The Identification of Spelling Variants in English and German Historical Texts: Manual or Automatic?

Thomas Pilz, Andrea Ernst-Gerlach and Sebastian Kempken
Department of Computer Science and Applied Cognitive Science,
Faculty of Engineering, University of Duisburg-Essen, D-47048
Duisburg, Lotharstr. 65, Germany

Paul Rayson
Computing Department, Infolab21, Lancaster University, Lancaster
LA1 4WA, UK

Dawn Archer
Department of Humanities, University of Central Lancashire,
Preston PR1 2HE, UK

In this article, we describe the respective approaches we have taken when addressing issues of spelling variation in German and English historical texts. More specifically, we describe an experiment to evaluate automatic techniques for the development of letter replacement heuristics against manually created gold standards of known letter replacements rules. As will become clear, the motivation for the research differs according to the team of researchers: the German researchers are seeking to develop a search engine for historical texts; the English researchers want to improve the results obtained when applying corpus linguistic techniques (developed for modern language) to historical data. However, the respective teams do share a longer term goal of assessing whether it is possible to develop a generic spelling detection tool for Indo-European languages.

Correspondence:
Paul Rayson,
Computing Department,
Infolab21, Lancaster
University, Lancaster
LA1 4WA, UK.
E-mail:
paul@comp.lancs.ac.uk

1 Introduction

In this article, 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 (seventeenth–nineteenth century). The approaches taken by the respective teams differ in a number of respects. Take the use of contextual rules, for example: in the German system, contextual 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 respecting the detection of real-word spelling

doi:10.1093/llc/fqm044 Advance Access Published on 24 November 2007

variants (such as when 'then' has been used where a modern reader might expect 'than'). Our motivations for undertaking the research are also different: the German researchers are seeking to develop a search engine for historical texts; the English researchers want to improve the results obtained when applying corpus linguistic techniques (developed for modern language) to historical data. These differences apart, we have noticed that we seem to be addressing similar issues when seeking to identify English and German variants. As we will report here, we also seem to be identifying similar letter replacement patterns in the two languages. The purpose of this article, then, is to highlight these similarities by comparing manual and automatic techniques for the development of letter replacement heuristics in German and English.

We are also using the opportunity that this collaborative endeavour affords to assess whether the overlap between the (German and English) letter-replacement heuristics is sufficient to allow for the development of a generic spelling detection tool for Indo-European languages (of which German and English are examples). In particular, we will apply the machine-learning approaches developed by the German team to the lists of manually derived 'historical variant'–'modern equivalent' pairs derived from existing corpora of English and German as a means of determining whether we can derive similar letter replacement heuristics automatically (to those derived manually). This means that we are using the manually derived heuristics (for German and for English) as gold standards against which to evaluate the automatically derived rules. Our prediction is that, if the technique works in both languages, it would suggest that we are one step closer to developing generic letter-replacement heuristics for the identification of historical variants for Indo-European languages.

We begin with a brief summary of the different approaches adopted by the German and English researchers (see Sections 2 and 3), and, after introducing the machine-learning approaches (Section 4), go on to determine their success rate in detecting variants as well as the similarities between the approaches (Section 5).

2 German Spelling Variation

The interdisciplinary project RSNSR (Regelbasierte Suche in Textdatenbanken mit nichtstandardisierter Rechtschreibung, ''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.*, 2006). The aim of RSNSR is to provide means to perform reliable full text-searches in historical documents[1]. There are numerous problems involved when dealing with old text, not least because the optical character recognition is error-prone (due to varying hot types and obsolete characters). This is further complicated in German, as German orthography was not officially regulated prior to the German unification of orthography in 1901. Consequently, one needs to have profound knowledge of historical spelling-variation to retrieve satisfying results.

On the basis of 12,621 manually collected one-to-one word mappings between non-standard and modern spellings, RSNSR follows three different paths to reach its goal: manual rule derivation, trained string edit distance and automatic rule learning. The existing manual rules use an alphabet of sixty-two 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. The context supports regular expressions using the *java.util.regex* formalism with some minor extensions to ease the input of linguistic data. Those cover the definition of phoneme-groups and special characters in Unicode. A graphical interface allows for easier development and customization of the rules. However, the design of a rule set for the period from 1803 to 1806, based on only 338 pairs of evidences, took about three days to create. 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. In spite of these potential 'problems', the most elaborate rules are still found within the manually derived gold standard. This, in part, explains how (and why) the automatic approaches came into focus. Before describing those automatic approaches in detail (Section 4), we will first introduce the variant detector developed by the English researchers.

