Bryan Jurish / Marko Drotschmann / Henriette Ast

Constructing a Canonicalized Corpus of Historical German by Text Alignment --- DRAFT ---

Abstract

Historical text presents numerous challenges for contemporary natural language processing techniques. In particular, the absence of consistent orthographic conventions in historical text presents difficulties for any system requiring reference to a static lexicon indexed by orthographic form. Canonicalization approaches seek to address these issues by assigning an extant equivalent to each word of the input text and deferring application analysis to these canonical forms. Quantitative evaluation of canonicalization techniques in terms of precision and recall requires reference to a ground-truth corpus in which the canonical form for each corpus token has been manually verified, but such manually annotated corpora are difficult to come by and in general both costly and time-consuming to create. In this paper, we describe a method for bootstrapping a ground-truth canonicalization corpus with minimal manual annotation effort by means of automatic alignment of historical texts with current editions of the same texts, coupled with a two-phase manual review process.

1 Introduction

Virtually all conventional text-based natural language processing techniques 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. Unconventional 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; Gotscharek et al. 2009a, b; Reffle et al. 2009; Jurish 2010, 2011) 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.

A quantitative evaluation of any canonicalization technique in terms of the information retrieval notions of precision and recall requires reference to both *retrieval* and *relevance* relations over corpus target items (tokens or types). In general, a retrieval relation can be defined for any elementary canonicalization function f as the equivalence kernel $\sim_f = f \circ f^{-1} = \{(x,y) : f(x) = f(y)\}$; for a given query, all and only those items are retrieved which share a canonical form with that query. The relevance relation however should be independent of the canonicalization technique chosen, in order to ensure comparability of evaluation results for different canonicalization methods. Further, relevance should be determined as far as possible by manual inspection in order to avoid confounding evaluation results and to ensure that problematic phenomena such as lexical ambiguity are adequately accounted for. Clearly, these desiderata would be fulfilled by a ground-truth corpus of historical text in which the canonical form for each corpus token has been manually verified, thus providing both canonicalization input data (the original corpus text) as well as a relevance relation (the equivalence kernel for the manually determined canonical forms), but such manually annotated corpora are difficult to come by and in general both costly and time-consuming to create.

In this paper, we describe a method for constructing such a ground-truth canonicalized corpus with minimal manual annotation effort using automatic text alignment coupled with a two-phase manual review process. The core intuitions underlying our approach can be summarized as follows:

(1) When they exist, contemporary editions of historical texts already incorporate the desired relevance relation; and

(2) since language change is a comparatively slow process, only a small subset of the relevance relation can be expected to consist of “interesting” non-identity pairs.

N	Text
12405	C. Brentano: *Geschichte vom braven Kasperl und dem schönen Annerl*. Berlin: Vereinsbuchhandlung, 1838.
1865	W. Busch: *Max und Moritz*. München: Braun & Schneider, 1865.
14490	J. W. von Goethe: *Iphigenie auf Tauris*. Leipzig: Göschen, 1787.
42970	J. W. von Goethe: *Wilhelm Meisters Lehrjahre*. Bd. 1. Berlin: Unger, 1795.
43933	J. W. von Goethe: *Wilhelm Meisters Lehrjahre*. Bd. 2. Berlin: Unger, 1795.
45255	J. W. von Goethe: *Wilhelm Meisters Lehrjahre*. Bd. 3. Berlin: Unger, 1795.
63215	J. W. von Goethe: *Wilhelm Meisters Lehrjahre*. Bd. 4. Berlin: Unger, 1796.
24771	J. W. von Goethe: *Torquato Tasso*. Leipzig: Göschen, 1790.
3164	I. Kant: "Beantwortung der Frage: Was ist Aufklärung?" In: *Berlinische Monatsschrift*, 1784, H. 12, S. 481-494.
5925	G. E. Lessing: *Die Erziehung des Menschengeschlechts*. Berlin: Voss, 1780.
30922	F. Schiller: *Kabale und Liebe*. Mannheim: Schwan, 1784.
50697	J. Spyri: *Heidi's Lehr- und Wanderjahre*. Gotha: Perthes, 1880.
9702	T. Storm: *Immensee*. Berlin: Duncker, 1852.

Table 1: Historical source texts used to construct the prototype corpus

Intuition (1) suggests that we can extract the desired relevance relation by aligning a historical text with a contemporary edition of the same text, while intuition (2) can be used to guide the alignment process by attempting to maximize the number of identity alignments.

