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.

ID: 31846

Last modified: 29-09-2018

Copyright information:

much of the text of this plan as they would like in their own plans, and customise it as necessary. You do not need to credit the creator(s) as the source of the language used, but using any of the plan's text does not imply that the creator(s) endorse, or have any relationship to, your project or proposal

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

Data

If you are re-using existing data, what licences or terms of use will you have to comply with?

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.

How will new data build on and relate to existing data?

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.

What types of new data will you create and in what format?

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.

Can you estimate the size of the data you will create?

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

500 GB.

What methods will you use to capture your data and how will these ensure that your data are high quality?

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

What contextual information is needed for you or someone else to understand your data?

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.

How will you capture contextual information?

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.

Will you use any metadata standards?

OAI-ORE Open Archives Initiative Object Reuse and Exchange

Data Protection

Where will you store your data and how will you ensure that they are backed up? Will you use University-managed data storage or will you need to set up your own back-up procedures?

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.

How will you secure your data? What methods will you use to restrict access to your sensitive data? Will you encrypt hardware when working off campus?

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.

How will you protect your research participants? Will you obtain informed consent for data retention and sharing? How will you anonymise data to safeguard the privacy of your participants?

N/A

Retention and preservation

Which subsets of your data will you keep at the end of your project? Will you retain anonymised versions but destroy personal data and identification keys? Will you retain all of the raw data or is a processed version more suitable to preserve? Do you need to keep all intermediary files or would you only need to refer back to input files or a final version?

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.

How will you prepare your data for long-term preservation? Are you able to convert your data to open file formats? What contextual information do you need to retain so that your data remain understandable and usable?

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.

Where will you archive your data to ensure that they are preserved and sustained for several years after your project ends? Will you submit your data to a specialist data repository/centre and if so, have you consulted them about your requirements?

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

How big will your final dataset be and will there be any costs associated with archiving them, such as data deposit charges?

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

Can you demonstrate that you'll plan ahead to maximise data sharing? For example, will you only share a subset of the data where informed consent was granted for 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.

Are there any reasons why you would not be able to share some of your data? Would they be covered by data protection legislation, licence restrictions, or contractual confidentiality clauses? Are there ethical reasons why data should not be released?

N/A

When will you share your data? Will data be made available upon first publication of findings or within a limited period after the end of the project? Do you need to delay publication to allow for commercialisation or patent applications? Will you embargo your data to allow for a limited period of exclusive use?

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

How will you disseminate your research? Will you include a data access statement in published articles? Does your chosen method of data preservation provide a persistent URL such as a Digital Object Identifier? What licences will you assign to your data?

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

What do you require this personal data for? What is the purpose of using the personal data?

N/A

How are you making people aware of how their personal data is being used? Do you need to update your privacy notice?

N/A

Which conditions for processing apply for your project? For Special Categories please ensure you select at least one from Section 1 and one from Section 2 below. Please select all that apply and provide any additional details. Section 1: Conditions for Personal Data

- **The data subject has given consent to the processing (please provide the consent wording and where it is stored)**
- **Contractual necessity (please confirm which contract this relates to)**

- **Compliance with any legal obligation (please document which legal obligation)**
- **To protect the vital interests of the data subject (please provide details)**
- **Functions of a public nature or task in the public interest (please provide details)**
- **Legitimate interest of the Data Controller (please provide details of legitimate interest)**

Section 2: Conditions for Special Categories Data

- **The data subject has given explicit consent to the processing**
- **Necessary so that you can comply with employment law**
- **To protect the vital interests of the data subject or other person**
- **The processing is carried out as part of the legitimate activities of a not-for-profit organisation**
- **The individual has deliberately made the information public**
- **The processing is necessary in relation to legal rights**
- **The processing is necessary for administering justice or for exercising statutory or governmental functions**
- **The processing is necessary for medical purposes**
- **The processing is necessary for monitoring equality of opportunity**

N/A

Is all the personal data you are using necessary? Are you collecting enough to carry out the work, is there any you could do without to limit the risks to the individuals?

N/A

How are you ensuring that personal data obtained from individuals or other organisations is accurate? How will you keep it updated?

N/A

How long will you keep the data and how will you dispose of it? Are the retention periods on the University Retention Schedule?

N/A

Where will the data be stored? If storage is in the cloud, where is the physical server? Will you need to transfer the data outside the EEA? If yes, how will you ensure adequate protection?

N/A

Will you be able to meet all the Data Subject Rights? Can you provide copies of data if requested? Are you able to fully delete the data (not just archive)?

N/A

Please briefly document below any risks with the use of personal data and how you will control such risks. Include technical controls (IT security, encryption etc), physical controls (location, locked room etc), personnel controls (training, access control etc), and procedural controls (contract, polices etc).

N/A