## Abstract
## The identification of spelling variants in English and German historical texts: manual or automatic?

Dawn ARCHER (University of Central Lancashire)
Andrea ERNST-GERLACH, Sebastian KEMPKEN, Thomas PILZ (Universität Duisburg-Essen)
Paul RAYSON (Lancaster University)

## The identification of spelling variants in English and German historical texts: manual or automatic?

1. Introduction

In this paper, we describe the approaches taken by two teams of researchers to the identification of spelling variants. Each team is working on a different language (English and German) but both are using historical texts from much the same time period (17th – 19th century). The approaches differ in a number of other respects, for example we can draw a distinction between two types of context rules: in the German system, context rules operate at the level of individual letters and represent constraints on candidate letter replacements or n-graphs; in the English system, contextual rules operate at the level of words and provide clues to detect real-word spelling variants i.e. 'then' used instead of 'than'. However, we noticed an overlap between the types of issues that we need to address for both English and German and also a similarity between the letter replacement patterns found in the two languages.

The aims of the research described in this paper are to compare manual and automatic techniques for the development of letter replacement heuristics in German and English, to determine the overlap between the heuristics and depending on the extent of the overlap, to assess whether it is possible to develop a generic spelling detection tool for Indo-European languages (of which English and German are examples).

As a starting point, we have manually-built letter replacement rules for English and German. We will compare these as a means of highlighting the similarity between them. We will describe machine learning approaches developed by the German team and apply them to manually-derived 'historical variant'-'modern equivalent' pairs (derived from existing corpora of English and German) to determine whether we can derive similar letter replacement heuristics. Using the manually-derived heuristics as a gold-standard we will evaluate the automatically derived rules.

Our prediction is that if the technique works in both languages, it would suggest that we are able to develop generic letter-replacement heuristics for the identification of historical variants for Indo-European languages.

2. German spelling variation

The interdisciplinary project "Rule based search in text databases with non-standard orthography" which is funded by the Deutsche Forschungsgemeinschaft [German Research Foundation] developed a rule-based fuzzy search-engine for historical texts (Pilz *et al.* 2005). Its aim of RSNSR is to provide means to perform reliable full text-search in documents written prior to the German unification of orthography in 1901.

On basis of around 4,000 manually collected one-to-one word mappings between non-standard and modern spellings, RSNSR follows three different paths to come up with an efficient rule set. Those are manual rule derivation, trained string edit distance and automatic rule learning. Additional mappings will be collected to further enhance the quality of those approaches. The manual derivation uses an alphabet of 62 different sequences, in parts historical *n*-graphs (e.g. <a>, <äu>, <eau>), built from combinations of the 30 standard graphemes of the German language. Being built manually, the alphabet considers linguistic restraints. Neither in context nor at the position of substitution non-lingual *n*-graphs (i.e. grapheme sequences that directly correspond to phonemes) are allowed. The context may also feature regular expressions using the *java.util.regex* formalism. The manually derived gold

standard features the most elaborate rules. However the design of a rule set for the period from 1803 to 1806, based on only 338 *evidences* took about three days to create. Furthermore, the manual derivation is prone to human-error. This is especially true as soon as the rule set exceeds certain limits where side effects become more and more likely.

The algorithm used to calculate the edit costs was proposed in 1975 by Bahl and Jelinek and taken up 1997 by Ristad and Yianilos who extended the approach by machine learning abilities. The authors applied the algorithm to the problem of learning the pronunciation of words in conversational speech (Ristad and Yianilos 1997). In a comparison between 13 different edit distances, Ristad and Yianilos' algorithm proofed to be the most efficient one. Its error rate on our list of *evidences* was 2.6 times lower than the standard Levenshtein distance measure and more than 6.7 times lower than Soundex (Kempken 2005).

The automatic generation of transformation rules uses triplets containing the contemporary words, their historic spelling variant and the collection frequency of the spelling variant. First, we compare the two words and determine so called 'rule cores'. We determine the necessary transformations for each training example and also identify the corresponding context. In a second step, we generate rule candidates that also consider the context information from the contemporary word. Finally, in the third step we select the useful rules by pruning the candidate set with a proprietary extension of the PRISM algorithm (Cendrowska 1987).

