Querying the *Deutsches Textarchiv*

Bryan Jurish Christian Thomas Frank Wiegand

Deutsches Textarchiv · Berlin-Brandenburgische Akademie der Wissenschaften
Jägerstrasse 22/23 · 10117 Berlin · Germany
`jurish|thomas|wiegand@bbaw.de`

Abstract

Historical document collections present unique challenges for information retrieval. In particular, the absence of consistent orthographic conventions in historical text presents difficulties for conventional search architectures which typically rely on a static inverted index keyed by orthographic form. Additional steps must therefore be taken in order to improve recall, in particular for single-term bareword queries from non-expert users. This paper describes the query processing architecture currently employed for full-text search of the historical German document collection of the *Deutsches Textarchiv* project.

1 Introduction

Historical document collections present unique challenges for information retrieval. In particular, the absence of consistent orthographic conventions in historical text presents difficulties for any system requiring reference to a static lexicon keyed by orthographic form. Conventional search architectures on the other hand typically rely on a static inverted index [Knu73, BCC10] mapping each actually occurring surface string to a list of its locations, implicitly assuming the source texts adhere to strict orthographic conventions. Since casual or non-expert users cannot be expected to be familiar with the many spelling variants to be found in historical document collections, and since the explicit enumeration of all possible variants can be a time-consuming and error-prone process even for language-historical experts, additional steps must be taken to improve recall [EGF06, HHL$^+$07, GNR$^+$09, Jur12, Efr13].

This paper describes the process architecture for full-text search in the historical German document collection of the *Deutsches Textarchiv* (DTA). Our approach makes use of an extensive corpus preprocessing phase to annotate the source texts with linguistically salient attributes such as "canonical" contemporary form, part-of-speech tag, and lemma. Building on the richly annotated corpus and a document index structure supporting multiple quasi-independent token-level attributes, naïve bareword searches are expanded into equivalence classes of historical spelling variants by a dedicated external expansion server.

The rest of this paper is organized as follows: section 2 describes the historical text corpus indexed by the DTA, section 3 describes the DTA query processing architecture in greater detail, and section 4 contains a conclusion and brief description of work currently in progress.

2 Text Corpora

The *Deutsches Textarchiv* ("German Text Archive")[1], a project funded by the *Deutsche Forschungsgemeinschaft* (DFG, "German Research Foundation") at the Language Research Center of the Berlin-Brandenburg Academy of Sciences and Humanities, provides a core corpus of more than 1300 significant German texts from various disciplines originally published between ca. 1600 and 1900. Due to the project's primary focus on the history of the German language, full-text transcriptions document the original printed works, of which the earliest edition accessible was digitized. The transcriptions were acquired for the most part using the highly accurate double-keying method; optical character recognition (OCR)

In: U. Kruschwitz, F. Hopfgartner and C. Gurrin (eds.): Proceedings of the MindTheGap'14 Workshop, Berlin, Germany, 4-March-2014, published at http://ceur-ws.org

[1] `http://www.deutschestextarchiv.de`

was used for only ca. 200 volumes, together with extensive manual pre-structuring and post-correction phases. The corpus as a whole therefore displays an exceptionally high accuracy not only on the level of transcription, but also on the annotation level.

The DTA core text sources are published via the Internet as digital facsimiles and as XML-annotated transcriptions together with comprehensive bibliographic meta-data. The annotation consistently follows the well-documented DTA "base format" (DTABf),[2] a TEI subset developed for the representation of (historical) written corpora [CLA12]. As of January 2014, the DTA core corpus comprises 1301 digitized volumes (ca. 680M characters, 100M tokens).

In addition to the core corpus, the DTA currently includes 473 high-quality textual resources provided by cooperating projects or curated from existing text collections such as Wikisource and Project Gutenberg.[3] Further additions include the *Polytechnisches Journal* (1820–1931; 370 volumes, 490M characters, 78M tokens).[4] In total, the DTA and its extensions comprise approximately 1.2B characters in 195M tokens. In the context of a DFG-funded project, the existing OCR text of the journal *Die Grenzboten* (1841–1922) is currently being structured according to the DTABf and automatically corrected on the character level.[5] The resulting optimized text base will be integrated as an extension to the DTA corpora as well.

