Historical Treebanks and Linguistic Theory/Schemes, Interoperability and Applications

Organizers
Lisa Gotthard (l.gotthard@ed.ac.uk), University of Edinburgh
Dag T. T. Haug (d.t.t.haug@ifikk.uio.no), University of Oslo
Nikolaos Lavidas (nlavidas@enl.uoa.gr), National and Kapodistrian University of Athens
Christopher Sapp (csapp@iu.edu), Indiana University
Carola Trips (carola.trips@uni-mannheim.de), University of Mannheim
George Walkden (george.walkden@uni-konstanz.de), University of Konstanz

Annotated corpora have become the main empirical basis of historical linguistics. Over the last twenty-five years, parsed corpora in the Penn tradition, the dependency treebanks of the PROIEL family and their descendants, and the historical treebanks released within Universal Dependencies enable researchers to retrieve instances of an abstract structure across a language's recorded history and to test hypotheses about change on counts rather than solely on collections of examples (Taylor 2020: 195–212; Eckhoff et al. 2018: 29–65; de Marneffe et al. 2021: 255–308).

Treebank construction is a methodological intervention, intricately linked to linguistic theory (Trips & Stokes 2023). Annotation is not a neutral layer laid over the text. Every tag in a historical treebank records an analytical decision, made under conditions that differ from those of living languages: no native speaker can be consulted, texts survive in editions that are themselves interpretations, and the relevant categories often changed during the period that the corpus covers. On the one hand, research questions can influence the choice of annotation schemes (e.g., the Penn corpora's lack of a VP), and, on the other hand, the composition of the corpus may shape what research questions can be asked and what answers can be obtained (Haug 2015).

Recent developments in treebank construction are opening up new research possibilities. First, as more treebanks come online, new avenues are created for comparative work across related corpora (e.g., Walkden's (2025) investigation of parataxis and hypotaxis across seven languages). Second, treebanks are increasingly multi-dimensional, with some corpora adding information structure or discourse relations (e.g., release 1.0 of the *Corpus of Historical Low German*; Götze et al. 2007), lemmatization (e.g., Percillier & Trips 2020; Percillier et al. 2024), semantic information (e.g., Struik et al. 2026), sociolinguistic variables, and links to external resources such as dictionaries or maps. Increasing attention is also being paid to forms of metadata that capture the textual and material dimensions of language use. Finally, advances in computational linguistics can facilitate the creation of substantially larger datasets (see, e.g., Kulick et al. 2021, 2022), allowing infrequent syntactic patterns or fine-grained sociolinguistic differences to be queried.

However, the growth of resources has also produced a problem of comparability; the same phenomenon – whether null subjects, clitic placement, argument structure, aspectual composition, or information structure – is annotated in different and often incompatible ways across corpora, so that comparisons across historical stages or languages may in fact compare annotation schemes rather than grammars. Conversion between schemes does not remove the

problem, since it loses information: the empty nodes and secondary dependencies of the PROIEL scheme have no direct counterpart in the basic Universal Dependencies representation, and constituency and dependency annotation encode partly different distinctions. Beneath the linguistic layers lies the philological one. The choice of edition, the treatment of variants, and the degree of normalization of spelling and word division all change what is counted: an edition of the Greek New Testament based on the early majuscules and one based on the Byzantine text-type are, for the purposes of quantitative syntax, different corpora. Moreover, automatic annotation now extends corpora to stages and text types for which no gold data exist, and the reliability of such silver annotation varies sharply with period, genre, and tool.

The workshop brings together builders and users of annotated historical corpora to address these issues. We invite contributions on questions including (but not limited to):

- How have historical treebanks contributed to recent advances in theories of diachronic syntax?
- How do theoretical assumptions influence corpus design, annotation practices, and interaction between annotation layers?
- How can treebanks support cross-linguistic comparison while ensuring interoperability across schemes, formats and annotation categories?
- How can we ensure diachronic comparability, that is, the use of a single scheme across stages of one language where the category inventory changes, and across translations and parallel texts?
- How should treebanks deal with the philological layer, i.e. editorial practices, manuscript variation, text normalization, and the segmentation of texts into sentences and tokens?
- What innovative layers of treebank annotation are resulting in new discoveries about syntactic variation and change?
- Can methods from other disciplines such as philology, manuscript studies, and historical sociolinguistics inform the design of future treebanks?
- What are the limitations of parsed corpora – what can they not tell us, and why not?
- What forms of metadata are most valuable for enriching historical corpora and investigating language history?
- How can computational advances be used to facilitate construction and annotation of the next generation of corpora? What problems do these advances pose? Can automatically annotated data serve as evidence for change?

