Using an Alignment-based Lexicon for Canonicalization of Historical Text —DRAFT—

Bryan Jurish, Henriette Ast
jurish@bbaw.de
Berlin-Brandenburgische Akademie der Wissenschaften

1 Introduction

Virtually all conventional text-based natural language processing techniques – from traditional information retrieval systems to full-fledged parsers – require reference to a fixed lexicon accessed by surface form, typically trained from or constructed for synchronic input text adhering strictly to contemporary orthographic conventions. Unorthodox input such as historical text which violates these conventions therefore presents difficulties for any such system due to lexical variants present in the input but missing from the application lexicon. *Canonicalization* approaches (Rayson et al., 2005; Jurish, 2012; Porta et al., 2013) seek to address these issues by assigning an extant equivalent to each word of the input text and deferring application analysis to these canonical cognates.

Traditional approaches to the problems arising from an attempt to incorporate historical text into such a system rely on the use of additional specialized (often application-specific) lexical resources to explicitly encode known historical variants. The simplest form such lexical resources take is that of simple finite associative lists or "witnessed dictionaries" (Gotscharek et al., 2009b) mapping each known historical form w to a unique canonical cognate $\tilde{w}$. Since no finite lexicon can fully account for highly productive morphological processes like German nominal composition, and since manual construction of a high-coverage lexicon requires a great deal of time and effort, such resources are often considered inadequate for the general task of canonicalizing arbitrary input text (Kempken et al., 2006).

In this paper, we investigate the utility of a finite deterministic canonicalization lexicon semi-automatically constructed from a corpus of historical and contemporary editions of the same texts (Jurish et al., 2013), comparing it to the robust generative finite-state canonicalization architecture described in Jurish (2012), and to a hybrid method which uses a finite lexicon to augment a generative canonicalization architecture.

1.1 Related Work

Rayson et al. (2005) describe an automatic "variant detector" for canonicaliza-

tion[1] of historical English, reporting a substantial improvement in accuracy on a small test set compared to conventional spell-checkers. Inverse canonicalization approaches mapping modern query words to (potential) historical variants have been described by Gotscharek et al. (2009b) and Ernst-Gerlach and Fuhr (2007). Recent work on canonicalization for historical text has focused on the use of context for disambiguation of historical "false friends" (Jurish, 2012; Reffle et al., 2009), the induction of rule-sets for mapping historical to modern forms (Baron and Rayson, 2009; Bollmann et al., 2011), and the rigorous characterization of the mapping task (Jurish, 2012; Porta et al., 2013).

The use of specialized canonicalization lexica for historical document collections has been described by Hauser et al. (2007), who note in particular that explicit lexical mappings may succeed where pattern-based cognate approaches fail, as in the case of extinct historical word forms such as *marcken* ("to trade"). Gotscharek et al. (2009a) describe the manual construction of a canonicalization lexicon or "attestation dictionary" and its application in the context of information retrieval. Scheible et al. (2011); Dipper and Schultz-Balluff (2013) describe manually constructed corpora annotated with canonical cognates, and Jurish et al. (2013) present a semi-automatic bootstrapping procedure for canonicalization corpora using historical and modern editions of the same texts.

2 Materials

For the current investigations, we used the semi-automatic procedure described in Jurish et al. (2013) to bootstrap a canonicalized corpus of historical German in which each (historical) token w is explicitly associated with a (modern) canonical form $\tilde{w}$ by aligning historical texts with contemporary editions of the same texts. The construction is based on the assumptions that the contemporary editions used in the construction adhere to modern orthographic conventions and can therefore be interpreted as the desired canonical output for the respective historical texts on the one hand, and that a large portion of the canonicalization pairs can be expected to be identity pairs in which the historical form is in fact a valid modern form on the other. In attempt to minimize manual annotation effort while maximizing the accuracy of the relevance relation implied by the canonicalization pairs, the historical and contemporary editions were first automatically aligned, and subsequently subjected to a two-phase manual review process of the non-identity alignments. The procedure was applied to 126 volumes of historical German originally published between 1780 and 1901 drawn from the *Deutsches Textarchiv*[2] and contemporary editions of the selected volumes provided by the online libraries Project Gutenberg[3] and Zeno.[4] The resulting corpus contained 5,642,813 tokens of 212,028 distinct $(w, \tilde{w})$-pair types.