2 Construction

2.1 Sources

We applied our construction to a prototype corpus of 13 volumes of historical German text published between 1780 and 1880 (Table 1).[1] The text of the historical editions was drawn from the *Deutsches Textarchiv* ("German Text Archive"; Geyken and Klein 2010),[2] encoded according to the Text Encoding Initiative (TEI) P5 Guidelines.[3] Contemporary editions of the selected volumes were provided by the online libraries Project Gutenberg[4] and Zeno.[5]

The raw historical corpus was heuristically tokenized into 417,249 tokens of 30,101 distinct surface types in 20,872 sentences. Of these, 349,541 tokens (84%) of 28,146 distinct surface types (94%) contained only alphabetic and hyphenation characters and were thus considered "word-like".

2.2 Text Alignment

The first phase of the construction process is the heuristic alignment of a historical source text with a contemporary edition of the same text. The contemporary edition or "target text" is assumed to adhere to contemporary orthographic conventions, and the purpose of the alignment phase is to extract a significant portion of the canonicalization relevance relation exhibited by the editorial changes in the target text. Effectively, the alignment of source and target texts should bootstrap a relevance relation based on the linguistic competence of the human editor(s) responsible for the contemporary edition.

Input to the alignment phase were pairs of files representing the raw historical source and contemporary target editions of each corpus text. The optimal alignment itself was computed by GNU

1 The 13-volume prototype corpus represents only a small portion of an ongoing corpus construction project using the methods described here. We use the 13-volume corpus as an example throughout this article because it represents the most thoroughly annotated subset of the corpus at the time of writing.

2 http://deutschestextarchiv.de

3 http://www.tei-c.org/Guidelines/P5/

4 http://www.gutenberg.org

5 http://www.zeno.org

`diff` (Hunt and McIlroy 1976; MacKenzie et al. 2003) under the '`--minimal`' switch,[6] which returns an alignment based on the longest common subsequence (LCS) for the given source and target texts. Since `diff` aligns its argument files line-by-line, both source and target texts were heuristically tokenized into a one-word-per-line format before alignment in order to abstract over differences in formatting. Additionally, since `diff` aligns input lines exclusively on the basis of surface string identity, the conservative transliteration function from Jurish (2010) was applied to the historical text to help account for extinct graphemes.[7]

The initial alignment returned by `diff` is a sequence of *hunks*, where each hunk is either:

- an *identity hunk*, a string of adjacent (transliterated) tokens occurring in both source and target texts;[8]
- a *deletion hunk*, a string of tokens occurring only in the source;
- an *insertion hunk*, a string of tokens occurring only in the target; or
- a *change hunk*, equivalent to a simultaneous deletion and insertion with no intervening identity hunk.

Insertion hunks were ignored during the alignment phase. Identity hunks were marked as valid canonicalizations and copied verbatim to the aligned output file, in accordance with intuition (2) from section 1.[9] For each token in a deletion hunk, a corresponding token with an empty canonical form was included in the output file, where it was marked as unaligned and therefore requiring further manual attention.

In order to extract potential canonicalizations beyond those for which strict identity of (transliterated) string forms applies, each change hunk was inspected more closely. First, each change hunk was tested for identity modulo token boundaries in order to accommodate common concatenative morphological phenomena such as exhibited by the canonicalizations *zwei und vierzig* ↦ *zweiundvierzig* ("forty-two") or *allzuweit* ↦ *allzu weit* ("all too far"). Change hunks which were entirely accounted for by identity of concatenated transliterated forms were flagged as such and accepted with the corresponding substring identities into the output file.[10]

Each remaining change hunk was passed to an additional fine-grained alignment subroutine using the Wagner-Fischer (1974) algorithm for computing string edit-distance (Levenshtein 1966) to align the deletion and insertion portions of the change hunk on the character level. The resulting character-wise alignment was used to determine the most likely target word in the insertion portion for each source word in the deletion portion, using a scoring function based on the empirical probability of a (case-insensitive) match operation per source token character. Word alignments thus extracted from change hunks were copied as candidate canonicalizations to the output file, but flagged as non-identity alignments in need of further attention.

The final output of the text alignment phase for each pair of source and target files was a single XML file containing one token for each token of the source text. Each output token was assigned attributes for both the original source string (before transliteration) and the aligned target word string (if any), in addition to the administrative flags described above.

---

6 The GNU `diff` manual glosses this option as "Try hard to find a smaller set of changes".

7 In particular, the transliterator was responsible for mapping the historical long 'ſ' to a conventional round 's', as well as superscript 'e' to the conventional *Umlaut* diacritic '¨', as in the transliteration *Abſtände* ↦ *Abstände* ("distances").

