JS Journal of Historical Syntax
Volume 6, Article 4: 1–5, 2022

CREATING ANNOTATED CORPORA FOR HISTORICAL LANGUAGES*

Marieke Meelen
University of Cambridge

David Willis
University of Oxford

This edited volume derives from a workshop 'Creating annotated corpora for historical languages', held at Selwyn College, Cambridge on 26–27 September 2019. The workshop formed part of a wider project 'Developing a Welsh Historical Treebank', funded by the British Academy and Leverhulme Trust, which aimed to develop conventions and procedures that might form the basis for a fully parsed representative corpus of historical Welsh texts. The workshop was designed to share experience of building annotated historical corpora, focusing in particular on the technical issues involved.

Contributions to the workshop focused both on corpus creation (text creation or the difficulties involved in creating parsing conventions, for instance) or on the issues involved in using corpora in linguistic research. There were sessions on current work on text creation and parsing for Welsh in the morning of the first day, while the afternoon was devoted to an introduction to the Parsed Historical Corpus of the Welsh Language. The second day was devoted to sessions discussing technical issues in corpus creation in other languages, such as Old and Middle English, Southern Dutch, Russian and Old Church Slavonic. This special issue of the *Journal of Historical Syntax* presents a selection of work by contributors to the workshop and their collaborators on

* We are grateful to the British Academy and Leverhulme Trust for the Small Research Grant award SRG18R1\181450 'Developing a Welsh Historical Treebank' and to the Arts and Humanities Research Council for funding the UK portion of the AHRC–DFG UK–German collaborative research project in the humanities 'The history of pronominal subjects in the languages of northern Europe' (award no. AH/V00347X/1). Our thanks also go to the anonymous reviewers of this volume and to George Walkden as editor-in-chief of the *Journal of Historical Syntax* for his help with the production of this volume.

topics related to research with historical annotated corpora, from challenges and solutions in the creation of annotated corpora to research based on the output of such project-creation activities.

Historical corpora within the tradition of the Penn Parsed Corpora of Historical English are tagged for both part-of-speech (POS) and for hierarchical syntactic constituency structure in the form of phrase-structure descriptions (PSDs). POS tagging allows easy extraction of particular grammatical elements, while addition of full PSDs to texts in order to construct treebanks provides researchers with reliable access to exhaustive searching of syntactic structures of particular relevance to their research questions. Such constructed corpora have also tended to aim at some form of representativeness, with a similar range of text extracts being used for each century or time period across the historical range of the corpus, so far as this is possible given the range of texts that were composed in and have survived from a given period. These types of annotated historical corpora have become an essential tool for comparative linguists working on morphology, syntax, information structure and patterns of language change. Historical treebanks of this kind have been created for a number of languages, including English, French, Icelandic, Portuguese, and Old Saxon/Low German, and others are in progress.

Two of the articles in this volume deal directly with the challenges of corpus creation, focusing on issues involved in tagging and, above all, parsing texts automatically. **Marieke Meelen and David Willis** look at a selection of issues encountered in providing syntactic structural descriptions (trees) for the texts in the Parsed Historical Corpus of the Welsh Language. Some of these are issues likely to arise in the creation of any parsed historical corpus, such as the extent to which the incorporation of hierarchical structure is a useful addition or an impediment to effective searching, or the best way of representing elements shared between coordinated clauses. Another issue common across historical corpora, and raised also in Eckhoff's contribution to this volume, is how to deal with elements whose grammatical status changes over time. They emphasise the need to adopt conventions that facilitate ease of searching and that can be applied consistently across as extensive a period of language history as possible.

In their article, **Melissa Farasyn, Anne-Sophie Ghyselen, Jacques Van Keymeulen and Anne Breitbarth** report on the construction of a tagged and parsed corpus of the southern Dutch dialects with obvious implications for diachronic research. They test different taggers and parsers (Frog, TreeTagger and Alpino). In line with the CLARIN philosophy, which encourages the sharing, use and sustainability of tools, they choose to adapt these tools, which were initially geared towards standard-language annotation, to make

them suitable for their dialect corpus. They furthermore discuss the challenges they encountered working with spoken, unstandardised language in general and stress the importance of researchers reporting both on the different paths they explored while annotating their data and on the amount and kind of manual postprocessing that is required. Their pilot study is an inspirational precursor of their current, much larger project, in which they aim to annotate over 750 recordings (570 hours) of spoken Southern Dutch dialect material.

