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

Marianne Hundt, Melanie Röthlisberger, Gerold Schneider & Eva Zehentner
(University of Zurich)

Keywords: historical corpus linguistics, computational linguistics, (semi-)automatic retrieval, annotation, precision and recall

This workshop aims to bring together researchers working at the interface between historical linguistics and computational linguistics, zooming in on the particular issues that arise when retrieving data automatically or semi-automatically from historical databases.

Developments in historical corpus linguistics have taken a similar route as in corpus-based research on present-day languages: from the creation of small reference corpora to increasingly larger databases and from text-only to richly annotated resources. However, historical data have always posed particular challenges for the development of corpus resources, their annotation, and their analysis. Corpus representativeness and balancedness, for instance, has been impaired by the limited availability of texts, particularly for the very early stages of written attestation. This is often generally referred to as the 'bad data' problem of historical linguistics (cf. e.g. Labov 1994: 11). Additionally, the highly variable orthography typical of earlier texts has meant that the tools developed for more uniform data cannot be applied in a straightforward manner to historical corpora. In the case of smaller corpora, this has resulted in grammatical annotation through manual annotation or post-editing. For the increasingly larger resources, however, manual annotation is tedious, and researchers have developed tools for pre-processing like spelling normalisation (Baron & Rayson 2008) and lemmatisation (Burns 2013) to enable automatic tagging and parsing. Matters are complicated further by the fact that a range of different annotated resources exist (*Penn Treebank*, *Penn Parsed Corpora, Universal Dependency Treebanks*) and different parsing tools (e.g. Schneider 2012) have been applied to historical corpora, which are likely to require different retrieval strategies, which in turn make comparisons across corpora difficult. While the list of syntactic parsers is large (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), few have been used on, or adapted to historical texts.

The goal of this workshop is to discuss the challenges that (semi-)automatic retrieval of data from historical corpora pose for the study of grammatical change, specifically in Germanic languages. In particular, we invite contributions addressing questions such as (but not limited to) the following:

- Mapping of different annotation schemes:

- Many corpora use different annotation schemes: for example, the Penn Parsed corpora are annotated in a different way than the Penn Treebank (Marcus et al. 1993), or dependency-parsed corpora like ARCHER. If this is the case, up to which point, and for which phenomena can the data be compared?
- How is change over time dealt with or reflected in annotation? For example, the annotation of certain items differs between the Penn-Parsed corpora (covering different periods), indicating e.g. grammaticalisation processes. Is this problematic, and what can we do to make sure it does not affect retrieval?
- Automatic parsers use 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?
- One way to avoid language-specific annotation may be to employ tools like the Universal Dependency label (in progress).1 Using such highly underspecified, coarse sets of tags (and e.g. dependency labels) may indeed increase comparability, but may also lead to a loss of granularity. Is this worth the risk, and are the labels really directly mappable between different languages? Does historical data add further problems?

- <u>Evaluation of bottom-up approaches to data retrieval for language change</u>:
  - Data-driven approaches to language change (e.g. Hilpert and Gries 2016) can detect new patterns, increasing recall. But can we also detect changes in rare constructions, and how much insight does this really add?
  - How much benefit is there in using fine-grained annotations for exploratory research? Could we get the same (or more representative) results with automatically parsed data/ POS tags only?
  - Query tools like Stanford Tregex (Levy and Andrew 2006) allow us to come up with very elaborate, detailed queries to extract relevant data. This can be useful, but may also be problematic: Where can we draw the line between too generic and too specific queries?
  - What is the 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?

- <u>Issues of precision and recall in historical corpora</u>:
  - Precision and recall is lower for most phenomena in historical corpora when using automatically annotated data. How much does this affect results? For example, an increase of a specific lexical item may be due to low recall caused by unidentified spelling variants in earlier texts.

- Precision is often less affected than recall. Up to which point are we ready to infer conclusions from annotation which is largely correct but only annotates prototypical cases?
- While automatic annotation typically involves more errors, it has the 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?

Ultimately, this workshop seeks to provide a platform for researchers working within these subject areas to exchange ideas and to jointly address the challenges (and chances) we are faced with.

References

Alberti, Chris, Daniel Andor, Ivan Bogatyy, Michael Collins, Dan Gillick, Lingpeng Kong, Terry Koo, Ji Ma, Mark Omernick, Slav Petrov, Chayut Thanapirom, Zora Tung, and David Weiss. 2017. "Syntaxnet models for the CoNLL 2017 shared task". *arXiv:1703.04929*, 2017.3

Baron, Alistair and Paul Rayson. 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, Martin and Stefan Th. Gries 2016. Quantitative approaches to diachronic corpus linguistics. In Merja Kytö and Paivi Pahta (Eds.), *The Cambridge Handbook of English Historical Linguistics*. Cambridge: Cambridge University Press, 36-53.

Kroch, Anthony, and Ann 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, Anthony, Beatrice Santorini, and Lauren 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, William. 1994. *Principle of change, internal factors*. Oxford: Blackwell.

Levy, Roger and Galen Andrew. 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, Mitch, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. "Building a Large Annotated Corpus of English: the Penn Treebank". *Computational*

Linguistics, 19, 313-330.
Schneider, Gerold. 2008. *Hybrid Long-Distance Functional Dependency Parsing*. Doctoral Thesis. University of Zurich.
Schneider, Gerold. 2012. “Adapting a parser to Historical English”. In Jukka Tyrkkö, Matti Kilpiö, Terttu Nevalainen & Matti Rissanen (Eds.) *Studies in Variation, Contacts and Change in English, Volume 10: Outposts of Historical Corpus Linguistics: From the Helsinki Corpus to a Proliferation of Resources*. Helsinki: VARIENG.
Sennrich, Rico, Gerold Schneider, Martin Volk, and Martin Warin. 2009. “A New Hybrid Dependency Parser for German”. *Von der Form zur Bedeutung: Texte automatisch verarbeiten / From Form to Meaning: Processing Texts Automatically. Proceedings of the Biennial GSCL Conference 2009*, 115-124.
van Noord, Gertjan. 2006. “At last parsing is now operational”. In Piet Mertens, Cedrick Fairon, Anne Dister & Patrick Watrin (Eds.) *TALN06. Verbum Ex Machina. Actes de la 13e conference sur le traitement automatique des langues naturelles*, 20-42.