Spelling Normalization of Historical German with Sparse Training Data

Marcel Bollmann
Department of Linguistics
Ruhr-Universität Bochum, Germany
bollmann@linguistics.rub.de

1 Introduction[1]

Recently, there has been a growing interest in historical language corpora. Projects to create such corpora exist for a variety of languages such as German (Scheible et al. 2011), Spanish (Sánchez-Marco et al. 2010), or Slovene (Erjavec 2012). Annotation of these corpora is complicated by the fact that specialized tools for these language stages are typically not available. A common approach is to employ spelling normalization to map historical wordforms to modern ones (e.g., Adesam et al. 2012, Baron et al. 2009, Jurish 2010), so that existing tools for modern language (e.g., modern POS taggers) can be used on the normalized data.

This paper presents an approach to spelling normalization that combines three different normalization algorithms and evaluates it on a diverse set of texts of historical German. The evaluation shows that this approach produces acceptable results even with comparatively small amounts of training data. The normalization methods were previously described in Bollmann (2012), though with a much more restricted evaluation.

2 Normalization Methods

Spelling normalization is performed using three different methods: wordlist mapping, rule-based normalization, and weighted Levenshtein distance (WLD). All methods operate on a single wordform at a time, without taking token context into account.

Wordlist mapping refers to the use of a pre-defined list (or "dictionary") of historical wordforms to perform word-by-word substitutions, without any notion of characters or spelling variation. Mappings can be manually defined or learned from training data.

Rule-based normalization was first presented in Bollmann et al. (2011). It is based on the concept of character rewrite rules which operate on one or more characters and take their immediate context into account. They can be written in a form similar to phonological rewrite rules, e.g.:

- $v \rightarrow u \,/\, \#\_n$

This rule describes the substitution of "v" by "u" between a word boundary and the character "n".

Rewrite rules can be extracted automatically from an aligned training corpus. For this purpose, a modified Levenshtein algorithm is used, where instead of counting the number of edit operations, the actual edit operations are recorded. During normalization, each historical wordform is processed character by character, with rewrite rules being applied at each position depending on their frequency in the training data. Additionally, to prevent the generation of nonsense words, normalization candidates are checked against a modern lexicon.

Weighted Levenshtein distance (WLD) is a measure of distance between two wordforms. It is a variant of classic Levenshtein distance where each edit operation can be assigned an individual weight (typically between 0 and 1). As an example, when normalizing Early New High German texts, the substitution $v \rightarrow u$ could be assigned a relatively low weight, modelling the fact that it is a common spelling variant and much more likely to be a correct normalization than substitutions of unrelated characters, such as $v \rightarrow x$. The notion of WLD can be extended to n-grams as well, assigning weights to substitutions such as $ow \rightarrow au$; in this study, n-grams of a length up to three characters are considered. Additionally, as spelling variations between historical and modern texts are typically not symmetric, a directed version of WLD is used here, meaning that the substitutions $v \rightarrow u$ and $u \rightarrow v$ are not required to have the same weight.

Using these definitions of WLD, the measure can be used for normalization by finding the entry in a modern lexicon which has the lowest distance to the historical input string. While weights were defined manually in Bollmann (2012), a learning algorithm has been implemented for this study. It roughly follows the approach outlined in Adesam et al. (2012). First, word pairs in a training corpus are aligned on a character level using iterated Levenshtein distance alignment (Wieling et al. 2009). Afterwards, weights are calculated using the following formula:

$$-\frac{1}{d}\log\left(p_{\alpha}(RHS|LHS)\right) \qquad (1)$$

Here, LHS/RHS are the left- and right-hand sides of a character substitution, respectively; p_α refers to the (conditional) probability of the characters with additive smoothing (using $\alpha = 0.5$); and d is a scale factor to bring the weights in line with the default substitution cost of 1 (using $d = 7$).

[1] The research reported here was financed by Deutsche Forschungsgemeinschaft (DFG), Grant DI 1558/4-1.

