Syndicus, Johann Christopf Verdrieß, als Bevollmächtigte der Stätten
Gießen, Butzbach, Großenlinden, Biedenkopf, Battenberg, Hatzfeld, Stauffenberg, Allendorf, Roßbach und
Königsberg * Martin Adam Seip, Civitatis Darmstadiensis Syndicus, Johan Peter Quirin, als Bevollmächtigte
von der fürstlichen Residenz und übrigen Stätten der Obern- und Niederngrafschaft Catzenelebogen * Heinrich
Lang, Bürgermeister zu Alsfeld nahmens und aus Vollmacht dieser und der Städte Grebenau, Kirdorf und Romrod
* Friedrich Christopf Schröder, Syndicus und Stadtschreiber zu Grünberg, im Nahmen und aus Vollmacht dieser
und der Stadt Ullrichstein * Johann Valentin Zeckler, Stadtschreiber zu Nidda, und im Nahmen und Vollmacht
beyder Städte Schotten und Ließberg
```

Figure 1 Mapping the confidence value of the Bayesian classifier to the opacity of a word's background.

A trainable Naïve Bayes classifier was implemented using N-Grams. It is able to detect recognition errors as well as historical spellings by calculating their factor of variability while not necessarily categorizing every spelling. It provides a level of confidence in its classification. Mapping the confidence value to the opacity of a word's background color can assist users in the detection of spelling variants (see Figure 1). With the combination of different filters, we expect detection rates well beyond 80 percent.

No matter how elaborate a categorization algorithm becomes, some cases will always need human intervention: words with a standard translation but no standard spelling or real word errors.

The amount of historical variants in a text is valuable information for the deduction of a document's date of origin. Certain types of variation are often significant for specific eras or locations, like 'Letternhäufelung' in the baroque period. Methods such as authorship attribution and parsing the document for locations and dates can provide additional information. Inversely, knowledge about the origin of a text document is a means of deducing probable types of variation. Using this information, feature-based distance measures are able to determine the most probable standard spelling related to a spelling variant and to carry out such tasks as helping in the automatic construction of historical dictionaries. Pilz et al. (2007) showed an increased retrieval quality when using adequately adjusted measures. Rayson et al. (2005) have largely automatized the process of automatic translation annotation, but this has proven to be especially hard for heavily inflected languages. Implemented with dynamic programming, the interface is interactively able to present a list of candidate standard spellings whenever a nonstandard spelling is in focus. At least for German, the correct spelling will not necessarily appear in the top position, especially not with the correct inflection. But to combine spellings of identical inflection is of utmost importance since a stochastic distance measure reproduces every difference with no regard for semantics. If, for instance, diachronic variation is to be reproduced, it should not be mingled with inflection rules.

As described, the quality of the distance measures directly affects the quality of evidence collection. We hope that a continuous circle of collection and training with as little human intervention as possible will lead to search engines with even better retrieval results than today's.

References

Mischke, L.
2007 Teilautomatisierte Verschlagwortung von in altdeutschen Schriftfonts gesetzten Texten mit Hilfe lernender Verfahren. PhD thesis, University of Duisburg-Essen. Studien zur Mustererkennung, Vol. 24, Logos-Verlag, Berlin

Pilz, T., Philipsenburg, A. and Luther, W.
2007 Visualizing the Evaluation of Distance Measures SigMorPhon2007. Proceedings ACL 2007, Prague, 28 Jun 2007

Rayson, P., Archer, D. and Smith, N.
2005 VARD versus Word: A comparison of the UCREL variant detector and modern spell checkers on English historical corpora. Proceedings Corpus Linguistics 2005, Birmingham, 14-17 Jul 2005