(Semi-)automatic retrieval of data from historical corpora: Chances and challenges

Marianne Hundt, Melanie Röthlisberger, Gerold Schneider & Eva Zehentner

Aims

- bring together researchers working at the interface between historical linguistics and computational linguistics
- focus on particular issues that arise when retrieving data automatically or semi-automatically from historical databases (of Germanic languages)

Historical corpus linguistics

- similar developments as in corpus-based research on present-day languages:
  - small reference corpora → larger databases
  - text-only → richly annotated resources

Historical corpus linguistics: Challenges

- limited availability of texts: 'bad data' problem (cf. e.g. Labov 1994: 11)
  - corpus representativeness/ balancedness

- highly variable orthography: no straightforward application of tools developed for more uniform data to historical data
  - smaller corpora → grammatical annotation through manual annotation or post-editing
  - manual annotation for larger corpora: tedious/ impractical → development of pre-processing tools like spelling normalisation (Baron & Rayson 2008) and lemmatisation (Burns 2013) to enable automatic tagging/ parsing

Historical corpus linguistics: Challenges

- differently annotated resources (*Penn Treebank, Penn Parsed Corpora, Universal Dependency Treebanks*)
- different parsing tools (e.g. Schneider 2012)
- different retrieval strategies required
- comparisons across corpora difficult
  - lack of use and adaptation of wide range of syntactic parsers to historical texts (e.g. Schneider 2008 for English, Sennrich et al. 2009 for German, van Noord 2006 for Dutch, Alberti et al. 2017 for Universal Dependency parsing)

Main questions

Mapping of different annotation schemes

- **different annotation schemes**: e.g. Penn Parsed corpora vs Penn Treebank (Marcus et al. 1993) vs. dependency-parsed corpora (e.g. ARCHER) → Up to which point, and for which phenomena can the data be compared?
    - *automatic parsers*: different annotations schemes, particularly across different languages and/or periods → What can be mapped easily, and where do we need additional manual decisions in the mapping? Can probabilistic mappers help us, or do they just extend the issue?

Main questions

Mapping of different annotation schemes

- **language specific annotation**: increase in comparability through use of highly underspecified, coarse sets of tags (and e.g. dependency labels). → Is this worth the loss of granularity, and are the labels really directly mappable between different languages? Does historical data add further problems?

- **change over time**: annotation of certain items differs between different periods in Penn-Parsed corpora (indicating e.g. grammaticalisation processes) → How is change dealt with or reflected in annotation? Is this problematic, and what can we do to make sure it does not affect retrieval?

Main questions

Evaluation of bottom-up approaches to data retrieval for language change

- **data-driven approaches to language change**: increase in recall through facilitated detection of new patterns (e.g. Hilpert and Gries 2016) → But can we also detect changes in rare constructions, and how much insight does this really add?
- **fine-grained annotations**: How much benefit is there in using such schemes for exploratory research? Could we get the same (or more representative) results with automatically parsed data/ POS tags only?

Main questions

Evaluation of bottom-up approaches to data retrieval for language change

- **query tools**: very elaborate, detailed queries to extract relevant data (e.g. Stanford Tregex, Levy and Andrew 2006) → Do the benefits outweigh the problematic issues? Where can we draw the line between too generic and too specific queries?
- **role of ambiguity in language change** → Is there a correlation between ambiguity of a language model (e.g. low tagger and parser confidence) or human annotators (low inter-annotator agreement) and change?

Main questions

Issues of precision and recall in historical corpora

- **automatically annotated data**: lower precision/ recall for most phenomena in historical corpora → How much does this affect results? (e.g. increase of a specific lexical item may be due to low recall caused by unidentified spelling variants in earlier texts)
  - often smaller effect on precision than recall → Up to which point are we ready to infer conclusions from annotation which is largely correct but only annotates prototypical cases?
  - typically more errors, but benefit of allowing us to deal with almost unlimited amounts of data → Up to which point can sheer size compensate and even overcompensate the errors?
- **significance testing**: assumes no errors or homogenously spread noise in the data → What can we do when this assumption is clearly violated?

