Information Access to Historical Documents from the Early New High German Period

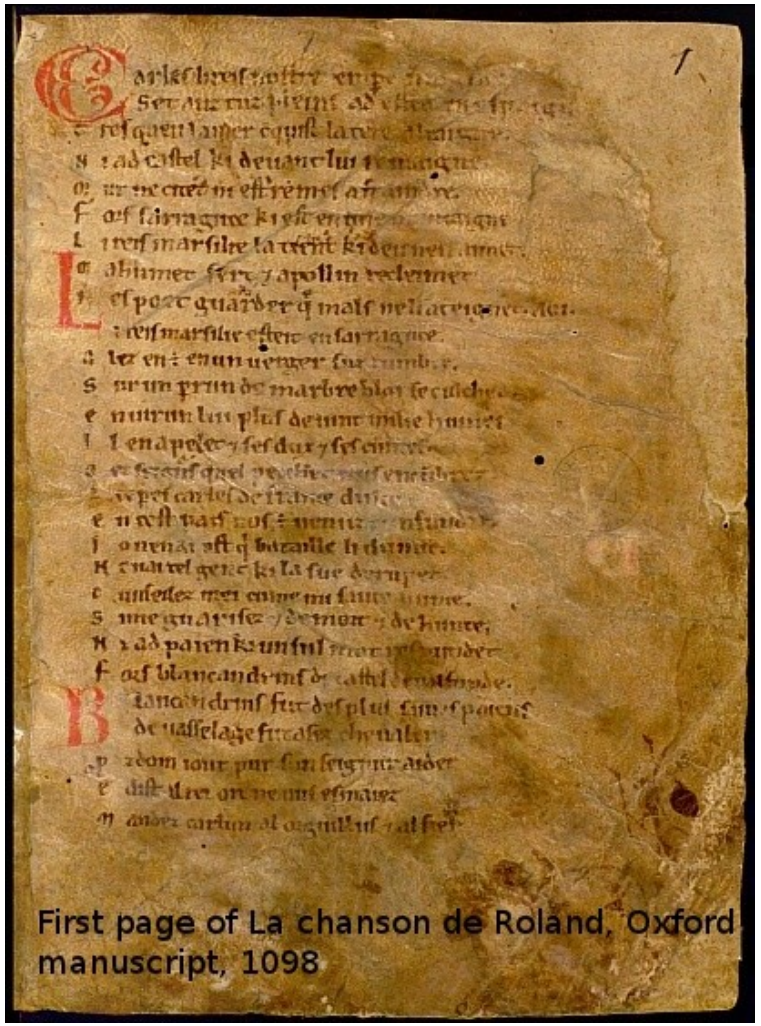
Andreas Hauser, Markus Heller, Elisabeth Leiss, Klaus U. Schulz, Chistiane Wanzeck

Mariona Coll Ardanuy
M.Sc. Language Science & Technology

Overview

- Cultural Heritage
  - Hidden in books and documents
  - The promises of digitalization
- Historical language change: German and levels
- Looking for a solution
- The authors' own work

Cultural Heritage



First page of La chanson de Roland, Oxford manuscript, 1098

«A huge part of the world-wide cultural heritage is hidden in historical books and documents»

- Linguistics: document's language
- Paleography: external and non-textual properties of the source
- History: source's contents

General public!