8 Strictly speaking, `diff` does not output identity hunks at all. The location and content of identity hunks can however easily be reconstructed from the line addresses associated with the adjacent non-identity hunks.

9 Violations of intuition (1) arising e.g. from use of non-standard orthography in both source and target texts will therefore result in spurious identity canonicalizations during this phase. Minor violations of intuition (2) such as might arise from a highly deviant source text will only increase the required manual annotation effort, while major violations of intuition (2) stemming e.g. from major grammatical discrepancies between source and target texts will cause the construction as given to fail, since an adequate treatment of these would require more sophisticated alignment techniques than a simple LCS-based method can provide.

10 Such treatment is justified to the extent one assumes (as we do) that despite diachronic changes in word boundary placement, the historical forms remain compositionally grammatical.

2.3 Manual Annotation

The automatic text alignment procedure discovered candidate canonicalizations for over 98% of word-like input tokens. Of these, over 77% were literal identity pairs and over 94% were identical after transliteration. Even accepting the validity of the transliterated-identity alignments,[11] we are still left with 23,205 word-like tokens requiring human attention. While this represents a substantial reduction in required manual annotation effort with respect to the full 349,541-word corpus, the situation can be further improved by splitting the manual annotation process into *type-wise* and *token-wise* phases.

2.3.1 Type-wise Confirmation

Natural language text is known to obey Heaps' Law (Heaps 1978; Baeza-Yates and Navarro 2000), a correlate of the more widely known Zipf rank-frequency correlation (Zipf 1949; van Leijenhorst and van der Weide 2005; Lü et al. 2010). The former empirical law states that there is a log-linear correlation between vocabulary size in types and corpus size in tokens. In the current context, Heaps' Law implies that a comparatively small number of alignment word-pair types can be expected to account for a large portion of the candidate tokens discovered by the alignment phase. Moreover, an incremental corpus construction process can be expected to encounter ever fewer novel candidate alignment types as the number of aligned tokens increases.

The next step toward minimizing the manual annotation effort required by our corpus construction is therefore a type-wise manual confirmation phase. In this phase, a human annotator is presented with a series of (*source* ↦ *target*) word-pair types representing candidate canonicalizations discovered by the alignment phase, and is asked to decide for each presented type whether or not the given *target* word is to be considered a valid equivalent contemporary form for the given *source*. Each alignment type is presented at most once,[12] and the annotator's decisions are saved to a persistent database and re-used for each newly aligned text, so that the effort required for type-wise confirmation decreases as the corpus grows.

Since each decision regarding the validity of an alignment type is final, achieving our goal of a high-quality output corpus suitable for use as a ground-truth relevance relation means that great care must be taken to ensure that the decisions made at this stage are based on conservative criteria. As an example, consider the canonicalization candidate (*über* ↦ *aber:* "over" ↦ "but"): the heuristics used by the text alignment phase can easily suggest the alignment of these two types by virtue of their common string suffix *-ber*, but given the high frequencies of the closed-class words involved, the potential for spurious alignments of the corresponding types is very great indeed.

For this reason, type-wise annotators were instructed to accept only those proposed alignments of which they were certain. Additional guidelines given to the type-wise annotators included the instructions:

(1) In general, accept changes in letter case and common historical allographs; e.g. accept any of the source forms *Bei, Bey, bei,* or *bey* for the target word *bei* ("by").

(2) Reject alignments involving a change in lexical root, part-of-speech, or morphosyntactic features; e.g. (*das* ↦ *dass:* "the" ↦ "that"), (*Ewigkeiten* ↦ *Ewigkeit:* "eternities" ↦ "eternity").

(3) Reject alignments of suspected graphical origin such as printing-, OCR-, or transcription errors; e.g. (*Gerechtigkelt* ↦ *Gerechtigkeit:* "justice"), (*zuükhalten* ↦ *zurückhalten:* "hold back").

(4) Reject alignments in which the proposed target is itself archaic or extinct; e.g. (*danach* ↦ *darnach:* "afterwards"), (*Licht* ↦ *Lichte:* "light") – the respective inverse alignments would however be acceptable.

(5) Reject alignments whose source components are surface-identical to non-equivalent

11 Note that automatic alignment with a contemporary text should serve to minimize any bias introduced by the transliteration function, since the contemporary target text provides independent evidence for any transliterations which are accepted by this heuristic.