Contributions: Schedule (pre-lunch)

9.00-9.25	*Introduction*
9.30-9.55	Contact-induced change in the diachrony of English and semi-automatic retrieval of data from historical corpora of translated vs non-translated texts (Nikolaos Lavidas, National and Kapodistrian University of Athens)
10.00-11.25	PLENARY & COFFEE BREAK
11.30-11.55	Mapping of lemmatisation annotation to multiple Middle English corpora (Michael Percillier, University of Mannheim)
12.00-12.25	Pattern matching or holistic retrieval: finding bare clefts in unannotated corpus data (Lara Verheyen, Sara Budts, Peter Petre & William Standing, University of Antwerp)
12.30-12.55	Comparing annotation schemes across time: The problem of syntactic mapping (Eva Zehentner, Marianne Hundt, Melanie Röthlisberger & Gerold Schneider, University of Zurich)

Contributions: Schedule (post-lunch)

13.00-14.00	LUNCH
14.00-14.25	Annotation quality assessment and error correction in diachronic corpora: Combining pattern-based and machine learning approaches (Tom S Juzek, Stefan Fisher, Pauline Krielke, Stefania Degaetano-Ortlieb & Elke Teich, Saarland University)
14.30-14.55	When polysemy is what a construction is (all) about: Exploring the use of neural language models for semantic search and classification in (diachronic) corpora (Lauren Fonteyn, University of Leiden)
15.00-15.25	Extraposition and information density in Early New High German corpora (Sophia Voigtmann, Richard Gerbracht, Dietrich Klakow & Augustin Speyer, Saarland University)
15.30	END

Contributions: Languages and time-periods

9.00-9.25	*Introduction*
9.30-9.55	Contact-induced change in the diachrony of English and semi-automatic retrieval of data from historical corpora of translated vs non-translated texts (Nikolaos Lavidas, National and Kapodistrian University of Athens)
10.00-11.25	PLENARY & COFFEE BREAK
11.30-11.55	Mapping of lemmatisation annotation to multiple Middle English corpora (Michael Percillier, University of Mannheim)
12.00-12.25	Pattern matching or holistic retrieval: finding bare clefts in unannotated corpus data (Lara Verheyen, Sara Budts, Peter Petre & William Standing, University of Antwerp)
12.30-12.55	Comparing annotation schemes across time: The problem of syntactic mapping (Eva Zehentner, Marianne Hundt, Melanie Röthlisberger & Gerold Schneider, University of Zurich)

Contributions: Languages and time-periods

9.00-9.25 *Introduction*

9.30-9.55 Contact-induced change in the diachrony of English and semi-automatic retrieval of data from historical corpora of translated vs non-translated texts
(Nikolaos Lavidas, National and Kapodistrian University of Athens)

10.00-11.25 PLENARY & COFFEE BREAK

11.30-11.55 Mapping of lemmatisation annotation to multiple Middle English corpora
(Michael Percillier, University of Mannheim)

12.00-12.25 Pattern matching or holistic retrieval: finding more clefts in unannotated corpus data
(Lara Verheyen, Sara Budts, Peter Petre & William Standing, University of Antwerp)

12.30-12.55 Comparing annotation schemes across time: The problem of syntactic mapping
(Eva Zehentner, Marianne Hundt, Melanie Röthlisberger & Yannick Frehner, University of Zurich)

Old English

Middle English

Early Modern English

Early/Late Modern English

Contributions: Languages and time-periods

13.00-14.00	LUNCH
14.00-14.25	Annotation quality assessment and error correction in diachronic corpora: Combining pattern-based and machine learning approaches (Tom S Juzek, Stefan Fisher, Pauline Krielke, Stefania Degaetano-Ortlieb & Elke Teich, Saarland University)
14.30-14.55	When polysemy is what a construction is (all) about: Exploring the use of neural language models for semantic search and classification in (diachronic) corpora (Lauren Fonteyn, University of Leiden)
15.00-15.25	Extraposition and information density in Early New High German corpora (Sophia Voigtmann, Richard Gerbracht, Dietrich Klakow & Augustin Speyer, Saarland University)
15.30	END