All corpus texts are available in the web based platform for collaborative quality assurance DTAQ. Within DTAQ, transcriptions can be proofread, and misprints, transcription or annotation errors as well as erroneous meta-data can be corrected [Wie13]. The DTA serves as a basis for a reference corpus of the historical New High German language and offers highly relevant primary sources for academic research in various disciplines in the humanities and sciences as well as for legal scholars and economists.

3 Methods

This section describes the process architecture underlying the DTA's full-text search functionality. Section 3.1 briefly describes the preprocessing techniques used to prepare the corpus for indexing, and section 3.2 deals with the index itself. The query expansion strategy used for runtime term conflation is presented in section 3.3, and section 3.4 describes some accessibility-oriented extensions.

3.1 Corpus Preprocessing

In order to provide a powerful and flexible retrieval environment, the raw text corpus was subjected to an extensive automatic preprocessing phase before being passed to the low-level retrieval engine for indexing. In particular, corpus text was automatically tokenized into paragraph-, sentence- and word-like units using the WASTE tokenizer [JW13], extinct historical spelling variants were mapped to "canonical" contemporary forms using a both a finite lexicon of known forms and a robust generative canonicalization cascade within the DTA::CAB framework[6] [Jur13], and the returned canonical forms were passed to conventional software tools for morphological analysis [GH06], part-of-speech tagging [Jur03], lemmatization, and named-entity recognition [DD09].

3.2 Index Structure

The richly annotated corpus data was passed to the free open-source DDC concordance tool[7] [Sok03] for indexing of selected document- and token-level attributes. In addition to document-level bibliographic meta-data fields such as title, author, publication date and genre, DDC also allows each token to be associated with a fixed number of quasi-independent local attributes, Boolean conditions over which may be conjoined in runtime queries. In contrast to many conventional search architectures, the DTA corpus index uses not only a raw text string to represent a corpus token, but also includes the following token-level attributes:

Utf8Token (u) contains the raw token text encoded in UTF-8 [Uni13].

Token (w) contains a deterministic transliteration of the raw token text into that subset of the latin alphabet used in contemporary German orthography. In the case of historical German, deterministic transliteration is especially useful for mapping the long-s character 'ſ' to a conventional round 's' and for mapping superscript 'e' to the conventional *Umlaut* diacritic '¨', as in the transliteration *Abſt*$\overset{e}{a}$*nde* ↦ *Abstände* ("distances"). This attribute was used as the default for literal string-identity searches.

[2] http://www.deutschestextarchiv.de/doku/basisformat

[3] These resources were integrated in the course of a BMBF-funded CLARIN-D "curation project". For documentation on the project and a list of resources integrated cf. http://www.deutschestextarchiv.de/clarin_kupro.

[4] The original project's web page is http://dingler.culture.hu-berlin.de. After DFG-funding expired, the complete text base was integrated into the BBAW corpus infrastructure and a DDC open search wrapper http://kaskade.dwds.de/dingleros was set up for corpus queries.

[5] http://brema.suub.uni-bremen.de/grenzboten

[6] http://www.deutschestextarchiv.de/demo/cab

[7] http://www.ddc-concordance.org

CanonicalToken (v) contains the estimated "canonical" contemporary form for the current token as determined by the corpus preprocessing phase; e.g. *Teil* for the raw text *Theyl* ("part") or *fragte* for the raw text *frug* ("asked").

Pos (p) contains the part-of-speech (POS) tag automatically assigned to the source token by the moot part-of-speech tagger [Jur03] using the STTS tag-set [STT95].

Lemma (l) contains the lemma or "base form" assigned to the source token by the corpus preprocessing phase, taking into account both the POS-tag and the analyses returned by the TAGH morphological analyzer [GH06], if any.

XPath (xpath) contains the "canonical" XPath to the deepest element node containing (the first character of) the current token in the original TEI source document.

Page (page) identifies the source facsimile, for administrative and cross-referencing purposes.

Line (lb) tracks the line number of the source token on the current page, for administrative and cross-referencing purposes.