Contributions may concern any language and any period, and case studies that show what a particular annotation decision does to a diachronic result are as welcome as papers on the design of schemes and tools. We expect contributions on Greek and Latin, on older Germanic, Romance, and Slavic languages, Indo-Iranian, and Semitic, including treebank projects of the PROIEL family, the Penn tradition and Universal Dependencies, and projects that annotate semantics and discourse. The convenors' own work supplies some of the cases the workshop will discuss: the AthDGC corpus (*Athens Digital Glossa Chronos*, Athens-PROIEL), which extends PROIEL annotation to every stage of Greek from Homer to the present and documents the drop in reliability of automatic annotation for the modern stage (Lavidas et al. 2026); the parallel treebank of the early Indo-European Bible translations, whose alignment layer was designed for the kind of cross-linguistic comparison that translation corpora make possible (Haug and Jøhndal 2008: 27–34); new constituency-parsed corpora including the *Indiana*

Parsed Corpus of Historical (High) German (Sapp et al. 2026) and the INTSAYT project (Wallenberg et al.: https://gtr.ukri.org/projects?ref=UKRI2198#); and ongoing efforts to extend this framework to Middle Norwegian in the Norchron project.

We invite up to fifteen papers in 30-minute slots, which will be followed by a closing general discussion of about 45 minutes on standards for comparability across historical corpora. We intend to publish selected papers in a thematic journal issue or edited volume.

References

Eckhoff, Hanne Martine, Kristin Bech, Gerlof Bouma, Kristine Eide, Dag Trygve Truslew Haug, Odd Einar Haugen & Marius Jøhndal. 2018. The PROIEL treebank family: A standard for early attestations of Indo-European languages. Language Resources and Evaluation 52(1). 29–65.

Götze, Michael, Thomas Weskott, Cornelia Endriss, Ines Fiedler, Stefan Hinterwimmer, Svetlana Petrova, Anne Schwarz, Stavros Skopeteas & Ruben Stoel. 2007. Information structure. In Stefanie Dipper, Michael Götze & Stavros Skopeteas (eds.), *Information structure in cross-linguistic corpora: Annotation guidelines for phonology, morphology, syntax, semantics and information structure*, 147–87. Potsdam: Universitätsverlag Potsdam

Haug, Dag. 2015. Treebanks in historical linguistic research. In Carlotta Viti (ed.), *Perspectives on Historical Syntax,* 185–202. Amsterdam: Benjamins.

Haug, Dag T. T. & Marius L. Jøhndal. 2008. Creating a parallel treebank of the Old Indo-European Bible translations. In *Proceedings of the Second Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2008)*, 27–34. Marrakech: ELRA.

Kulick, Seth, Neville Ryant & Beatrice Santorini. 2021. Penn-Helsinki Parsed Corpus of Early Modern English: First Parsing Results and Analysis. *Society for Computation in Linguistics* 5(1), 143–57. https://doi.org/10.7275/twww-ef90

Kulick, Seth, Neville Ryant & Beatrice Santorini. 2022. Parsing Early Modern English for Linguistic Search. In Allyson Ettinger, Tim Hunter & Brandon Prickett (eds.), *Proceedings of the Society for Computation in Linguistics (SCiL) 2022*, 143–57. Association for Computational Linguistics. https://api.semanticscholar.org/CorpusID:211296688

Lavidas, Nikolaos, Kiki Nikiforidou, Dag T. T. Haug, Leonid Kulikov et al. 2026. AthDGC: Athens Digital Glossa Chronos (Athens-PROIEL), v0.4.0. Zenodo. https://doi.org/10.5281/zenodo.20439182

de Marneffe, Marie-Catherine, Christopher D. Manning, Joakim Nivre & Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics* 47(2). 255–308.

Percillier, Michael, Yela Schauwecker, Achim Stein & Carola Trips. 2024. *Carrying Verbs Across the Channel: Modelling Change in Bilingual Medieval England*. Cham, Switzerland: Palgrave Macmillan. https://doi.org/10.1007/978-3-031-50806-6

Percillier, Michael, & Carola Trips. 2020. Lemmatising verbs in Middle English corpora: The benefit of enriching the Penn-Helsinki Parsed Corpus of Middle English 2 (PPCME2), the Parsed Corpus of Middle English Poetry (PCMEP), and A parsed Linguistic Atlas of Early Middle English (PLAEME). In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk & Stelios Piperidis (eds.), *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 7170–8. Marseille, France: European Language Resources Association.

Sapp, Christopher, Rex Sprouse, & Elliott Evans. 2026. *Indiana Parsed Corpus of Historical (High) German*. https://ipchg.iu.edu/

Struik, Tara, Lena Kaltenbach, & Susanne Lang. 2026. On the benefits of enriching the Penn Parsed Corpora of Historical English to study change at the syntax-semantics interface: Introducing the Roots & Results (RoRe) database. *Journal of Open Humanities Data* 12(105), 1–13. https://doi.org/10.5334/johd.546

Taylor, Ann. 2020. Treebanks in historical syntax. *Annual Review of Linguistics* 6. 195–212. https://doi.org/10.1146/annurev-linguistics-011619-030515

Trips, Carola, & Peter A. Stokes. 2023. From Original Sources to Linguistic Analysis: Tools and Datasets for the Investigation of Multilingualism in Medieval English. In Sara M. Pons-Sanz & Louise Sylvester (eds.), *Medieval English in a Multilingual Context*, 49–91. Cham, Switzerland: Palgrave Macmillan. https://doi.org/10.1007/978-3-031-30947-2_3

Walkden, George. 2025. Parataxis and hypotaxis in historical corpora. *Language* 101(4). e343–e368.