Contributions: Languages and time-periods

13.00-14.00 LUNCH

14.00-14.25 Annotation quality assessment and error correction in diachronic corpora: Combining pattern-based and machine learning approaches (Tom S Juzek, Stefan Fisher, Pauline Krielke, Stefania Degaetano-Ortlieb & Elke Teich, Saarland University)

14.30-14.55 When polysemy is what a construction is (all) about: Exploring the use of neural language models for semantic search and classification in (diachronic) corpora (Lauren Fonteyn, University of Leiden)

15.00-15.25 Extraposition and information density in Early New High German corpora (Sophia Voigtmann, Richard Gerbracht, Dietrich Klakow & Augustin Speyer, Saarland University)

15.30 END

Late Modern English

Modern (American)English

Early New High German

Contributions: Methodological issues

- **annotation (lemmatisation, parsing, spelling normalisation)**: Percillier, Zehentner et al., Juzek et al., Voigtmann et al.
- **corpus coverage/ representativeness/ balancedness**: Percillier
- **document classification**: Lavidas
- **precision & recall**: Zehentner et al., Verheyen et al., Voigtmann et al.
- **error detection/ correction**: Juzek et al.
- **data classification**: Fonteyn, Voigtmann et al.

Contributions: Methodological tools

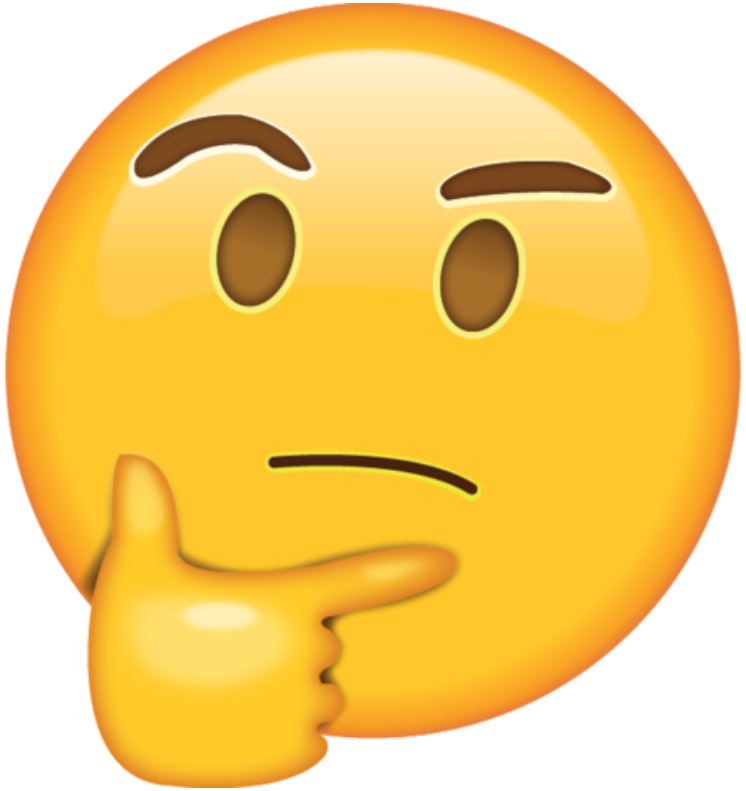
- **machine learning approaches**
  - document classification: Lavidas
  - vector-/ neural network-based: Verheyen et al., Fonteyn (BERT)
  - active learning: Juzek et al., Voigtmann et al.
  - Gradient Tree Boosting/ Conditional Random Fields: Voigtmann et al.
- **different parsers**: Percillier, Zehentner et al., Juzek et al.
- **information-theoretic/information-density measures**: Juzek et al., Voigtmann et al.

