Preface

Recent years have seen a surge of interest in the application of computational methods to problems in historical linguistics. To date, much of this work has been based on the application of simple similarity measures to short lists of lexical items or grammatical features for achieving large-scale genetic grouping of languages. While highly publicized and demonstrably useful, such approaches are inherently limited both by the narrow range of linguistic features examined and the low-level processing methods used.

At the same time, language technology for dealing with modern languages has developed apace, with automatic language tools now achieving a degree of accuracy that has enabled both popular online services such as Google translate and the rapid accumulation of linguistically annotated monolingual and multilingual corpora for many languages. Much less has been done on historical texts: there is little commercial interest in these language varieties, there is often limited amounts of data (making purely data-driven annotation approaches unfeasible), and they are less well-behaved than modern print corpora, due to lack of standardization on all linguistic levels, starting with orthography. Digitized older texts also often suffer from OCR errors.

The basic premise of the workshop is that historical linguistics can benefit greatly from having access to historical and diachronic corpora with rich linguistic annotations, but this is a field where researchers have barely scratched the surface of what is possible. However, because of the nature of the material and of the research questions, interesting questions of theory and method arise in connection with this work, which often are relevant to work on modern data as well (e.g., linguistic variation in spoken language or in web genres). The workshop aimed at providing a forum where these questions can be discussed. The target audience of the workshop were researchers – linguists and computational linguists – involved in the creation and utilization of richly annotated historical and diachronic text corpora, in the context of historical-comparative (diachronic, genetic) linguistic research.

We invited papers presenting original research relating to computational historical linguistics, on topics such as:

- theoretical and methodological aspects of automatic annotation for historical linguistic research, e.g.:
  - the influence and significance of annotation errors
  - which kinds of annotation are needed and useful for historical linguistics
  - how to deal with variation and multilinguality
  - annotation transfer between diachronic language stages or between languages
  - issues of standardization, interoperability and data sharing
- innovative user interfaces for computational historical linguistics (including search and visualization solutions)
- design of optimal annotation workflows with manual and automatic components for creating historical and diachronic corpora
- linguistic processing of annotated historical and diachronic corpora for historical linguistic research, e.g.:
  - methods for tracking change in vocabulary and grammar in diachronic corpora
  - grammar extraction and comparison on historical and diachronic treebanks

Six submissions were accepted for presentation at the workshop and inclusion in this proceedings volume after a thorough review procedure and subsequent revision by the authors of the papers. Each submission was reviewed by three (anonymous) members of the program committee:

- Yvonne Adesam (University of Gothenburg)
- David Bamman (Carnegie Mellon University)
- Lars Borin (University of Gothenburg)
- Gerlof Bouma (University of Gothenburg)
- Stefanie Dipper (Ruhr-Universität Bochum)
- Michael Dunn (MPI for Psycholinguistics, Nijmegen)
- Þórhallur Eyþórsson (University of Iceland)
- Markus Forsberg (University of Gothenburg)
- Dag Haug (University of Oslo)
- Seth Kulick (Linguistic Data Consortium)
- Hrafn Loftsson (Reykjavik University)
- Marco Passarotti (Catholic University of the Sacred Heart, Milan)
- Michael Piotrowski (Leibniz Institute of European History)
- Eiríkur Rögnvaldsson (University of Iceland)

The workshop featured two invited speakers: **Seth Kulick** (Linguistic Data Consortium), who gave a presentation with the title *Treebank analysis using derivation trees*, and **Michael Piotrowski** (Leibniz Institute of European History), who talked about *Historical NLP and the Digital Humanities*.

The workshop organizers:
Þórhallur Eyþórsson
Lars Borin
Dag Haug
Eiríkur Rögnvaldsson

WS website: `http://spraakbanken.gu.se/swe/nodalida-chl-ws-2013`

Acknowledgements: Financial support for the organization of the workshop has come in part from NOS HS (the exploratory workshop grant *Diachonic Syntax Corpus (DiaSynCorp)*).