Towards a semantic and syntactic search engine for electronic corpora of Greek and Latin

A Data Management Plan created using DMPonline

Creator: William Short

Affiliation: University of Exeter

Template: University of Exeter

Project abstract:

Existing applications for searching electronic corpora of ancient languages have greatly facilitated research and pedagogy by permitting users to rapidly query large collections of Greek and Latin texts, and by keying matching results to dictionary entries and morphological analysis. However, most of these tools were designed exclusively for word-form queries and cannot accommodate grammatical specifications as search parameters. Moreover, although computational semantic search has been explored using parallel bilingual dictionaries, no service yet exists permitting users to query meanings in either corpus, nor has any taken advantage of the precision that a lexical database could afford. This project will design and implement a software system that integrates conceptual-semantic (‘WordNet’) data with the syntactic annotations of ‘treebanks’ to enable users, for the first time, to search Greek and Latin texts based on their semantic and syntactic properties – opening these texts to new kinds of linguistic, literary, and cultural study.

Last modified: 29-09-2018

Copyright information:

Towards a semantic and syntactic search engine for electronic corpora of Greek and Latin

Data

This project involves two largely separate datasets. The first consists of conceptual-semantic information for the Latin language in the form of an SQL database (the Latin WordNet '2.0'). This dataset re-uses -- but significantly builds on -- the data created by Stefano Minozzi for the Fondazione Bruno Kessler's MultiWordNet Project in 2008, which is distributed under a Creative Commons Attribution-ShareAlike 4.0 International (CC BY-SA 4.0) license.

The second dataset will consist of 'treebank' syntactic mark-up for Latin (and eventually Greek) texts, in the form of TEI-encoded XML files. Treebank data will be aggregated from three sources: the *Greek and Latin Dependency Treebank* created by the Perseus Project at Tufts University, the *Index Thomisticus,* and the *PROIEL* project, all of which operate under similar Creative Commons Attribution-ShareAlike licenses.

Our expansion of the Latin WordNet willl increase the size of this knowledge-bank from 9000 to over 40000 lemmas. According to the terms of the ShareAlike license, this new data will also be made publically and freely available under CC BY-SA 4.0.

To integrate treebank data into our search engine architecture (ANNIS), XML files will be modified in order to standardized annotation according to our technical design. In particular, synset data from the Latin WordNet '2.0' will be added, along with annotations for ranking synset assignments. Like the WordNet data, these new treebank files will be made available under a CC BY-SA 4.0 license.

New WordNet data is being created through an internal Django web site, with a PostgresSQL database backend. The data consists of lexical, morphological, and semantic information for over 35,000 Latin words.

Including the WordNet database and treebank files, our dataset will likely be less than 500 GB.

The WordNet data will be manually curated and entered via our internal tooling. Once synset information has been integrated into the treebank mark-up, this data will be reviewed for accuracy by the project team.

Documentation and description

The data of the Latin WordNet '2.0' has been created according to the specifications of the MultiWordNet project (which is itself a multi-lingual version of the Princeton WordNet for English). Presently existing documentation will be expanded to reflect our modifications of this specification, however. For instance, our modification of the WordNet includes differentiation of literal, metonymic, and metaphorical senses of words and new documentation will explain our revised database format (to aid the creation of third-party APIs).

New documentation is also now being created to explain our treebank annotation structure and mark-up procedures. This documentation will explain, in particular, our system for integrating semantic 'synset' data with syntactic annotations.

A separate Software Design and Functional Specification Document detailing the search engine's architecture will be created. This SDFSD will be made available on the project's blog, GitHub site, and will be included with data distributions.

OAI-ORE Open Archives Initiative Object Reuse and Exchange

Data Protection

During development, and throughout the period of this grant, the Latin WordNet '2.0' and treebank datasets will be deposited in the University of Exeter ORE data management system. In the production phase, the linguistic data that our search engine will aggregate – the semantically and syntactically annotated texts in TEI XML format, along with the WordNet SQL database supporting the search interface – will be hosted on University of Exeter production servers. This data, along with all documentation, will also be made available to the open-source community on GitHub.

Data will be automatically backed-up as part of the regular University data management regime.

In production, modification and maintenance of databases on the University’s web servers will be restricted by secure authentication methods to appropriate team members and university IT staff. During development, data will centralised on university 'cloud' servers and access will be delegated by the project lead. Collaborators will be given appropriate temporary access to this data when they are not directly associated with the host institution.

N/A

Retention and preservation

Data is not sensitive and is intended to be enduring and maintained for public access following the completion of the project. Once the production databases have been created, development databases will be erased.

WordNet data will be made available in the form of an SQL dump and treebank data will be made available as XML files. As this amounts to a collection of text files, no conversion or special software is required for access.

On university servers; on the project's blog; and on GitHub.

Data volume is relatively minimal; we project no more than 500 GB. No costs will be related specifically to deposit of the datasets, however in production the search engine will need to be maintained on university servers.

Data sharing

All data will be shared on the project's blog and GitHub during development. The project's WordNet data is already available, in preliminary form, on GitHub at https://github.com/wmshort/latinwordnet. This data will be periodically updated to reflect on-going progress until the dataset is complete. Once the source code for our search engine is available and treebank data has been completed, this will be made available on a new project page.

N/A

Data is already available, and will continue to be updated over the course of the project. No plan for commercialization. No embargo is planned.

Research concerning the design of our software system or dealing with theoretical issues that arise during its implementation will be submitted to academic journals. All data will be made available under a Creative Commons license, and the project is entirely open-source and open-access.

Data Protection Impact Assessment

N/A

N/A

N/A

N/A

N/A

N/A

N/A

N/A

N/A