A traditional inverted index is constructed for each attribute at corpus indexing time. Unlike conventional query interpreters supporting only document-level dependencies however, the DDC runtime query interpreter ensures that dependencies in a given user query are resolved at the token level. For example, the query *(@Böttcher WITH $p=NN)* would retrieve all and only those instances of the literal string *Böttcher* annotated with the part-of-speech tag *NN* indicating a common noun ("cooper"), whereas *(@Böttcher WITH $p=NE)* would retrieve those instances tagged as proper names. A conventional query evaluation architecture on the other hand would only be capable of retrieving those documents containing some instance of the target word (*Böttcher*) and some instance of the target part-of-speech tag (*NN* or *NE*), regardless of whether or not the tag was assigned to the target word, or to some other word in the document.

3.3 Runtime Query Expansion

Despite the rich annotations offered by indexed corpus, the majority of actual searches are in fact single-term "bareword" queries. Of 29,410 total queries between September 2013 and January 2014, 15,977 (54.3%) were single-term bareword searches, 3,302 (11.2%) were phrases composed exclusively of bareword terms, and 9,219 (31.4%) were bareword queries of the 'Lemma' attribute, together accounting for 96.9% of user searches.[8] In order to improve recall for such queries[9] – especially from non-expert users who cannot be expected to be familiar with the great diversity of spelling variants to be found in historical texts – while still retaining the flexibility of the multi-attribute DDC index, we extended the DDC query language to include user-defined *term expansion pipelines* with attribute-dependent defaults for both explicit and implicit runtime term conflation.

In addition to built-in term expanders for e.g. letter-case normalization or legacy rule-based stemming, we introduced a new extendable class of external term expanders accessed via HTTP as well as a class for chains or "pipelines" of multiple expanders. Each expander x receives as input a finite set T of strings (terms)[10] and returns a finite set $x(T)$ of "equivalent" strings, for some expander-dependent conflation relation $\sim_x$. The query interpreter evaluates an expanded query as it would any set-valued query as the Boolean disjunction over all elements of the (expanded) set: $[\![x(T)]\!] = \bigcup_{t \in x(T)} [\![t]\!]$. Prototypically, $\sim_x$ will be a true equivalence relation and $x(T)$ will be a superset of T, so that literal matches to a user query will always be retrieved.

Each token attribute is associated with a default expansion pipeline, so that bareword queries can be assigned equivalence classes in an attribute-dependent manner: it would be counter-productive for example to attempt to analyze XPath attribute values as natural language text, whereas Token attribute values are expected to be historical word-forms and may be analyzed as such. The current DTA corpus index configuration defines the following term expanders, among others:

tolower Letter-case expander generating lowercase variants of its input.

toupper Letter-case expander generating uppercase variants of its input. This is the default expander for the Pos attribute.

case Letter-case expander generating upper-, lower-, and initial-uppercase variants of its input. This is the default expander for the Lemma attribute.

[8] The dominance of simple bareword queries is not surprising, being as it is well attested in the literature on generic web searching, e.g. [JSS00, SWJS01, WM07].

[9] [JA12] reported an improvement in type-wise recall from 55.7% to 95.7% for canonical-form queries *vs.* raw string-identity queries in an artificial retrieval task over a small test corpus of 18th- to 19th-century German text, corresponding to a token-wise recall improvement from 78.5% to 99.3%.

[10] A bareword query is treated as a singleton set for purposes of term expansion.

morphy Legacy rule-based stemming and re-inflection using Morphy [Lez00].

tagh TAGH-based lemmatization and re-inflection [GH06] via external server.

pho Phonetic equivalence via external DTA::CAB server.

rw Rewrite equivalence via external DTA::CAB server.

eqlemma TAGH-based best-lemma match using a pre-compiled index via external DTA::CAB server. This is the default expander for both the `Token` and `Utf8Token` attributes.

Of particular interest are the external CAB-based expanders such as `pho`, `rw`, and `eqlemma`. In order to function efficiently, the associated expansion servers must restrict the strings returned to those actually occurring in the corpus. Since each of the CAB-based expanders are equivalence relations of the form $f_a \circ f_a^{-1}$ for some function f_a on source tokens (e.g. phonetic-form or best-lemma), the bulk of the task can be accomplished during the corpus preprocessing phase by constructing a database mapping the image of the corpus under f_a to the associated surface types; i.e. an extensional inverse map $f_a^* : f_a[W] \to \wp(W) : a \mapsto f_a^{-1}(a) \cap W$ for a source attribute $f_a : W \to A$ from corpus words W to some characteristic set of possible attribute values A. Runtime expansion can then be performed by analyzing each input term t with the function f_a and performing a simple lookup in the extensional database, setting $x_a(T) = [f_a \circ f_a^*](T) = \bigcup_{t \in T} f_a^*(f_a(t))$.