The promises of digitization

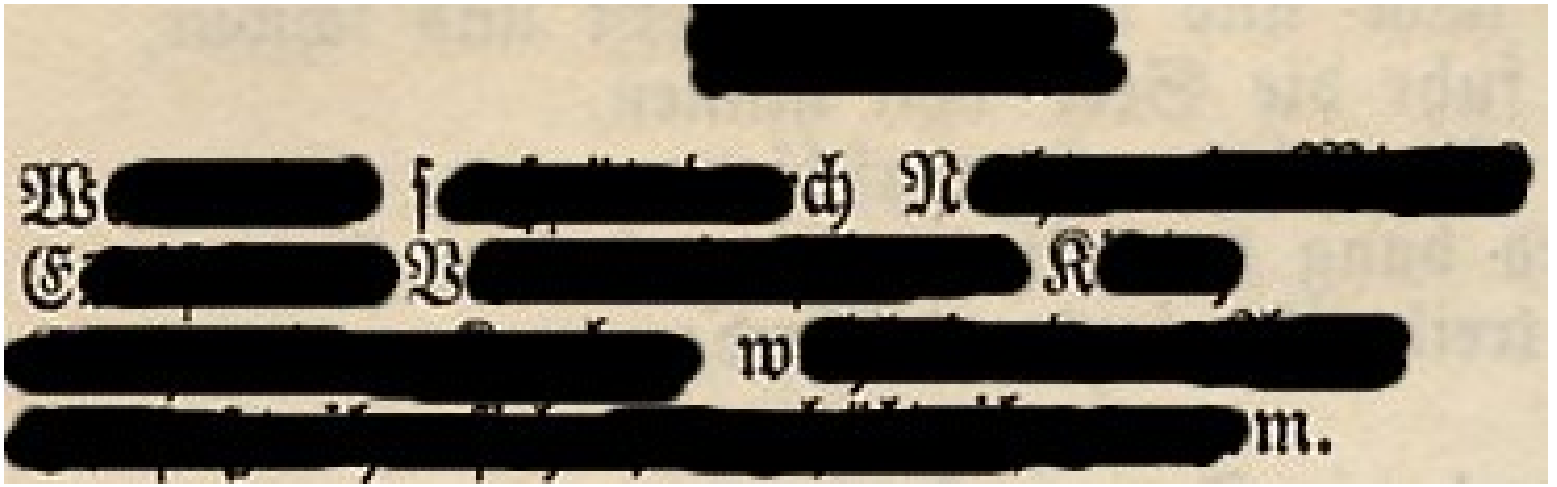
- Three degrees of digitization:
  - Scanned image
  - Textual representation from transcription
  - Structured and possibly annotated version of the textual representation, often in XML
- Assumption: a digitized version of a document will allow Information Retrieval, Text Mining and other NLP technologies

The (false) promises of digitization

- Unstandardized language
  - Diachronic variation
  - Topographic variation
  - Languages without normalized spelling
- Documents with physical ruin
- Difficult OCR for Gothic print