3 English Spelling Variation

The existing English system called VARD (VARiant Detector) has three components. Firstly, 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. Second, a small set of contextual rules that 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, which were designed during the collection of the one-to-one mapping table and are intended to reduce the manual overhead for detection of unseen variants in new corpora.

The rationale behind the VARD tool is to detect and normalize 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.

4 The Automatic Approaches

Distance metrics are commonly used in dialectrometry to calculate the distance/similarity between different dialect variants. There are numerous different methods, which have certain advantages and disadvantages. The *FlexMetric* datatype allows the implementation of all those metric, their direct comparison and the subsequent use in our environment.

To determine the most efficient algorithm for the problem at hand, thirteen different distance metrics were compared with each other (Kempken, 2005). The algorithm that proved best to calculate the edit costs between modern and historical spellings was proposed in 1975 by Bahl and Jelinek and taken up again in 1997 by Ristad and Yianilos, who extended the approach by machine learning abilities. One consequence of this is that it can be easily trained on a list of evidences. Ristad and Yianilos (1997) originally applied the algorithm to the problem of learning the pronunciation of words in conversational speech. In our latest evaluation, its error rate on the 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). Our research showed that about 4,500 evidences are sufficient to represent the spelling variation of a given timeframe with satisfying accuracy. The development of enhanced metrics is still ongoing.

The automatic generation of transformation rules (described in detail in Ernst-Gerlach and Fuhr, 2006) uses triplets containing the contemporary words, their historic spelling variant and the collection frequency of the spelling variant. Our approach has been to compare the two words as a means of determining so called *rule cores*. We then determine the necessary transformations for each training example and also identify the corresponding context. A historical variant of the word '*unnütz*' (= useless), for example, is '*unnuts*'. For this pair, we get the following two-element set of rule cores:

$$\{('unn', ('\ddot{u}', 'u'), 't'), ('t', ('z', 's'), '')\}.$$

As a second step, we generate rule candidates that also takes account of the context information (e.g. *consonant* or *word-ending*[2]) of the contemporary word. If we use the example shown above, we find that the following candidate rules are generated (amongst others):

$$('z', 's'), ('tz', 'ts'), ('z\$', 's\$'), ('tz\$', 'ts\$'),$$
$$('Cz\$', a'Cs\$')$$

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). The corresponding algorithms offer various possibilities for optimizations (e.g. utilization of phonetic methods), which are currently under development.

5 Comparisons

For this article, we compared the German gold standard, mentioned earlier, with the two different machine-learning algorithms and also the English gold standard. Please note that, when discussing the comparison of edit operations and letter replacements, we will not speak of matches but *correspondences* so that we pay attention to the fact that edit operations and replacement rules are in *n:m* relation to each other. That said, a relation will only be called *corresponding* if both sides can be said to form a complete match.

The English replacements were built with the intention of detecting spelling variants more easily. In contrast, the German rules are used to generate such variants. The German rules were gathered by taking into account an overview of the evidences, whereas the English rules were meticulously built to fit specific variants. Taking all these differences into account, the similarities between the two data sets are remarkable.

5.1 Gold standards

The first apparent similarity between the two gold standards is the amount of replacements, i.e. sixty-eight for German and fifty-two for English. It should be noted that this slight difference in amount may be due to the German umlaut-characters and also the more general nature of the English letter replacements. Our experiences in training rule sets suggests that, after only about ten variant spellings, even quite specific rules tend to apply repeatedly. An amount of fifty—eighty rules is often generated by 100–200 variants. After this, the rule sets cease to grow, and tend to become more and more generalized. In respect to content, the gold standards are obviously related. A total of 46% (English) and 48% (German) of the operations are concerned with vowel sounds and of those 23% (English) and 20% (German) with <e> alone. The remaining patterns within the gold standards tend to capture (consonants that relate to the First and Second Germanic) consonant shifts (<d>, <t>, <b>, <s>, <z>) and common gemination (e.g. <t>→<tt>).