12 Identity alignments, identity-of-transliteration alignments, and unaligned source words are not presented at this stage.

contemporary words. This criterion applies chiefly to ambiguities involving the archaic dative *-e* suffix and contemporary plurals; e.g. (*Orte* ↦ *Ort:* "place(s)"), (*Lande* ↦ *Land:* "land(s)").

(6) Reject alignments of proper names which involve any graphematic changes beyond transliteration of extinct characters, e.g. (*Franciska* ↦ *Franziska*) and (*Oehi* ↦ *Öhi*), but (*Goṫhe* ↦ *Goethe*) is allowed.

For the prototype corpus described in section 2.1, the 23,205 unconfirmed token alignments were reduced to a set of 7,166 alignment pair types of which only 5,780 elements representing 17,839 tokens corresponded to successful alignments arising from change hunks whose source and target components were not surface-identical modulo transliteration. Of these, 4,483 alignment types (77%) representing 16,083 tokens (90%) were accepted in the type-wise confirmation phase, thus eliminating over 69% of the remaining uncanonicalized tokens by manually inspecting less than one quarter of the available unconfirmed items.

The annotation effort required for type-wise confirmation was estimated by explicitly measuring the time needed for confirmation of a random sample of 100 corpus types. Annotation of the sample proceeded at an average confirmation rate of 3.95 seconds per pair, corresponding to a projected total annotation time of about 6.3 hours for the entire corpus. In terms of the original input corpus size, the type-wise confirmation phase proceeded at an estimated rate of over 15 words per second, so the corpus construction up to and including the type-wise confirmation phase does indeed display a very high throughput.

2.3.2 Token-wise Review

Although the combination of automatic text alignment and type-wise manual confirmation is able to provide canonicalizations for the vast majority of input tokens (ca. 98%) with only very little manual annotation effort, a small fraction of input tokens do remain unaccounted for by these techniques. These as-yet uncanonicalized words however are likely to be of particular interest for diachronic corpus-based studies since they include those canonicalization patterns which cannot be reduced to simple string identities or common "run-of-the-mill" allography relations, as well as those which involve ambiguities with valid contemporary forms. In order to achieve a more accurate model of the canonicalization relevance relation, we therefore introduced an additional manual review phase for direct annotation of canonical cognates for as-yet uncanonicalized word-like tokens in sentential context.

Not all of the uncanonicalized tokens returned by the type-wise confirmation phase represent "interesting" non-trivial canonicalization patterns, however. In particular, editorial changes to the original text involving front or back matter, marginalia, speaker designations or stage directions were purged from the corpus by means of a simple XPath filter. Later investigations showed that in some cases – especially in verse collections – chunks of source text spanning multiple pages failed to be automatically aligned at all, usually due to heavy editorial intervention (re-ordering) in the contemporary edition. An additional filter was developed to heuristically detect and remove such unaligned chunks from the corpus using a moving window of n=3 sentences and a minimal alignment threshold of p=75%. It was also noted that the change-hunk-internal heuristic scoring function used in the text alignment phase often failed for short closed-class words such as *der* ("the"), *und* ("and"), or *nicht* ("not"), causing an inordinate inflation of uncanonicalized tokens due to these words' high frequencies. For this reason, a lexicon of 213 high-frequency closed-class items and appropriate canonicalizations was created and applied to the uncanonicalized portion of the corpus.

After pruning and application of the closed-class exception lexicon, the corpus contained a total 405,150 tokens of which 341,798 (84%) were "word-like". Of these, only 3,476 (1.1%) were uncanonicalized. The pruning and closed-class lexicon heuristics together eliminated over half of the remaining uncanonicalized tokens by discarding a mere 2.2% of word-like corpus material as "uninteresting". The pruned corpus was separated into blocks of roughly ten pages which were then randomly sorted and concatenated into a single corpus file for token-wise annotation.

Token-wise annotation itself was performed using a dedicated graphical interface in conjunction with the character-level text-to-image coordinate mapping used by the *Deutsches Textarchiv* online corpus search utility. The annotator was presented with each as-yet uncanonicalized token together with its immediate sentential context in document order, and was asked to assign each such token a lexically equivalent extant cognate. If an automatically discovered alignment was present for the token, it was presented as the default canonical form. The annotator was also asked to provide additional administrative data for each canonicalization if and when appropriate, specifically:

- Whether the token presented is in fact a valid token, or whether it instead represents an error on the part of the heuristic tokenizer.
- Whether the sentence containing the token presented is in fact a valid sentence-like unit, or whether it represents a tokenization error.
- Which of a set of eight pre-defined coarse-grained lexical classes the current token is to be considered an instance of. The set of lexical classes from which the annotator could choose were:

  LEX: a "normal" lexical word; this was the default class assigned if no other class was explicitly chosen.