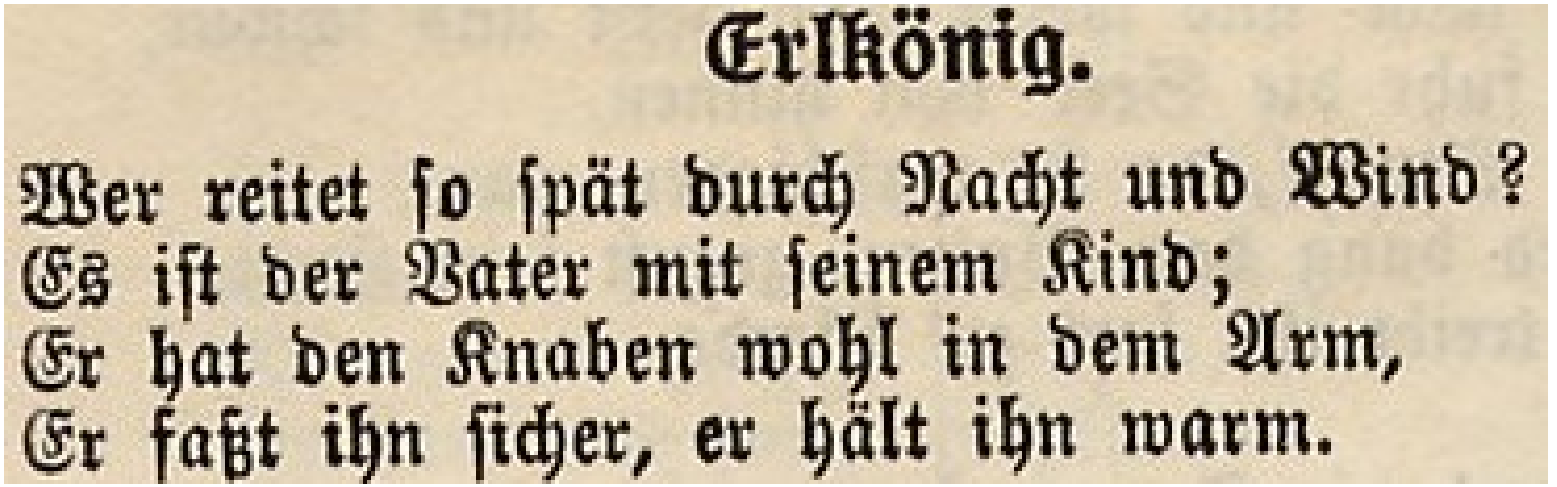
The (false) promises of digitization

- Unstandardized language
  - Diachronic variation
  - Topographic variation
  - Languages without normalized spelling
- Documents with physical ruin
- Difficult OCR for Gothic print



The (false) promises of digitization

- Unstandardized language
  - Diachronic variation
  - Topographic variation
  - Languages without normalized spelling
- Documents with physical ruin
- Difficult OCR for Gothic print



Erlkönig.

Wer reitet ſo ſpät durch Nacht und Wind?
Es iſt der Vater mit ſeinem Kind;
Er hat den Knaben wohl in dem Arm,
Er faßt ihn ſicher, er hält ihn warm.

Historical Language Change: German

- Old High German 8th Century until approx. 1100
- Middle High German 1100-1350
- Early New High German 1350-1600
- New High German 1600-present

Old High German

Fater unseer, thu pist in himile,
uuihi namun dinan,
qhueme rihhi diin,
uuerde uuillo diin,
so in himile sosa in erdu.
(St Gall Paternoster, 8th Century)

Vater unser, der Du bist im Himmel.
Geheiliget werde Dein Name.
Zu uns komme Dein Reich.
Dein Wille geschehe
wie im Himmel also auch auf Erden.
(Early Old Catholic Church Version, 1950)

Middle High German

Uns ist in alten mæren wunders vil geseit
von helden lobebæren, von grôzer arebeit,
von freuden, hôchgezîten, von weinen und von klagen,
von küener recken strîten muget ir nu wunder hœren sagen
(*The Song of the Nibelungs*, 13th Century)

Uns wurde in alten Erzählungen viel Wundersames gesagt
von ruhmreichen Helden, von großem Leid,
von Freuden, Festen, von Weinen und von Klagen,
vom Kampf kühner Recken sollt ihr nun Wunder hören sagen
(*The Song of the Nibelungs*, modern translation)

Early New High German

Unser Vater jnn den himel.
Dein name werde geheilig̃t.
Dein Reich kom̃e.
Dein wille geschehe
auff erden wie im himel
(*Pater Noster*, Luther Bible, 1534)

Vater unser, der Du bist im Himmel.
Geheiliget werde Dein Name.
Zu uns komme Dein Reich.
Dein Wille geschehe
wie im Himmel also auch auf Erden.
(Early Old Catholic Church Version, 1950)

New High German

J. W. Goethe, F. Schiller, J. and W. Grimm...

- Linguistic variation decreases significantly (Konrad Duden's reform of orthography)

  Wer reitet so spät durch Nacht und Wind?
  Es ist der Vater mit seinem Kind;
  Er hat den Knaben wohl in dem Arm,
  Er faßt ihn sicher, er hält ihn warm.

- High German vs. Low German

Historical Language Change: Recap

- Phonological/graphical:

Grapheme	Variants
<a>	< á, å, ah, aa, ai, ae, a̋ >
<e>	< eh, ee, ei, ey, e̊, ë, ä >
<i>	< j, y, ÿ, ie, iee, i͛ , ij, ye, ih, jh, ieh, yh >
<o>	< oh, ó, oe, oi, oy, oo >
<u>	< ú, ů, ů, v, w, uh, wh, ůh, uy >
<ä>	< a̋, e, a, æ, ae, äh >
<ü>	< ű, u, ů, v, ű, ÿ, y, w, ue, üe, üh, uy>
<ö>	< ő, ó, o, öh, oe, öe, ő̈e, œ >