Subsequent experience with the resulting corpus indicated that the assumptions underlying the construction procedure were not in fact borne out by the editions used in our construction. In particular, the assumption that the contemporary editions themselves systematically adhere to contemporary orthographic conventions appears to have been unjustified in the current case. While some

[1] Rayson et al. use the term "normalisation".
[2] `http://deutschestextarchiv.de`
[3] `http://www.gutenberg.org`
[4] `http://www.zeno.org`

orthographic normalization – such as the conversion from historical *th* to contemporary *t* as in the mapping pair *Theil*↦*Teil* ("part") – was indeed undertaken by the editors of the contemporary editions, these texts still exhibit a substantial number of unnormalized historical spelling variants. Such unnormalized historical spellings lead to identity canonicalizations of the form (w, w) during the alignment phase of the corpus construction procedure, which were accepted into the output corpus without manual confirmation.[5] The presence of such unnormalized words in the contemporary editions thus leads to identity canonicalizations which are not in fact valid contemporary forms, and thus do not accurately represent a ground-truth canonicalization, such as *andre* ↦ *andre* ≠ *andere* ("other"), *kömmt* ↦ *kömmt* ≠ *kommt* ("comes"), *nich* ↦ *nich* ≠ *nicht* ("not"), and *ward* ↦ *ward* ≠ *wurde* ("was").

In an attempt to ameliorate these shortcomings, the entire corpus was subjected to a document-level review phase. Five volumes (197,925 tokens) were dropped from the corpus due to pervasive orthographic violations in the contemporary editions used for alignment – typically, these were volumes of verse using non-standard capitalization conventions. Using page-wise diagnostic heuristics, a total of 204 pages in 41 volumes were manually selected and purged from the corpus, chiefly due to heavy use of pseudo-phonetic dialect or foreign-language material – for example the entirety of the story *Von den Fischer und siine Fru* ("of the fischer and his wife", written entirely in Low German) was purged from the Grimms' fairy tales in this fashion.

Many alignment errors were found to result from irregular hyphenation, explicit elisions or genitive marking using apostrophes, and tokenization errors involving beginning-of-line quotes. In order to remove these errors from the corpus, all 9,250 types (16,300 tokens) containing an apostrophe, quotation mark, or mixture of alphabetic and non-alphabetic characters were flagged as invalid, effectively removing them from further consideration. Finally, the remaining corpus tokens were heuristically checked for consistency with an independently constructed canonicalization lexicon derived from an online error database, and target forms were checked against the TAGH morphology system for contemporary German (Geyken and Hanneforth, 2006). Inconsistent pairs and unknown target forms were flagged as suspicious and are currently undergoing an additional manual review phase. A total of 12,121 types (57,542 tokens) were flagged as suspicious in this manner, and at the time of writing (November, 2012), 55,059 tokens of 9,686 suspicious types have been manually checked and re-incorporated into the corpus. In its current state, the trimmed corpus contains 5,444,888 tokens of 205,055 distinct pair-types. Of these, 4,916,639 tokens of 173,532 distinct types occurred in sentences containing no suspicious or purged material.

2.1 Test Corpus

We used that subset of the corpus which had been subjected to the most thorough manual scrutiny[6] as a ground-truth test corpus for evaluation. After applying the corpus trimming heuristics described above, the test corpus contained 378,300

[5] Nearly half of the output corpus types representing over 81% of tokens were identity pairs, and over 59% of types representing over 87% of tokens were identical modulo transliteration.

[6] The "prototype corpus" as described in Jurish et al. (2013)

tokens of 28,012 distinct pair types in 17,472 sentences. Of these, 319,866 tokens of 27,561 distinct pair types contained only alphabetic characters and were thus considered "word-like". Identity canonicalizations accounted for 250,382 word-like tokens (78%) of 15,454 distinct types (56%).