5.2 OCR errors

Real-world examples often tend to be overly complex. Indeed, the historical spellings we are working with in both projects exhibit a great deal of variation (Pilz *et al.*, 2006). An additional problem we face is the character variation in optical character recognition (OCR). Consequently, we use the language-independent problem of faulty OCR as a first step to the automatic approaches. This problem is also very common in retro-digitization projects of course and, as such, is in no way limited to this field of research.

The manual rule set for OCR consists of seventy-three weighted rules. As previously mentioned, the edit costs used in the distance metric cannot be directly mapped onto those rules. We therefore add single edit operations together as a means of allowing for the performance of more complex replacements. Bi- or tri-graph operations, for example, are reflected by the subsequent application of letter replacements. A current diploma thesis is developing methods to gather this inherent context-information from the most cost-efficient transformations. A direct comparison reveals a large amount of analogy between manual rules and edit costs as shown in Table 1. This fact is not overly surprising, since the efficiency of distance metrics in spelling variation has already been stated (Heeringa *et al.*, 2006). Sorted by their edit costs, the ten topmost edit operations are all within the maximally weighted rules. It should be noted that the rank of the edit operations refers to the operation's costs, and that the rank of the replacement rules represents the sum of the rule weights of all rules the edit operation is used within. Also, since some of the results are equal, there are multiple second and third places.

Some thirty-nine of the fifty-seven generated rules directly match with the prominent manual ones. Additionally, nine match if we ignore the context and five manual rules could be built from combined generated rules. All ten top ranked manual rules have also been generated. In contrast to the edit operations procedure, we also gather *n*-grams and context information. Table 1 shows the degree of correlation between manual and automatic rules. The differences in the ranks are caused through different ranking strategies. The manual replacement rules are ranked

Table 1 Comparison of the manual and automatic approaches

Edit operations		Replacement rules	
Rank	**Operation**	**Rank**	**Operation used in**
1.	r→i	1.	r→i
2.	r→ε	2.	ri→n, ri→u
3.	i→ε	3.	in→m, n→ii
4.	n→m	3..	in→m, n→m
5.	i→n	2.	i→n, n→ii, ri→n
6.	f→s	3.	f→s
7.	n→u	2.	n→u
8.	ü→u	4.	ü→u
9.	s→f	3.	s→f
10.	ä→a	3.	ä→a
Replacement rules		**Automatic Rules**	
Rank	Operation	Rank	Operation
1.	r→i	5.	r→i
2.	n→u	1.	n→u
3.	f→s	9.	f→s [e, r, ä]
3.	s→f	9.	s→f
3.	ä→a	1.	ä→a
4.	in→m	1.	in→m
4.	ri→n	1.	ri→n
4.	ü→u	7.	ü→u
5.	b→h	1.	b→h

by the occurrence, i.e. how often a rule is used. The ranking of the automatic rule generation is based on the probability ranking principle (following Robertson, 1977). Following this principle, optimum retrieval is achieved when the ranking of the documents is ordered by decreasing values of the probabilities of relevance. We automatically generated false one-to-one word mappings that are used for calculating precision values for the rules, which document the probabilities of relevance for the ranking process. It is worth noting that automatic rule generation already shows good results independently of the different ranking schemes.

5.3 Historical spelling variation

The comparison of the German gold standards with the edit operations yields similar or even better results. The first replacements are already very familiar to the German historical linguist (cf. insert or remove <e>, insert <h>, replace <s> with <ß>, replace <z> with <c>…). Within the thirty-one most frequent operations that cover 99.8% of all probable replacements, there is not a single one that does not correspond to at least one entry of the manual rule set. Moreover, these operations often correspond to two or more rules (e.g. _→h: f→ph, g→ch). 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 also show obvious similarities. Indeed, twelve of the twenty most frequently used rules from the automatic approach are also included in the manually built rules. Moreover, we can easily determine equivalent rules for a further six within the manual. For example, the rule t→et from the automatic approach corresponds to the more generalized form _→e taken from the manual approach. It is also worth noting that the first four rules match the four most frequent gold standard ones.