- Morphological: inflection, compounding, word formation
- Lexical: doublets German and Latin, word losings, borrowings
- Syntactical: regularization, word order, punctuation

Looking for a solution

- Special dictionaries
- Rule-based generative matching
- Matching based on word similarity

The 3 approaches can be combined

Related work

- *Information Retrieval on historical text collections* [Ernst 2006]
- *Rule-based search in databases with nonstandard orthography* [Pilz 2006]
- *Information Retrieval on text collections for languages without fixed orthography* [Strunk 2003]
- *Matching variants, approximate name matching* [Zobel 1995]

Work and resources developed so far

- Workshops, conferences
  - E.g.: “Workshop on Historical Text Mining”, Lancaster Uni. 2006
- Dictionaries for historical language
  - Deutsches Wörterbuch von J. und W. Grimm
  - 4 dictionaries for Middle High German
  - Goethe Wörterbuch
  - Deutsche Rechtswörterbuch
  - Dialectal dictionaries
- Electronic corpora for other languages
  - Helsinki-Corpus → English
  - Frantext → French

The authors' own work

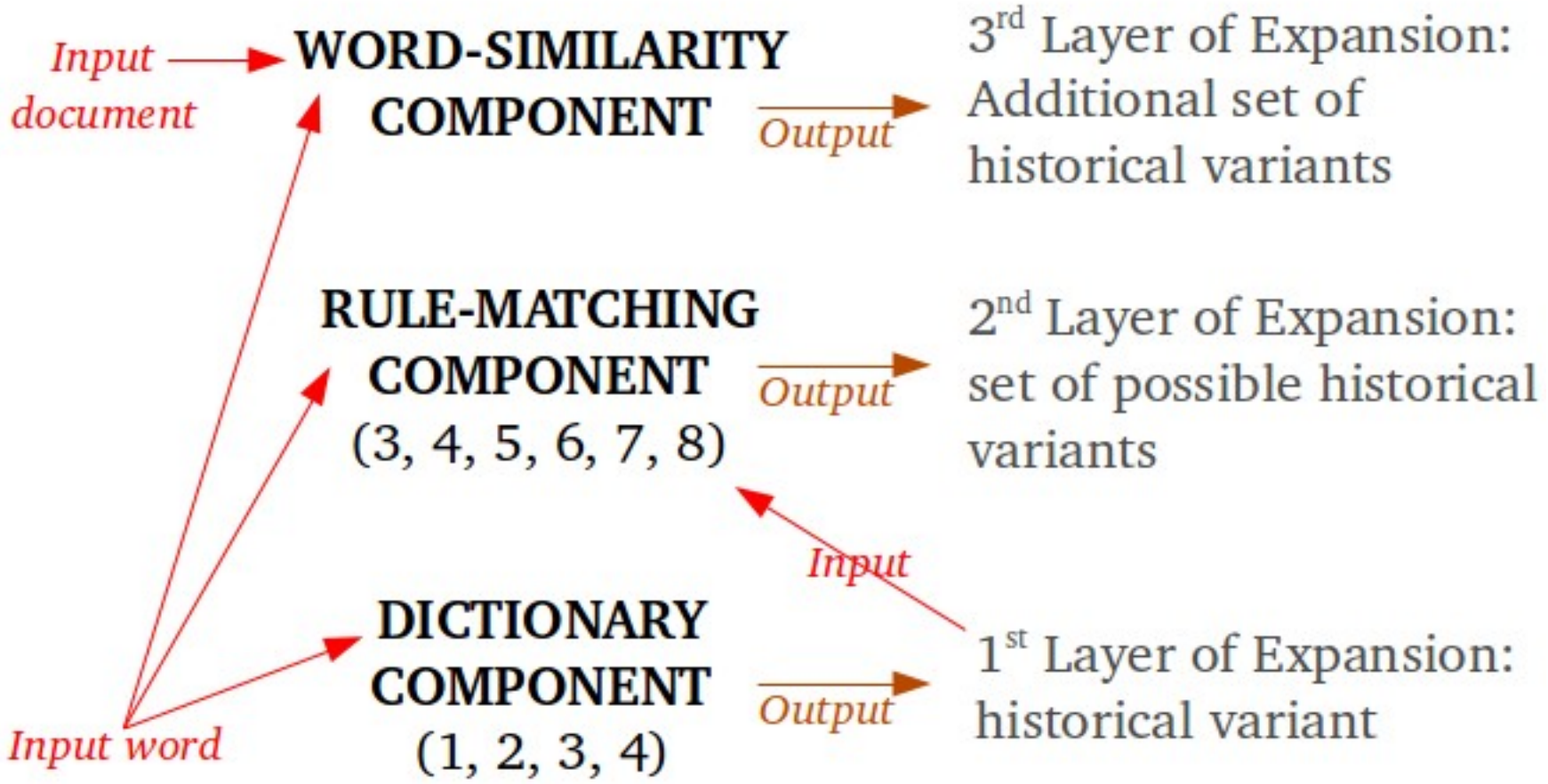
- 14th-17th century prints from ENHG (no manuscripts)
- Corpus of 11 digitized texts from ENHG (130,000 words) out of a selection of 23 texts. 4 texts also have information about category, translation into NHG, underlying ENHG lemma, underlying NHG lemma
- Iterative process:
  - Manually create a small corpus
  - Handle spelling and compound variations
  - Create an usable electronic dictionary
  - Incorporate morphology and syntax
  - Incorporate document structure and meta-information
  - Use all this to improve OCR and digitize more texts