2.2 Training Corpus

In order to achieve as accurate as possible a picture of the effectiveness of a corpus-induced canonicalization lexicon, all works by or about any author represented in the test set were excluded from the training set. The final training corpus used for the current experiments contained material from 45 distinct authors distributed over 101 volumes published between 1785 and 1901. After removing all sentences containing questionable material, the training corpus contained 4,180,924 tokens of 161,148 distinct pair types in 194,678 sentences. Of these, 3,511,679 tokens of 158,074 distinct pair types were "word-like". Among word-like tokens, 2,737,398 (78%) of 79,882 distinct types (51%) were identity canonicalizations of the form $(w, \tilde{w})$.

2.3 Canonicalization Lexicon

The training corpus described above was used to bootstrap a finite canonicalization lexicon. Raw frequency counts $\mathrm{f}(w, \tilde{w})$ for pairs of historical source word w and contemporary target word $\tilde{w}$ were computed over the pruned training corpus. The finite corpus-based canonicalization lexicon was defined by the simple expedient of mapping each source type w represented in the training corpus to that target type $\textsc{clex}(w)$ with which it occured most frequently:[7]

$$\textsc{clex}(w) = \arg\max_{\tilde{w} \in \mathcal{A}^*} \mathrm{f}(w, \tilde{w}) \tag{1}$$

Of course, such a deterministic type-wise mapping cannot account for any ambiguity whatsoever, but the frequency-maximization heuristic should act to ensure that the correct target form is returned for most input tokens of any known source type. Only 856 the training corpus source types ($< 1\%$) had ambiguous canonicalizations modulo letter case, and only 1,626 training corpus tokens ($< 0.1\%$) would have been incorrectly canonicalized by the frequency maximization heuristic from equation (1). For the effectiveness of a corpus-trained lexicon on previously unseen text, unknown words – words present in the input for which no training data was available – have a far greater impact. In order to extend the finite function $\textsc{clex}(\cdot)$ to a total canonicalization function $\textsc{lex} : \mathcal{A}^* \to \mathcal{A}^*$ which produces some output string for every possible input string, a fallback strategy was implemented which maps any unknown input word to itself:

$$\textsc{lex}(w) = \begin{cases} \textsc{clex}(w) & \text{if defined} \\ w & \text{otherwise} \end{cases} \tag{2}$$

[7] $\mathcal{A}$ is a finite character alphabet. In case multiple maximally frequent target forms were found, one was chosen randomly.

2.4 HMM Canonicalizer

The robust generative canonicalization architecture described in Jurish (2012, Ch. 4) employing a dynamic Hidden Markov Model to disambiguate type-conflation hypotheses was used here to provide a generic corpus-independent token-level canonicalization function.[8] For the current experiments, an "intensional" phonetic equivalence cascade with an infinite weighted target lexicon derived from the TAGH morphology transducer (Geyken and Hanneforth, 2006) was used in place of a finite target lexicon.

3 Method

We used the relevance relation derived from the test corpus to compare three different canonicalization techniques: the generic corpus-independent canonicalizer (HMM) from section 2.4, the corpus-based canonicalization function with identity fallback (LEX) from section 2.3, and a hybrid method (HMM+LEX) which canonicalizes known words according to the finite corpus canonicalization function CLEX$(\cdot)$, passing any unknown words to the generic HMM canonicalizer. More precisely, the hybrid method passed all sentences in their entirety through the HMM canonicalizer, but each token instantiating a known word type w was assigned a singleton set of canonicalization hypotheses containing only the unique lexicon entry CLEX(w) for that type, effectively restricting the output of the model for known words while still allowing context-dependent disambiguation of unknown words, and even allowing the model to make direct use of the corpus-based canonicalizations in its computation of path probabilities.