  JOIN: used together with sentence-level attributes to indicate a string of multiple source tokens to be canonicalized into a single target token. The annotator was additionally asked to map the individual source tokens to compositionally plausible contemporary equivalents where possible.

  SPLIT: used together with an auxiliary target attribute to indicate a single source token to be canonicalized into multiple target tokens. The annotator was additionally asked to map the source token to a single compositionally plausible (e.g. hyphenated) target token where possible.

  FM: foreign-language material.

  GONE: an extinct lexeme without any contemporary cognate.

  GRAPH: an error of graphical origin.

  NE: a proper name, e.g. a person or place name.

  BUG: an encoding error in the source corpus.

Canonical cognates were determined by direct etymological relation of the source root in addition to matching morphosyntactic features. Proper names were canonicalized in accordance with guideline (6) from section 2.3.1. Otherwise, proper names, extinct lexemes, and foreign-language material were treated as their own canonical cognates. Problematic tokens were explicitly marked as such and later subjected to review by an expert.

For efficient annotation of (potentially ambiguous) medium- and high-frequency words, the interface supported batch-level edit operations with optional user selection of target tokens based on a fixed-width context window. As additional visual aids, the annotator was presented with colour-coded "traffic light" status frames for the current source and target forms which indicated whether or not the corresponding word was known to the high-coverage TAGH morphology for contemporary German (Geyken and Hanneforth 2006), and whether or not it satisfied a set of morphological security heuristics (Jurish 2011: A.4). Finally, each edit operation was logged together with its timestamp and the annotator's user-name to a local history list in order to provide basic revision control functionality.

Of the 3,746 uncanonicalized tokens passed into the token-wise review phase, 3,263 (87%) were directly assigned canonical cognates by the original annotator, and the remaining 483 (13%) were flagged and subjected to expert review. 43 word-like tokens and 102 sentences were marked as tokenization errors. Since only complete sentences containing no invalid tokens were included in the final output corpus, tokenization errors resulted in the elimination of 2,827 word-like tokens (<1%) from the corpus. The distribution of the lexical classes assigned to the annotated tokens is given in Table 2.

Class	N	% Edited
LEX	2684	59.22 %
NE	874	19.29 %
JOIN	792	17.48 %
GRAPH	101	2.23 %
SPLIT	72	1.59 %
BUG	40	0.88 %
GONE	8	0.18 %
FM	1	0.02 %

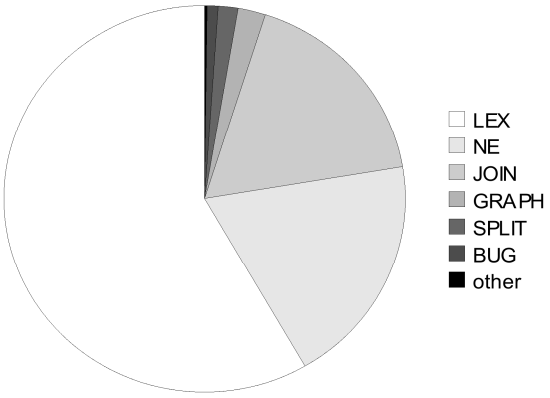

Table 2: Distribution of word classes assigned during token-wise review

Annotation effort was estimated using the intervals between timestamps associated with each manual edit operation. Edit intervals of less than 1 second or greater than 30 minutes were ignored for purposes of the computation. The original annotator applied 5,253 edit operations[13] in editing sessions totaling 55.9 hours. Expert review involved 964 edit operations in sessions totaling 11.7 hours. The manual annotation effort for the token-wise review phase was therefore 67.6 hours, and the total manual annotation effort for the entire corpus was only 74 hours, roughly 2 full-time work weeks. This corresponds to an average throughput of about 1.3 words per second for the whole prototype corpus from start to finish.

3 Conclusion

We have presented a method for constructing a ground-truth corpus of canonicalized historical text with minimal manual annotation effort using automatic text alignment techniques coupled with a two-phase manual review process. Automatic text alignment with a contemporary edition provided an efficient means of discovering non-trivial historical spelling variants, and allowed the subsequent manual review process to draw on the linguistic intuitions of the contemporary edition's editor(s). Manual review was divided into a conservative type-wise confirmation phase and a subsequent token annotation phase in order to leverage the logarithmic growth of vocabulary size for natural language text conforming to Heaps' Law. We estimated an annotation rate of approximately 1.3 words per second for a fully annotated corpus of 13 volumes of 18th-19th century German text.