3.4 Accessibility Extensions

On their own, none of the innovations discussed above "challenge the paradigm of information access as being a single-shot search request submitted to a web search engine."[11] On the contrary, the costly corpus preprocessing techniques, the indexing of multiple, partially redundant token attributes, and the use of implicit attribute-dependent default term expansion pipelines can be seen as workarounds for the overwhelming dominance of bareword searches from assumedly non-expert users.

In an attempt to promote user query-language literacy, an attribute-sensitive auto-completion widget was added to the prototype HTML search form.[12] In the absence of a user-specified target attribute, the auto-completion procedure performs a simple prefix search of the `Lemma` attribute, incorporating the appropriate explicit syntax into the suggestions it returns. Assumedly, this suggestion strategy is largely responsible for the comparatively high ratio of explicit lemma searches (31.4%) we observed.

Additionally, we implemented a simple web-based GUI for visualization, debugging, and fine-tuning of the term expansion process.[13] This so-called "query lizard" allows users not only to see the effects of changes in the expansion pipeline, but also to fine-tune the term sets actually queried by de-selecting undesirable target values such as miscanonicalizations, foreign-language material, etc. Unlike the auto-completion widget, the query lizard does not seem to have acquired a particularly wide user-base: only 321 accesses were observed between September 2013 and January 2014.

4 Conclusion and Outlook

We have described a flexible architecture for full-text search in historical document collections, especially those exhibiting a high degree of spelling variation. By using a corpus preprocessing phase to annotate the source documents with linguistically salient features and incorporating these into the corpus index as quasi-independent token attributes, we were able to implement a query interpreter which robustly interprets naïve bareword queries as equivalence classes of historical spelling variants, while still retaining the full precision of a raw string index.

We are interested in performing a more thorough evaluation of the online term expansion strategy's utility for actual user searches, and in comparing our approach to alternative methods for approximate search in historical document collections, e.g. [Efr13]. We are currently engaged in the development of semantically motivated term expanders and visualizations using both induced distributional semantic models [BDO95, BL09] and the manually constructed lexical network GermaNet [KL02, LK07].

Acknowledgements

The current work was supported by *Deutsche Forschungsgemeinschaft* grant KL 337/12-2. We are grateful to our colleagues Alexander Geyken, Susanne Haaf, Matthias Schulz, and Kai Zimmer, and to this article's anonymous reviewers for their many helpful comments and suggestions.

References

[BCC10] Stefan Büttcher, Charles L. A. Clarke, and Gordon V. Cormack. *Information Retrieval:*

[11] `http://mindthegap2014.dai-labor.de/?page_id=8`

[12] `http://kaskade.dwds.de/dtaos`

[13] `http://kaskade.dwds.de/dtaos/lizard`

Implementing and Evaluating Search Engines. MIT Press, Cambridge, MA, 2010.

[BDO95] Michael W. Berry, Susan T. Dumais, and Gavin W. O'Brien. Using linear algebra for intelligent information retrieval. *SIAM Review*, 37(4):573–595, December 1995.

[BL09] Marco Baroni and Alessandro Lenci. One distributional memory, many semantic spaces. In *Proceedings of the Workshop on Geometrical Models of Natural Language Semantics*, GEMS '09, pages 1–8, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics.

[CLA12] CLARIN-D AP 5. CLARIN-D user guide, version 1.0.1. Technical report, Berlin-Brandenburgische Akademie der Wissenschaften, 19 December 2012.

[DD09] Jörg Didakowski and Marko Drotschmann. Proper noun recognition and classification using weighted finite state transducers. In Jakub Piskorski, Bruce W. Watson, and Anssi Yli-Jyrä, editors, *Proceedings of FSMNLP 2008 (Ispra, Italy, 11-12 September 2008)*, volume 19 of *Frontiers in Artificial Intelligence and Applications*, pages 50–61. IOS Press, 2009.