For this paper, we compared the German gold standard, mentioned above, with the two different machine learning algorithms. The string learning algorithm produces a fixed amount of single letter replacement probabilities. It is not yet possible to gather contextual information. Bi- or tri-graph operations are reflected by subsequent application of letter replacements. Therefore they do not map directly onto the manual rules. However, the four most frequent replacements, excluding identities, correspond to the four most frequently used rules. For the period from 1800 to 1806 these are *T*→*TH*, *Ä*→*AE*, _→*E* and *E*→_.

The manual and the automatic derived rules show obvious similarities, too. 12 of the 20 most frequently used rules from the automatic approach are also included in the manually built rules. For six other rules equivalent rules in the manual rule set exist. The rule *T*→*ET* from the automatic approach, for example, corresponds to the more generalised form _→*E* taken from the manual approach. And again do the first four rules match the four most frequent gold standard ones.

The automatic approaches, rule generation as well as edit distance, could be enhanced by a manual checking. Nevertheless, even a semi-automatic algorithm allows us to save time and resources. It is furthermore obvious, that the machine learning is already able to provide with a highly capable rule set for historical documents of German language.

## 3. English spelling variation

The existing English system called VARD (VARiant Detector) has three components. First, a list of 45,805 variant forms and their modern equivalents, built by hand. This provides a one-to-one mapping which VARD uses to insert a modern form alongside the historical variant which is preserved using an XML 'reg' tag. Secondly, a small set of contextual rules which take the form of templates of words and part-of-speech tags. The templates are applied to find real-word variants such as 'then' instead of 'than', 'doe' instead of 'do', 'bee' for 'be' and detection of the genitive when an apostrophe is missing. The third component consists of manually crafted letter replacement heuristics designed during the collection of the one-to-one mapping table and intended to reduce the manual overhead for detection of unseen variants in new corpora.

The rationale behind the VARD tool is to detect and normalise spelling variants to their modern equivalent in running text. This will enable techniques from corpus linguistics to be applied more accurately (Rayson et al., 2005). Techniques such as frequency profiling, concordancing, annotation and collocation extraction will not perform well with multiple variants of each word type in a corpus.

The English manual and automatically derived rules show a great deal of similarity. Nine of the twenty most frequent automatically derived rules are in the manual set. Eight other automatically derived rules have equivalents if we ignore context. Three automatically derived rules do not have a match in the manual version.

## 4. Conclusion

The motivation behind the two approaches of VARD and RSNSR differs. This reflects on the overall structure of rules as well. While VARD is used to automatically normalise variants and thus takes more accurate aim to determine the correct modern equivalent, RSNSR focuses on finding and highlighting those historical spellings. Therefore its demands for precision are diminished while recall is the much more important factor. However, the approaches are highly capable of supporting each other and expanding their original field of application.

## References


Cendrowska, J. (1987). PRISM: an algorithm for inducing modular rules. *Int. J Man-Machine Studies*, 27(4), pp.349-370.

Kempken, S. (2005). *Bewertung von historischen und regionalen Schreibvarianten mit Hilfe von Abstandsmaßen*. Diploma thesis. Universität Duisburg-Essen

Pilz, T., Luther, W., Ammon, U., Fuhr, N. (2005). Rule-based search in text databases with nonstandard orthography, *Proceedings ACH/ALLC 2005*, Victoria, 15 - 18 Jun 2005.

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. In *proceedings of the Corpus Linguistics 2005 conference*, July 14-17, Birmingham, UK.

Ristad, E., Yianilos, P. (1997). Learning String Edit Distance, *IEEE Trans. PAMI*, 1997

Hessisches Staatsarchiv Darmstadt. http://www.stad.hessen.de/DigitalesArchiv/anfang.html (accessed 25 Nov. 2005)

Bibliotheca Augustana. FH Augsburg. http://www.fh-augsburg.de/~harsch/augustana.html (accessed 25 Nov. 2005)

documentArchiv.de http://www.documentarchiv.de (accessed 25 Nov. 2005)