The authors' own work

- Starting point: manually collect correspondences in one text
- Classifying matching problems:
  1. New word form (*handeln* → *marcken*)
  2. Non-normalized latin words (*appellacionn, appellation, appellationn*)
  3. Variations in word splitting (*Winters zeiten* → *Winterzeit*)
  4. Partial new word form (*Grosßteil* → *Mehrteil*)
  5. Variation of prefixes/suffixes (*-chen* → *-lein*)
  6. Typesetting variations (*j* → *i*)
  7. Graphemic-phonetic variations (*Abertheur* → *Abenteuer*)
  8. New character (*fůr* → *für*)

The authors' own work

- Optimizing precision and recall:



The authors' own work

- Dictionary construction and linguistic workbench
  - Text analysis and annotation
  - Linguistic workbench with underlying SQL database
- Design of matching rules
  - Linguistic literature
  - Rules observed from the creation of the dictionary
- Special word distance
  - Modified Levenshtein distance measures (wrt kind of operation – insertion, deletion, substitution – and the particular symbol to be acted on)

Natural interplay between rule-based matching and distance weights

Operations are based on strings instead of characters (i → y only if lein → leyn)

Conclusion

- Ongoing research
- Describes a possible path to improve the matching strategy that helps relating modern language keywords with old variants
- General approach, portable to other languages
- Further work:
  - How to improve the OCR results from historical texts
  - How to deal with electronic historical texts annotated using different XML dialects

Thank you for your attention

References

- Andrea Ernst-Gerlach and Norbert Fuhr. “Generating search term variants for text collections with historic spellings”. In *28th European Conference on Information Retrieval Research (ECIR 2006)*, 2006
- Andreas Hauser, Markus Heller, Elisabeth Leiss, Klaus U. Schulz, Christiane Wanzeck. “Information Access to Historical Documents from the Early New High German Period”. In: L. Burnard, M. Dobreva, N. Fuhr, A. Lüdeling (eds): *Digital Historical Corpora - Architecture, Annotation, and Retrieval*. Dagstuhl Seminar Proceedings, 2007
- Thomas Pilz, Wolfram Luther, Norbert Fuhr, and Ulrich Ammon. “Rule-based search in text databases with nonstandard orthography”. *Literary and Linguistic Computing*, 21(2):179–186, 2006
- Jan Strunk. *Information retrieval for languages that lack a fixed orthography*. Technical report, Linguistics Department, Stanford University, 2003
- Justin Zobel and Philip Dart. *Finding approximate matches in large lexicons. Software–Practice and Experience*, 25(3):331–345, 1995