The 13-volume corpus described above constitutes only the initial portion of an ongoing corpus construction project. We are currently working on incrementally extending the canonicalized corpus using the methods described here based on the historical texts from the *Deutsches Textarchiv*. At the time of writing, an additional 116 volumes containing 5,843,664 tokens in 286,091 sentences have been automatically aligned and passed through the type-wise confirmation phase, requiring manual annotation of an additional 58,644 alignment pair types. Of these, 3,730,781 tokens in 177,390 sentences have also passed through the initial token-wise annotation phase and are awaiting expert review.

Acknowledgements

The work described here was funded by a *Deutsche Forschungsgemeinschaft* (DFG) grant to the project *Deutsches Textarchiv*. Additionally, the authors would like to thank Alexander Geyken, Susanne Haaf, Thomas Hanneforth, Lothar Lemnitzer, and Kai Zimmer for helpful feedback, questions, comments, and assistance with various stages of the work described here.

References

Baeza-Yates, Ricardo / Navarro, Gonzalo (2000): Block-addressing indices for approximate text retrieval. In: Journal of the American Society for Information Science (JASIS), 51(1):69-82.

13 Multiple edit operations were applied to some tokens, and 786 tokens were edited which had already been canonicalized in the type-wise alignment phase. The latter cases were for the most part batch operations which set administrative flags.

Geyken, Alexander / Hanneforth, Thomas (2006): 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. (Lecture Notes in Computer Science 4002). Berlin: Springer, 55-66.

Geyken, Alexander / Klein, Wolfgang (2010): Deutsches Textarchiv. In: Jahrbuch 2009 der Berlin-Brandenburgische Akademie der Wissenschaften. Berlin: Akademie Verlag, 320-323.

Gotscharek, Annette / Neumann, Andreas / Reffle, Ulrich / Ringlstetter, Christoph / Schulz, Klaus U. (2009a): Enabling information retrieval on historical document collections: the role of matching procedures and special lexica. In: Proceedings of AND '09. New York: ACM, 69-76.

Gotscharek, Annette / Reffle, Ulrich / Ringlstetter, Christoph / Schulz, Klaus U. (2009b): On lexical resources for digitization of historical documents. In: Proceedings of DocEng '09. New York: ACM, 193-200.

Heaps, Harold S. (1978): Information Retrieval: Computational and Theoretical Aspects. Orlando: Academic Press.

Hunt, James W. / McIlroy, M. Douglas (1976): An algorithm for differential file comparison. Computing Science Technical Report 41. Bell Laboratories, June 1976.

Jurish, Bryan (2010): More than words: Using token context to improve canonicalization of historical German. In: Journal for Language Technology and Computational Linguistics, 25(1):23-40.

Jurish, Bryan (2011): Finite-State Canonicalization Techniques for Historical German. PhD thesis, Universität Potsdam.

Levenshtein, Vladimir I. (1966): Binary codes capable of correcting deletions, insertions, and reversals. In: Soviet Physics Doklady, 10:707-710.

Lü, Linyuan / Zhang, Zi-Ke / Zhou, Tao (2010): Zipf's Law leads to Heaps' Law: Analyzing their relation in finite-size systems. In: *PLoS ONE*, 5(12):e14139.

MacKenzie, David / Eggert, Paul / Stallman, Richard (2003): Comparing and Merging Files with GNU Diff and Patch. Bristol: Network Theory Ltd.

Rayson, Paul / Archer, Dawn / Smith, Nicholas (2005): 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, July 14-17.

Reffle, Ulrich / Gotscharek, Annette / Ringlstetter, Christoph / Schulz, Klaus U. (2009): Successfully detecting and correcting false friends using channel profiles. In: International Journal on Document Analysis and Recognition, 12:165-174.

van Leijenhorst, Dirk C. / van der Weide, Theo P. (2005): A formal derivation of Heaps' Law. In: Information Sciences, 170(2-4):263-272.

van Rijsbergen, Cornelis J. (1979): Information Retrieval. Newton, MA: Butterworth-Heinemann.

Wagner, Robert A. / Fischer, Michael J. (1974): The string-to-string correction problem. In: Journal of the ACM, 21(1):168-173.

Zipf, George K. (1949): Human Behaviour and the Principle of Least-Effort. Cambridge, MA: Addison-Wesley.