Nilo Pedrazzini's paper forms a bridge between the creation of corpora and their use for linguistic research, examining what knowledge can be gained from shallow as opposed to deep parsing of a corpus. It thereby alludes to an issue fundamental to corpus creation, namely the constant tradeoff faced by researchers between time and effort expended on annotation and time spent investigating their central research questions. Put simply, does the time saved by using a richly annotated corpus justify the time spent creating it? Pedrazzini compares results for Early Slavic dative absolutes derived from small, deeply annotated corpora that include details of a range of information-structural features with those derived from large corpora with shallow annotation. He demonstrates how deeply annotated treebanks can be exploited to make informed predictions about a given construction in new texts in larger corpora that lack this type of annotation. Based on his analysis of dative absolutes, he concludes that deep annotation of small treebanks can be useful to test hypotheses, before investing time in deep annotation of large corpora.

The other articles in the volume give examples of the use of corpora for the study of historical linguistic questions. In the first, **Hanne Eckhoff** uses the PROIEL and TOROT corpora to examine the emergence of the category of animacy in Russian. She analyses this development by comparing definiteness-driven differential object marking (DOM) in Old Church Slavonic with the change from constructionally conditioned variation in Old East Slavonic to animacy-subgender marking in late Middle Russian. This research addresses the interesting question of how to annotate constructions that undergo change in historical corpora. Eckhoff defends a conservative approach to annotation in these cases: in order to investigate change in a clear and consistent way, she advocates adherence to the annotation schema that best captures the first-attested stage as long as possible. In addition to this, her study relies on deep annotation in the form of semantic and information-structural features, because conventional treebank annotation (i.e. POS tags and parsed structure) is not sufficient to capture the conditions of the observed variation and change in the emergence of the category of animacy in Russian.

Three articles provide examples of what can be achieved with the rela-

tively unenriched textual corpus resources already available for Welsh. All three of these articles build on the authors' previous research at Marburg as part of the project 'Translations as language contact phenomena' in collaboration with the Parsed Historical Corpus of the Welsh Language on the late medieval manuscript *Llyfr yr Ancr* ('Book of the Anchorite').

Erich Poppe reports on the expression or suppression of finiteness in the second and subsequent conjuncts of clausal coordination in Early Modern Welsh. In doing so, he shows how simple lexical searches can be used to examine change in relatively abstract syntactic structures such as coordination. In this case, for instance, he used the text of the 1588 Welsh Bible translation in the Early English Books Online database as a starting point, comparing the wording found with clausal complementisers with the wording found in the earlier 1567 New Testament and later revised 1620 Bible, as well as corresponding passages in the Greek New Testament or Hebrew text of the Old Testament as appropriate.

Elena Parina uses a POS-tagged version of a text that is intended to become part of PARSHCWL, namely the late-16th or early-17th-century translated Welsh collection of tales, the *Gesta Romanorum*. She searches for specific lexical items associated with a particular type of relative-clause marking, and compares each instance with the parallel passage in the 1510 and 1577 English editions of the same text from ProQuest's Early English Books. These English texts approximate to the source from which the Welsh was translated. Using relatively straightforward search procedures, she is able to demonstrate an increase in frequency of explicit marking of relative pronouns and an association between use of such pronominal marking with nonrestrictive relative clauses and with presentational constructions. The study as a whole suggests a possible future avenue for research investigating the emergence of register variation.

In the final contribution to this volume, **Raphael Sackmann** investigates patterns of subject marking in nonfinite clauses in an Early Modern Welsh text that forms part of the Historical Corpus of the Welsh Language 1500–1850. Through close textual analysis alongside descriptive statistical evidence, he shows that marking of subjects in this text is already closer to Present-day Welsh than to Middle Welsh, and attempts to explain the use of different strategies in terms of such semantic factors as tense, anteriority, generic/future reference and telicity. He further notes that creation of larger corpus resources will be the way to establish whether these results generalise to the entirety of the language attested at this period.

Overall, these articles showcase the range of current work in the creation and use of historical parsed corpora. They demonstrate that, while the pro-

duction of such corpora involves significant effort and necessitates the careful consideration of many practical and theoretical issues, the rewards in terms of empirical contribution to research are also substantial. Future development of this research agenda will surely continue to be a major aspect of work in historical syntax in the coming years.

Marieke Meelen
University of Cambridge
Trinity Hall
Trinity Lane
Cambridge
CB2 1TJ
mm986@cam.ac.uk

David Willis
University of Oxford
Jesus College
Turl Street
Oxford
OX1 3DW
david.willis@ling-phil.ox.ac.uk
davidwillis.net