3.1 Evaluation Measures

The various canonicalization methods were evaluated using the ground-truth test corpus from Section 2.1 to simulate an information retrieval task. Formally, let $G = \langle g_1, \ldots, g_{n_G} \rangle$ represent the test corpus,[9] where each token g_i is a pair $\langle w_i, \tilde{w}_i \rangle$ such that $\tilde{w}_i$ is the (modern) canonical cognate for the (historical) word w_i. Let $C = \{\text{HMM}, \text{LEX}, \text{HMM+LEX}\}$ be the finite set of canonicalizers under consideration. Then, for each test corpus token g_i and for each canonicalizer $c \in C$, let $[\![\tilde{w}_i]\!]_c$ represent the unique canonical form returned by the canonicalizer c for the token g_i. Let $Q = \bigcup_{i=1}^{n_G} \{\tilde{w}_i\}$ be the set of all canonical cognates represented in the corpus, and define for each canonicalizer $c \in C$ and query string $q \in Q$ the sets $\text{relevant}(q), \text{retrieved}_c(q) \subset \mathbb{N}$ of *relevant* and *retrieved* corpus tokens as:

$$\text{relevant}(q) = \{i \in \mathbb{N} : q = \tilde{w}_i\} \quad (3)$$

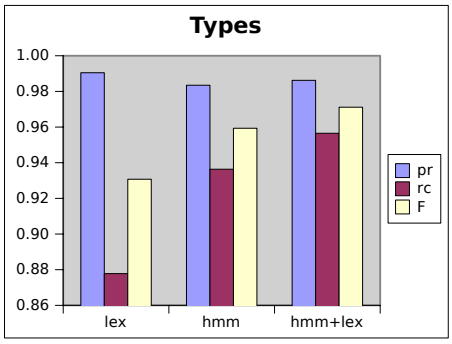
$$\text{retrieved}_c(q) = \{i \in \mathbb{N} : q = [\![w_i]\!]_c\} \quad (4)$$

Token-wise precision ($\text{pr}_{\text{tok},c}$) and recall ($\text{rc}_{\text{tok},c}$) for the canonicalizer c can

[8] A "hypothetical dictionary" in the terminology used by Gotscharek et al. (2009b).

[9] More precisely, only the "word-like" tokens of the test corpus were considered for evaluation purposes, and differences in letter case were ignored.

	Types			Tokens		
	$\mathrm{pr}_{\mathrm{typ}}$	$\mathrm{rc}_{\mathrm{typ}}$	$\mathrm{F}_{\mathrm{typ}}$	$\mathrm{pr}_{\mathrm{tok}}$	$\mathrm{rc}_{\mathrm{tok}}$	$\mathrm{F}_{\mathrm{tok}}$
LEX	**.990**	.878	.931	**.998**	.985	.992
HMM	.983	.936	.959	.996	.985	.991
HMM+LEX	.986	**.957**	**.971**	**.998**	**.993**	**.995**



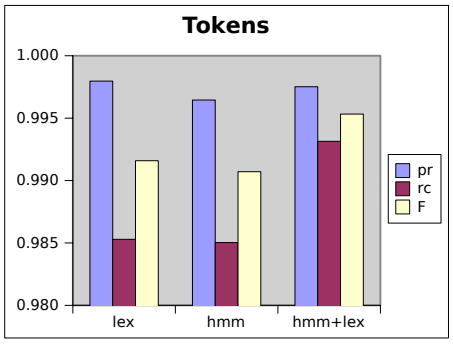


Table 1: Comparison of three canonicalization techniques: a generic Hidden Markov Model canonicalizer (HMM), a corpus-induced exception lexicon (LEX), and a generic canonicalizer supplemented by a corpus-induced lexicon (HMM+LEX). The maximum value in each column appears in boldface type.

then be defined as:

$$\mathrm{pr}_{\mathrm{tok},c} = \frac{\left|\bigcup_{q\in Q} \mathrm{retrieved}_c(q) \cap \mathrm{relevant}(q)\right|}{\left|\bigcup_{q\in Q} \mathrm{retrieved}_c(q)\right|} \tag{5}$$

$$\mathrm{rc}_{\mathrm{tok},c} = \frac{\left|\bigcup_{q\in Q} \mathrm{retrieved}_c(q) \cap \mathrm{relevant}(q)\right|}{\left|\bigcup_{q\in Q} \mathrm{relevant}(q)\right|} \tag{6}$$

