Polyphemus, a lexical database of the Ancient Greek papyri, and the Madrid Wordlist of Ancient Greek

Riaño Rufilanchas, Daniel
danielrianno@gmail.com
CSIC, Spain

Polyphemus: A Lexicographic Database of Greek Papyri

Currently, there is no comprehensive way to search the complete corpus of digitally edited Greek papyri for lemmata, specific grammatical forms, or examples of a grammatical category (Although since 2022, a portion of that corpus is covered by morphosyntactic analysis accessible in the Trismegistos web site). Polyphemus aims to address these shortcomings and more by processing all the papyrus texts from PapyInfo. The processing of these texts is done in conjunction with the processing that results in the Callimachus database. This paper summarizes the procedure by which Polyphemus is obtained, highlighting its key features.

First, each line of papyrus is analyzed to differentiate actual full words from gaps or non-textual elements. Next, complete words are identified and separated from fragments. Afterward, each token is lemmatized and assigned to a part of speech, with its morphological analysis. The Madrid WordList is used to aid this process, with proper nouns separated from common nouns. Lemma assignment and POS tagging is performed in three phases, with disambiguation using a system of morphosyntactic rules.

All this information is transferred to a SQL database and related to the data on the papyri obtained when creating the Callimachus database. For each lexical form (token), a lemma, a partially disambiguated morphological analysis, and a translation or gloss are obtained. Each of these parameters can be searched in combination with over fifty categories, such as date, origin, category, extension, subject, etc., made available by Callimachus. To date, more than the 97% of the complete words, including proper names, have been analyzed.

The Madrid Ancient Greek Word List

The lemmatization and POS tagging are performed by comparing each record in the Polyphemus database with the Madrid Ancient Greek Wordlist, which has been created over the last three years. While many Ancient Greek wordlists are evolutions, simplifications, or improvements from the Morpheus list, the Madrid Ancient Greek Wordlist is enriched with data coming from several treebanks, the digital version of the *Greek-English Lexicon* of Liddell-Scott-Jones (LSJ), Bailly's *Dictionnaire Grec-Français*, about 100,000 proper names from *The Lexicon of Greek Personal Names*, the Trismegistos repository of papyrological and epigraphic resources, and other minor resources. This data was processed to obtain morphological information, generating automatically the Attic and Ionic paradigm for each nominal entry in LSJ and Bailly. The lemmas are assigned a translation or gloss, mainly from the LSJ, S.C. Woodhouse's *English-Greek dictionary* or the *Organa Papyrologica* web site.

Polyphemus Interface

Polyphemus can be accessed online, currently containing about 4,600,000 words from Ancient Greek papyri. POS tagging and lemmatization enable the user to query the database for any morphological feature, lemma, or translation, and to combine this data with the formal content of the papyri provided by the Callimachus database, allowing more than 80 search criteria. Since both the original readings and editorial regularizations are preserved, researchers can use Polyphemus to search for phonetic or morphological features of the papyri. Examples of searches include: (a) texts containing a Greek word that translates as "poison," "medicine," " *praetor*," "water," etc.; (b) texts in which any lemma appears in a specific grammatical form from Elephantine between the 2nd century BC and 3rd AD; (c) all adjectives in accusative plural; or the optative of verbs in -μι in all texts.

Bibliography

Bohnet, Bernd / **Joakim Nivre** (2012): "A Transition-Based System for Joint Part-of-Speech Tagging and Labeled Non-Projective Dependency Parsing" *EMNLP-CoNLL*, pp. 1455-1465 [https://aclanthology.org/D12-1133].

Celano, Giuseppe G.A. / **Gregory Crane** / **Saeed Majidi** (2016): "Part of Speech Tagging for Ancient Greek" *Open Linguistics* 2:393–399 [DOI 10.1515/opli-2016-0020].

Crane, Gregory (1991): "Generating and Parsing Classical Greek" *Literary and Linguistic Computing*, 6:4, pp. 243–245 [https://doi.org/10.1093/llc/6.4.243].

Riaño Rufilanchas, Daniel (2006): El complemento directo en griego antiguo in: Anejos de Emerita, XLVII. Madrid: CSIC.