The English letter replacements used in VARD are not weighted: every operation features an equal probability of application. Because of the more general character of the English letter replacements, a comparison with the edit costs is eased. The six most probable edit operations directly correspond to the replacements. All of these fifty-two replacements are covered by the most probable 6.7% (147 instances) of the edit operations. Comparing the English automatic rules with those the VARD letter replacements, nine of the twenty most frequent automatically derived rules are in the manual set. Eight additional automatically derived rules have equivalents if we ignore context. Three automatically derived rules do not have a match in the manual version. We always have to take into account that the automatic methods gather their knowledge through direct evidences from input data: put simply, they represent a more objective view and may even find variation that was not discovered manually. It should be noted that we are of the opinion that, even if rules or edit operations are found that are unlikely or unwanted, it is of value to discover what may have led to those results.

Table 2 Evaluation of German metrics on English variants

Metric	Std. Levenshtein	German trained			English trained
		15th–16th century	13th–16th century	13th–15th century	
Recall (normalized)	0.89	0.91	0.93	0.94	1

5.4 Languages

Since the automatic approaches are already able to reproduce the manual rules to a satisfying extent (albeit in much less time), the automatic creation of generic letter-replacement heuristics for both English and German seems to be a worthwhile task. Indeed, it seems that heuristics for constricted regions or time-spans as well as specific purposes (e.g. typos or faulty OCR) can be trained within minutes. It is also worth noting that a recent cross-language evaluation proved that a deliberately reduced amount of training data can produce more capable metrics. Put simply, due to less divergent edit operations, the metrics with a narrowed focus often yield better results than more elaborate replacements. Table 2 shows the result of this evaluation.

Notice that five different distance metrics were evaluated on the same data set of historical English. If we normalize the recall of a metric trained on English historical spelling variants to one, the standard Levenshtein algorithm is 11% less efficient. A metric trained on German spelling variants of the fifteenth and sixteenth century yields slightly better results, and the efficiency can be increased still further, if we expand the time-span to the thiteenth century. To apply metrics in this manner reminds one of the approaches followed in authorship attribution, i.e. it is a way of reflecting the differences (and also the similarities) between two related languages. Given the kinship of English and German the result should not be unduly surprising: in addition to the obvious collective usage of the Latin alphabet of both languages, their letter frequencies (fingerprints) are noticeably related. Only <o> in English and <u> in German are not shared within the ten most frequent letters (while German <u> substitutes English <oo>: *Buch – book*). Furthermore, some similarities are based in general phonetic rules that apply to both languages. Neef (2005, p. 16)[3] states that the graphemic form of a spelling has to ensure the recodability of the corresponding phonetic form: this means the utilization of graphemes is limited in reference to the grapheme–phoneme correspondence (GPC) of a certain language. As a result, the amount of possible spelling variants—Neef (2005, p. 12) defines this as the graphematic space of solution—is also very limited. The relatedness of the articulatory features of vowels and stop classes of constants in English and German also allows for interesting similarities: *persewe – pursue* (engl.) and *trewen – treuen* (germ.), *preste – priest* (engl.) and *wilch – welche* (germ.), *abandom'd – abandoned* (engl.) and *samfft – sanft* (germ.). The substitution of <i>/<y> and gemination can be found in numerous cases in both languages: *abilyties – abilities* (engl.) and *sy – sie* (germ.), *agaain – again* (engl.) and *Maaβ – Maβ* (germ.).

Interestingly, the recall of distance measure can be raised by at least a further percent by narrowing the time-span to the thirteenth–fifteenth century. We would therefore suggest that German spelling variation of the sixteenth century does not fully represent English variation of the same period. There are socio-historical (as well as linguistic) reasons for this, of course. For example, although English and German are both West Germanic languages, a number of German consonants were particularly affected by the *High German* consonant shift. However, as the name of the shift intimates, it did not particularly affect Low German until after the ninth century (i.e. the fourth phase) and then only minimally. While this meant that many Low German words sounded more similar to their English counterparts than High German words, Low German began to suffer decline from the sixteenth century (after a climax during the time of the Hanseatic League), allowing High German to become the 'standard'.