Type-wise measures $\mathrm{pr}_{\mathrm{typ},c}$ and $\mathrm{rc}_{\mathrm{typ},c}$ are defined analogously, by mapping the token index sets of Equations (3) and (4) to corpus types before applying Equations (5) and (6). We use the unweighted harmonic precision-recall average F (van Rijsbergen, 1979) as a composite measure for both type- and token-wise evaluation modes:

$$\mathrm{F}(pr, rc) = \frac{2 \cdot pr \cdot rc}{pr + rc} \tag{7}$$

4 Results & Discussion

Type- and token-wise precision (pr), recall (rc), and harmonic precision-recall average F for the three canonicalization techniques with respect to the test corpus are given in Table 1. Immediately obvious from the observed data is that while both the HMM and corpus-based methods are quite effective in their own right ($\mathrm{F}_{\mathrm{tok}} > .99$ in both cases), the best performance across the board is achieved by the hybrid method HMM+LEX, as anticipated in light of the data from Gotscharek et al. (2009b). The data also show a clear discrepancy in type-wise recall between

the LEX and HMM methods. This effect can be attributed to insufficient data in the training corpus: the finite corpus-based canonicalization lexicon itself provided canonicalizations for only 81.7% of test corpus types representing 97.3% of test corpus tokens; the remaining types were handled by the string identity fallback strategy for the LEX condition. Less than half (40.9%) of the unknown word types were correctly canonicalized by the fallback strategy, representing slightly more than half of the unknown tokens (51%). It is worth noting that the recall of the identity fallback strategy was substantially poorer on unknown words than test-corpus globally, where it achieved a type-wise recall of 55.7% and a token-wise recall of 78.5%. This implies that a disproportionately large number of the test corpus types not present in the training corpus were in fact non-trivial historical spelling variants, since valid contemporary forms would be canonicalized correctly by the identity fallback strategy.

Replacing the naïve identity fallback strategy with the HMM canonicalization architecture in the condition HMM+LEX resulted in correct canonicalization for 80.1% of the unknown types representing 77.8% of unknown tokens. This is relatively unsurprising, since the HMM canonicalizer is explicitly designed to deal with previously unseen input types, whereas the corpus-based canonicalization lexicon can only be hoped to correctly canonicalize those types for which training data was available with any reliability. The benefits of combining corpus-based and robust generative techniques were not all one-way however: the HMM canonicalizer also benefited from inclusion of the corpus-based exception lexicon. The hybrid method HMM+LEX incurred 18–31% fewer type-wise errors and 33–53% fewer token-wise errors than the HMM canonicalizer on its own, although these differences are of smaller absolute magnitude compared to the effects on type-wise LEX recall. Differences in this region of the evaluation scale must be viewed with a modicum of skepticism for a test corpus of the current size, since the observed discrepancies result from differences in the canonicalizations of only 511 types (2595 tokens). Nonetheless, we believe that given the quality of our test corpus, the observed recall improvements at least are robust enough to survive replication on a larger scale.

5 Conclusion

We used a simulated information retrieval task over a semi-automatically constructed ground-truth corpus of historical German text to compare the performance of three different canonicalization techniques: a generic dynamic Hidden Markov Model disambiguation cascade, a static type-wise canonicalization lexicon trained from a canonicalized corpus, and a hybrid architecture which uses the generic method to canonicalize only those input words for which no training data was available. The observed results showed that while both the HMM and corpus-based techniques were quite effective on their own, the hybrid technique outperformed both of them in both type- and token-wise F. The most drastic improvements were observed in type-wise recall for the hybrid method with respect to the corpus-based lexicon, assumedly due to data sparsity problems for the corpus-based method from which the HMM method does not suffer as acutely). Substantial improvements were observed in both precision and recall for the hybrid method with respect to the HMM canonicalizer as well, which suggests that these two methods complement one another if both a large canonicalized

training corpus and a robust canonicalization cascade are available.

References

A. Baron and P. Rayson. Automatic standardization of texts containing spelling variation, how much training data do you need? In M. Mahlberg, V. González-Díaz, and C. Smith, editors, *Proceedings of the Corpus Linguistics Conference (CL2009)*, University of Liverpool, UK, 20–23 July 2009. URL `http://ucrel.lancs.ac.uk/publications/cl2009/314_FullPaper.pdf`.

M. Bollmann, F. Petran, and S. Dipper. Rule-based normalization of historical texts. In *Proceedings of Language Technologies for Digital Humanities and Cultural Heritage Workshop*, pages 34–42, Hissar, Bulgaria, 16 September 2011.