	Dating	Size	Baseline	Mapper	Rule-based	WLD	Combined
Berlin	15c	4,700	23.05%	62.05%	63.17%	60.71%	75.07%
Melk	15c	4,541	39.32%	63.15%	64.14%	69.34%	74.49%
Sermon1	1677	2,178	72.71%	76.46%	78.67%	76.40%	79.56%
Sermon2	1730	2,137	79.47%	85.22%	88.52%	88.15%	91.81%
Sermon3	1770	1,953	83.41%	86.58%	90.50%	95.46%	95.73%

Table 1. Dating, size (in number of tokens, excluding punctuation and foreign words), and normalization accuracy per text for different normalization methods after training on the first 500 tokens and evaluating on the rest.

Finally, these three normalization methods are combined in the form of a chain. First, the wordlist mapping approach is used to check if a modern equivalent for the historical input string is already known. Only if this is not the case, the rule-based method is applied. However, as the character rewrite rules depend on contexts, it is possible for this method to fail as well if the input word contains a previously unseen combination of characters. In this case, the WLD algorithm is used, which is guaranteed to find a normalization candidate. This order of normalization methods was found to perform best on average; also, chaining the methods in this way performed better than using a majority vote approach.

3 Corpora

Texts from two different corpora were used for the evaluation: the Anselm corpus and the GerManC-GS corpus (Scheible et al. 2011).

The Anselm corpus consists of more than 50 German manuscripts and prints of the text “Interrogatio Sancti Anselmi de Passione Domini” (“Questions by Saint Anselm about the Lord's Passion”). It is being created in the context of an ongoing, interdisciplinary research project, which also aims to provide a digital, annotated edition of the corpus. The texts were written between the 14th and 16th centuries in various German dialects, showing a great deal of spelling variation not only compared to modern German, but also between each other: e.g., spellings of the modern word *Frau* “woman” include *fraw*, *frouw*, *vrowe*, and *vrouwe*.

Two of the texts were manually normalized and are used for the evaluation: a manuscript in an Eastern Upper German dialect kept in Melk, Austria; and an Eastern Central German manuscript kept in Berlin.

The GerManC corpus aims to be a representative corpus of historical, written German from 1650 to 1800. It contains texts from different dialectal regions and genres. GerManC-GS is a subcorpus of GerManC containing gold standard annotations of normalization, lemmatization, and POS tags. In this paper, the texts of the genre “sermon” are used[2], as they are of a religious nature similar to the Anselm data. With the oldest of the texts being written in 1677, they are more recent than the Anselm texts and much closer to modern German spelling.

4 Evaluation

For each text, all normalization methods are evaluated both separately and in the chain combination described in Sec. 2. Punctuation and foreign words were removed before the evaluation: punctuation marks are trivial to normalize and could bias the results, while the spelling of foreign words is unlikely to be relevant for the spelling in the main language.

Furthermore, all normalization methods presented above require training data before they can be applied. Therefore, for each text, the first *n* tokens are used to train the normalizers, and evaluation is performed on the remainder of the text.

Table 1 shows the results per text for each normalization approach when the first 500 tokens are used for training. Accuracy is given in percentage of matching tokens compared to the gold standard normalization. The baseline accuracy, i.e., the number of tokens already identical to modern spelling, differs greatly between the texts. The Berlin text only has a baseline of 23.05%, i.e., it is very far from modern German spelling, while the more recent Sermon texts are much less problematic, with baselines as high as 83.41%.

For the automatic normalization, the combination of all three methods produces the best result in each case, achieving an increase of up to 12 percentage points compared to just using a single method. In general, the best results are achieved for the texts which have the least variation to start with, e.g., 95.73% for the most recent Sermon text from 1770. The biggest increase, however, is found for the Berlin text, which is brought from 23% to 75% accuracy with the combined normalization approach.

[2] I would like to thank Silke Scheible for kindly providing me with these texts.

	Berlin	Melk	Serm.1	Serm.2	Serm.3
Base	23.05%	39.32%	72.71%	79.47%	83.41%
100	57.48%	69.15%	77.53%	86.55%	91.37%
250	73.71%	73.90%	79.62%	88.18%	94.54%
500	75.07%	74.49%	79.56%	91.81%	95.73%
1,000	77.57%	76.76%	82.60%	92.79%	96.12%
2,000	81.52%	78.16%	—	—	—

Table 2. Accuracy per text using the combined normalization approach, for various sizes of the training portion.