6 Conclusion

Our study of historical German and English texts has shown that the automatic approaches we adopt (i.e. rule generation and edit distance) can be enhanced in several ways. For example, we would argue that manual intervention is a good option for the rule generation process, but not advisable for processes that involve edit distance. We would also suggest that, while we accept that the effects of manual changes are very difficult to grasp (not least because of the metrics work involved in a full (i.e. |alphabet| × |alphabet|) substitution matrix), we believe that our study has demonstrated that a semi-automatic algorithm can allow us (as researchers) to save valuable time and resources. It is therefore comforting to note that further work is being undertaken: indeed, a diploma thesis is currently investigating the overlay and combination of different metrics as well as the use of metrics and rule sets. We would also further suggest that machine learning can already provide us with a means of identifying a highly capable rule set for the identification of historical spelling variants within German and English historical texts and, in turn, sophisticated metrics for the investigation of those spelling variants within these languages.

As we highlighted in our introduction, the motivation behind the two approaches we introduce in this chapter—VARD and RSNSR—are not the same. Inevitably, this will reflect on the overall structure of the rules we have identified: indeed, while the VARD tool is used to automatically normalize variants and thus seeks to take a more accurate aim in respect to determining the correct modern equivalent, RSNSR focuses on finding and highlighting those historical spellings (i.e. the aim is not to normalize them but only to identify them, so that a given user of the tool might appreciate the link between the latter and their modernized form). As such, the demands for precision are diminished in the German tool, while recall is the much more important factor. Even so, we believe that the two approaches are highly capable of supporting each other and, by so doing, also serve to expand their original field of application. Indeed, we would contend that such research provides an important step in identifying generic rules for letter replacement heuristics that can be identified as Indo-European in origin (or, more specifically, influenced by Germanic languages). We would also advocate that this research needs to be expanded, in turn, to the romantic languages so that we can better determine those rules that potentially capture the relationship between romance languages and languages (like English) which have historically borrowed from them.

References

Bibliotheca Augustana. FH Augsburg. http://www.fh-augsburg.de/~harsch/augustana.html (accessed 29 March 2007).

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

documentArchiv.de. http://www.documentarchiv.de (accessed 29 March 2007).

Ernst-Gerlach, A. and Fuhr, N. (2006). Generating Search Term Variants for Text Collections with Historic Spellings. Advances in Information Retrieval, 28th European Conference on IR Research, ECIR 2006, London, UK, April 10–12 2006, Proceedings in *Lecture Notes in Computer Science 3936*. Springer 2006, pp. 49–60.

Heeringa, W., Kleiweg, P., Gooskens, C., and Nerbonne, J. (2006). *Evaluation of String Distance Algorithms for Dialectology. Proceedings of the Workshop on Linguistic Distances*, Sydney, AUS, July 2006, Association for Computational Linguistics, pp. 51–62.

Hessisches Staatsarchiv Darmstadt. http://www.stad.hessen.de/DigitalesArchiv/anfang.html (accessed 29 March 2007).

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

Neef, M. (2005). *Die Graphematik des Deutschen. Niemeyer (Linguistische Arbeiten, 500), Tübingen.*

Pilz, T., Luther, W., Ammon, U., and Fuhr, N. (2006). Rule-based search in text databases with nonstandard orthography. *Literary and Linguistic Computing*, **21**: 179–86.

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 of the Corpus Linguistics 2005 conference*. UK: Birmingham. July 14–17.

Ristad, E. S. and Yianilos, P. N. (1997). 'Learning string edit distance'. In Fisher, D. (ed.) *Machine Learning: Proceedings of the Fourteenth International Conference* (San Francisco, 8–11 July 1997), Morgan Kaufmann, pp. 287–295. http://citeseer.ist.psu.edu/ristad97learning.html.

Robertson, S. (1977). The probability ranking principle in IR. *Journal of Documentation*, **33**: 294–304.

Notes

[1] The German training data was provided courtesy of the following archives: Hessisches Staatsarchiv Darmstadt, Bibliotheca Augustana, and documentArchiv.de.

[2] $ denotes the word-ending, C denotes a consonant

[3] ''Eine graphematische Form muss die Rekodierbarkeit der korrespondierenden phonologischen Form gewährlesisten.''