Named Entity Recognition for a Text-Based Catalog of Ancient Greek Authors and Works

Berti, Monica
monica.berti@uni-leipzig.de
Leipzig University, Germany

Introduction

This poster proposal presents a project whose results are the linguistic extraction, analysis, and annotation of ancient Greek bibliographic references. This project has been producing a structured knowledge resource about ancient Greek authors and works, which is constituted by names of authors and titles or descriptions of works with variants and ambiguities in the original contextual language. The project makes use of methods and technologies from NLP and computational linguistics. In particular, it applies NER for semi-automatic extraction and annotation of entities related to bibliographic references to ancient Greek authors and works.

Project content

The poster summarizes characteristics of the database behind the project, which contains lemmatized and annotated named entities extracted from literary works with a focus on the *Deipnosophists* of Athenaeus of Naucratis (http://www.digitalathenaeus.org). The text of Athenaeus has been chosen because it is a reference resource for exploring language and patterns of ancient Greek bibliographic references. Moreover, its citations cover about 50% of the total number of Greek authors for the period of time between the 8th century BC and the 3rd century CE. The poster will also show methods and tools for named entity disambiguation, linking, and coreference resolution in order to extract and annotate real entities corresponding to author names (e.g., Ἀριστοφάνης ὁ Βυζάντιος) and work titles (e.g.,ἐν τῇ Λακεδαιμονίων Πολιτείᾳ). This project is producing a know-ledge base for linking entity mentions to a structured vocabulary for ancient Greek authors and works that can be used to annotate other significant texts, as for example the *Lexicon of the Ten Ora-tors of Harpocration* and the*Suda* lexicon, which are currently analyzed and whose results will be presented and discussed with the attendees of the poster as part of the *Linked Ancient Greek and Latin* (LAGL) project (https://www.lagl.org/).

Methods

Through the illustration of concrete examples, the poster shows the following project phases: 1) extraction and lemmatization of NEs; 2) annotation of single NEs in generic entity classes (LOC and LOCderiv for place names and derivatives; ORG and ORGderiv for festivals, Panhellenic games, and derivatives; OTH for miscellaneous entities as works, months, constellations, currencies, languages, groups, etc.; PER and PERderiv for gods, persons, personifications, authors, etc. and derivatives); 3) disambiguation of NEs through external authority lists for personal names and places like the *Lexicon of Greek Personal Names* (LGPN) and the *Pleiades* gazetteer; 4) disambiguation of NEs related to ancient authors and works and their identification with canonical citations according to the CITE Architecture; 5) use of the web-based platform INCEpTION (https://inceptionproject. github.io/) for relation extraction, NE linking, and co-reference resolution; 6) data export according to Linked Open Data best practices.

Goals

The final goal of the project is the creation of a catalog of ancient authors and works based on linguistic annotations in order to visualize references in their original context and offer new and dynamic text-based tools that are not available in existing indices and catalogs of Classical literature. This project offers new results about the language of ancient Greek bibliographic citations by documenting them with an immediate, full, and complete contextual analysis of their occurrences. Attendees of the poster session at DH2023 will be able to interact with extant and forthcoming online resources of the project, so that it will be possible to exchange ideas for further data production and collaboration with related projects in the digital humanities community.

Bibliography

Berti, Monica (2021): *Digital Editions of Historical Fragmentary Texts*. Heidelberg: Propylaeum. DOI: 10.11588/propylaeum.898.

Berti, Monica (2019a): “Historical Fragmentary Texts in the Digital Age”. In *Digital Classical Philology. Ancient Greek and Latin in the Digital Revolution*. Ed. by M. Berti. Berlin and Boston: De Gruyter, 257-276. DOI: 10.1515/9783110599572-015.

Berti, Monica (2019b): “Named Entity Annotation for Ancient Greek with INCEpTION”. In *Proceedings of CLARIN Annual Conference 2019*. Ed. by K. Simov and M. Eskevich. Leipzig, Germany: CLARIN, 1-4.