Contributions: Linguistic issues

- **morphology/ word formation**: Lavidas
- **syntax**
  - Percillier (*to*-dative)
  - Zehentner et al. (PPs in argument structure)
  - Verheyen et al. (bare clefts)
  - Fonteyn (be + PP constructions)
  - Voigtmann et al. (extraposition)
  - Juzek et al. (grammatical complexity)
- **language contact**
  - Lavidas
    (English - Latin/ French)
  - Percillier
    (English - French)

Publication ?

Discussion ?



References

Alberti, C., D. Andor, I. Bogatyy, M. Collins, D. Gillick, L. Kong, T. Koo, J. Ma, M. Omernick, S. Petrov, C. Thanapirom, Z. Tung, and D. Weiss. 2017. “Syntaxnet models for the CoNLL 2017 shared task”. *arXiv:1703.04929*, 2017.3

Baron, A. and P. Rayson. 2008. 2008. “VARD 2: A tool for dealing with spelling variation in historical corpora”. *Proceedings of the Postgraduate Conference in Corpus Linguistics*.

Burns, Philip. 2013. “MorphAdorner v2: A Java Library for the Morphological Adornment of English Language Texts." Evanston, IL. Northwestern University. (http://morphadorner.northwestern.edu/morphadorner/documentation/citation).

Hilpert, M. and S. Th. Gries 2016. Quantitative approaches to diachronic corpus linguistics. In M. Kytö and P. Pahta (Eds.), *The Cambridge Handbook of English Historical Linguistics*. Cambridge: Cambridge University Press, 36-53.

Kroch, A., and A. Taylor. 2000. The Penn-Helsinki Parsed Corpus of Middle English (PPCME2). Department of Linguistics, University of Pennsylvania. CD-ROM, second edition, release 4 (http://www.ling.upenn.edu/ppche-release-2016/PPCME2-RELEASE-4).

Kroch, A., B. Santorini, and L. Delfs. 2004. The Penn-Helsinki Parsed Corpus of Early Modern English (PPCEME). Department of Linguistics, University of Pennsylvania. CD-ROM, first edition, release 3 (http://www.ling.upenn.edu/ppche-release-2016/PPCEME-RELEASE-3).

Labov, W. 1994. *Principles of Linguistic Change*. Vol. I*: Internal Factors*. Oxford: Blackwell.

Levy, R. and A. Galen. 2006. “Tregex and Tsurgeon: tools for querying and manipulating tree data structures”. *5th International Conference on Language Resources and Evaluation (LREC 2006)*. (https://nlp.stanford.edu/pubs/levy_andrew_lrec2006.pdf).

Marcus, M., B. Santorini, and M. A. Marcinkiewicz. 1993. “Building a Large Annotated Corpus of English: the Penn Treebank”. *Computational Linguistics*, 19, 313-330.

Schneider, G. 2008. *Hybrid Long-Distance Functional Dependency Parsing*. Zurich: University of Zurich PhD thesis.

Schneider, G. 2012. Using semantic resources to improve a syntactic dependency parser. In *LREC* 2012 Conference Workshop "Semantic Relations II", Istanbul, Turkey, 22 May 2012 - 22 May 2012, 67-76.

Sennrich, R., G. Schneider, M. Volk and M. Warin. 2009. ”A New Hybrid Dependency Parser for German,” In Chiarcos, C., de Castilho, R. E., Stede, M. *Von der Form zur Bedeutung: Texte automatisch verarbeiten / From Form to Meaning: Processing Texts Automatically. Proceedings of the Biennial GSCL Conference 2009*. Tübingen: Narr, 115-124.

van Noord, G. 2006. “At last parsing is now operational”. In P. Mertens, C. Fairon, A. Dister and P. Watrin (Eds.) *TALN06. Verbum Ex Machina. Actes de la 13e conference sur le traitement automatique des langues naturelles*, 20-42.