This is a remarkable improvement, especially if we consider the small amount of training data that was used, and should greatly facilitate further processing of the data.

Additionally, for the combined normalization approach, another evaluation was performed with different sizes of the training part. Table 2 presents the results. Unsurprisingly, accuracy typically increases with larger training parts; however, even with only 100 tokens for training, there is a notable increase from the baseline (e.g., from 39% to 69% for Melk). Using 250 tokens already yields scores similar to those for 500 tokens, and in general, the increase in accuracy is lower above 250 tokens compared to the leap from 100 to 250 tokens. These figures show that even very small amounts of training data can be useful for normalization using this method.

5 Conclusion and Future Work

I have presented an approach to spelling normalization for historical texts that utilizes a combination of three different normalization algorithms. The approach was evaluated on different types of historical German texts from the 15th to 18th century. When trained on a fraction of the same text that is to be normalized, it achieves normalization accuracies between 75% and 95%, depending on the extent of the spelling variation in the input data.

While this normalization method requires a part of the text to be manually normalized for training, it already achieves good results even with only a few hundred tokens as training data. This makes it especially suited for a semi-automatic normalization approach, where a user confirms or corrects suggestions made by the automatic normalizer, which in turn can be expected to gradually improve with more input from the user.

The most important aspect for future research is the inclusion of token context for the normalization process, similar to Jurish (2010). As spelling variants can be ambiguous, processing wordforms in isolation effectively puts a limit on the maximum accuracy that can be achieved. Furthermore, it is conceivable that the exact composition of normalization algorithms can still be improved. An interesting property of the current configuration is that it proceeds from larger units of operation (full wordforms for the wordlist mapper) to smaller ones (single characters for WLD), which could turn out to be a key for its good performance. Future work could try to expand upon this hypothesis by utilizing even more levels of granularity.

References

Adesam, Y., Ahlberg, M. and Bouma, G. 2012. “bokstaffua, bokstaffwa, bokstafwa, bokstaua, bokstawa... Towards lexical link-up for a corpus of Old Swedish”. In *Proceedings of the 11th Conference on Natural Language Processing (KONVENS 2012), LThist 2012 workshop*, pp. 365–369. Vienna, Austria.

Baron, A., Rayson, P. and Archer, D. 2009. “Automatic standardization for spelling of historical text mining”. In *Proceedings of Digital Humanities 2009.* Maryland, USA.

Bollmann, M., Petran, F. and Dipper, S. 2011. “Rule-based normalization of historical texts”. In *Proceedings of the International Workshop on Language Technologies for Digital Humanities and Cultural Heritage*, pp. 34–42. Hissar, Bulgaria.

Bollmann, M. 2012. “(Semi-)automatic normalization of historical texts using distance measures and the Norma tool”. In *Proceedings of the Second Workshop on Annotation of Corpora for Research in the Humanities (ACRH-2)*, pp. 3–14. Lisbon, Portugal.

Erjavec, T. 2012. “The goo300k corpus of historical Slovene”. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation*, pp. 2257–2260.

Jurish, B. 2010. “More than words: using token context to improve canonicalization of historical German”. *Journal for Language Technology and Computational Linguistics* 25 (1): 23–39.

Sánchez-Marco, C., Boleda, G., Fontana, J.M. and Domingo, J. 2010. “Annotation and representation of a diachronic corpus of Spanish”. In *Proceedings of the Seventh Conference on International Language Resources and Evaluation*, pp. 2713–2718.

Scheible, S., Whitt, R.J., Durrell, M. and Bennett, P. 2011. “A gold standard corpus of Early Modern German”. In *Proceedings of the ACL-HLT 2011 Linguistic Annotation Workshop (LAW V)*, pp. 124–128. Portland, Oregon, USA.

Wieling, M., Prokić, J. and Nerbonne J. 2009. “Evaluating the pairwise string alignment of pronunciations”. In *Proceedings of the EACL 2009 Workshop on Language Technology and Resources for Cultural Heritage, Social Sciences, Humanities, and Education (LaTeCH – SHELT&R 2009)*, pp. 26–34. Athens, Greece.