S. Dipper and S. Schultz-Balluff. The Anselm Corpus: Methods and perspectives of a parallel aligned corpus. In *Proceedings of the workshop on computational historical linguistics at NoDaLiDa 2013*, NEALT Proceedings Series 18 / Linköping Electronic Conference Proceedings 87, pages 27–42, 2013. URL `http://www.ep.liu.se/ecp_article/index.en.aspx?issue=087;article=003`.

A. Ernst-Gerlach and N. Fuhr. Retrieval in text collections with historic spelling using linguistic and spelling variants. In *Proceedings of the 7th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL '07)*, pages 333–341, New York, 2007. ACM. doi: 10.1145/1255175.1255242.

A. Geyken and T. Hanneforth. TAGH: A complete morphology for German based on weighted finite state automata. In *Finite State Methods and Natural Language Processing, 5th International Workshop, FSMNLP 2005, Revised Papers*, volume 4002 of *Lecture Notes in Computer Science*, pages 55–66. Springer, Berlin, 2006. doi: 10.1007/11780885_7.

A. Gotscharek, A. Neumann, U. Reffle, C. Ringlstetter, and K. U. Schulz. Enabling information retrieval on historical document collections: the role of matching procedures and special lexica. In *Proceedings of The Third Workshop on Analytics for Noisy Unstructured Text Data*, AND '09, pages 69–76. ACM, New York, 2009a. doi: 1568296.1568309.

A. Gotscharek, U. Reffle, C. Ringlstetter, and K. U. Schulz. On lexical resources for digitization of historical documents. In *Proceedings of the 9th ACM symposium on Document Engineering*, DocEng '09, pages 193–200. ACM, New York, 2009b. doi: 1600193.1600236.

A. Hauser, M. Heller, E. Leiss, K. U. Schulz, and C. Wanzeck. Information access to historical documents from the Early New High German period. In *Proceedings of IJCAI-07 Workshop on Analytics for Noisy Unstructured Text Data (AND-07)*, pages 147–154, 2007.

B. Jurish. *Finite-State Canonicalization Techniques for Historical German*. PhD thesis, Universität Potsdam, January 2012. URL `http://opus.kobv.de/ubp/volltexte/2012/5578/`.

B. Jurish, M. Drotschmann, and H. Ast. Constructing a canonicalized corpus of historical German by text alignment. In P. Bennett, M. Durrell, S. Scheible, and R. J. Whitt, editors, *New Methods in Historical Corpora*, volume 3 of *Corpus Linguistics and Interdisciplinary Perspectives on Language (CLIP)*, pages 221–234. Narr, Tübingen, 2013.

S. Kempken, W. Luther, and T. Pilz. Comparison of distance measures for historical spelling variants. In M. Bramer, editor, *Artificial Intelligence in Theory and Practice*, pages 295–304. Springer, Boston, 2006. doi: 10.1007/978-0-387-34747-9_31.

J. Porta, J.-L. Sancho, and J. Gómez. Edit transducers for spelling variation in Old Spanish. In *Proceedings of the workshop on computational historical linguistics at NoDaLiDa 2013*, NEALT Proceedings Series 18 / Linköping Electronic Conference Proceedings 87, pages 70–79, 2013. URL `http://www.ep.liu.se/ecp_article/index.en.aspx?issue=087;article=006`.

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

U. Reffle, A. Gotscharek, C. Ringlstetter, and K. U. Schulz. Successfully detecting and correcting false friends using channel profiles. *Int. J. Doc. Anal. Recognit.*, 12(3):165–174, October 2009. doi: 10.1007/s10032-009-0091-y.

S. Scheible, R. J. Whitt, M. Durrell, and P. Bennett. A gold standard corpus of Early Modern German. In *Proceedings of the 5th Linguistic Annotation Workshop*, pages 124–128, Portland, Oregon, USA, 2011. ACL. URL `http://www.aclweb.org/anthology/W11-0415`.

C. J. van Rijsbergen. *Information Retrieval*. Butterworth-Heinemann, Newton, MA, 1979.