[Efr13] Miles Efron. Query representation for cross-temporal information retrieval. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, pages 383–392. ACM, 2013.

[EGF06] Andrea Ernst-Gerlach and Norbert Fuhr. Generating search term variants for text collections with historic spellings. In Mounia Lalmas, Andy MacFarlane, Stefan Rüger, Anastasios Tombros, Theodora Tsikrika, and Alexei Yavlinsky, editors, *Advances in Information Retrieval*, volume 3936 of *Lecture Notes in Computer Science*, pages 49–60. Springer, Berlin, 2006.

[GH06] Alexander Geyken and Thomas 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.

[GNR$^+$09] Annette Gotscharek, Andreas Neumann, Ulrich Reffle, Christoph Ringlstetter, and Klaus 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, 2009.

[HHL$^+$07] Andreas Hauser, Markus Heller, Elisabeth Leiss, Klaus U. Schulz, and Christiane 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.

[JA12] Bryan Jurish and Henriette Ast. Using an alignment-based lexicon for canonicalization of historical text. In *Proceedings of the International Conference Historical Corpora 2012*, Frankfurt am Main, Germany, 6th–9th December 2012.

[JSS00] Bernard J. Jansen, Amanda Spink, and Tefko Saracevic. Real life, real users, and real needs: a study and analysis of user queries on the web. *Information Processing & Management*, 36(2):207–227, 2000.

[Jur03] Bryan Jurish. A hybrid approach to part-of-speech tagging. Technical report, Project "Kollokationen im Wörterbuch", Berlin-Brandenburgische Akademie der Wissenschaften, Berlin, 2003.

[Jur12] Bryan Jurish. *Finite-State Canonicalization Techniques for Historical German*. PhD thesis, Universität Potsdam, January 2012.

[Jur13] Bryan Jurish. Canonicalizing the deutsches Textarchiv. In Ingelore Hafemann, editor, *Proceedings of Perspektiven einer corpusbasierten historischen Linguistik und Philologie (Berlin, 12th–13th December 2011)*, volume 4 of *Thesaurus Linguae Aegyptiae*, Berlin, Germany, 2013.

[JW13] Bryan Jurish and Kay-Michael Würzner. Word and sentence tokenization with Hidden Markov Models. *Journal for Language Technology and Computational Linguistics*, 28(2):61–83, 2013.

[KL02] Claudia Kunze and Lothar Lemnitzer. GermaNet representation, visualization, application. In *Proceedings of the 3rd International Language Resources and Evaluation (LREC '02)*, pages 1485–1491, Las Palmas, Canary Islands, 2002.

[Knu73] Donald Knuth. *The Art of Computer Programming. Third Edition*. Addison-Wesley, Reading, MA, 1998 [1973].

[Lez00] Wolfgang Lezius. Morphy – German morphology, part-of-speech tagging and applications. In *Proceedings of the 9th EURALEX International Congress*, pages 619–623, 2000.

[LK07] Lothar Lemnitzer and Claudia Kunze. *Computerlexikographie: Eine Einführung*. Gunter Narr Verlag, Tübingen, 2007.

[Sok03] Alexey Sokirko. A technical overview of DWDS/dialing concordance. Talk delivered at the meeting *Computational linguistics and intellectual technologies*, Protvino, Russia, 2003.

[STT95] Anne Schiller, Simone Teufel, and Christine Thielen. Guidelines fur das Tagging deutscher Textcorpora mit STTS. Technical report, University of Stuttgart, Institut für maschinelle Sprachverarbeitung and University of Tübingen, Seminar für Sprachwissenschaft, 1995.

[SWJS01] Amanda Spink, Dietmar Wolfram, Bernard J. Jansen, and Tefko Saracevic. Searching the web: The public and their queries. *Journal of the American Society for Information Science and Technology*, 52:226–234, 2001.

[Uni13] Unicode Consortium. *The Unicode Standard*. The Unicode Consortium, Mountain View, CA, 2013.

[Wie13] Frank Wiegand. TEI/XML editing for everyone's needs. In *TEI Members Meeting 2013 (poster session)*, Sapienza, Italy, 2nd–5th October 2013.

[WM07] Ryen W. White and Dan Morris. Investigating the querying and browsing behavior of advanced search engine users. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 255–262. ACM, 2007.