William Michael SHORT

Next-Gen search for Ancient languages: Meaning- and syntax-based querying of electronic corpora of Greek and Latin

Currently available tools for searching and browsing electronic corpora of Greek and Latin have greatly facilitated research, pedagogy, and learning by permitting scholars, teachers, and students to rapidly query words or phrases within large collections of texts, and by providing annotation-rich 'hits' that are automatically keyed to lexical data and morphological analysis. These tools permit word, phrase, or lexical co-occurrence searches within single texts or collections of texts, and (in the best cases) include the option of querying all possible inflections of words. However, most of these tools were designed exclusively for word-form (*n*-gram) queries and cannot accommodate morpho-syntactic properties as search parameters. What's more, although computational semantic search has been explored, no service yet exists permitting users to query *meanings* of words in either corpus, nor has any taken advantage of the precision and performance that a lexical database could afford.

The University of Exeter's TExtual Project aims to develop an entirely new search engine that will open Latin (and, later, Greek) texts to queries based – for the first time – on both their semantic and their syntactic properties. In addition to basic word-form queries, this engine will allow users to specify search parameters that include word senses or that capture broad semantic domains. In other words, it will enable searching texts for meanings independent of their lexical instantiation. For instance, a query could be constructed to 'find all occurrences of words meaning *love*' (optionally specifying the part of speech) or to 'find all occurrences of any word belonging to the domain of AGRICULTURE or MILITARY'. If desired, the user could even narrow the query to a particular sense of *love*, distinguishing, 'be enamoured with' (as in *Joanie loves Chachi*) from 'derive pleasure from' (as in *I love cooking*). Additionally, the engine will permit searches to be executed based on the morpho-syntactic properties of words. That is, it will enable searching texts for 'constructions' again independent of the lexical element or elements that happen to instantiate them: as in, for example, 'find all (sequential) occurrences of GENITIVE CASE and *ergo*' or 'find all (proximal) occurrences of PASSIVE VOICE and DATIVE'. Meaning- and syntax-based parameters could also be combined into complex multi-term queries such as 'find occurrences of any word meaning *attack* in the future tense' or 'find occurrences of any word in the domain of HUMAN ANATOMY in the accusative singular'.

In this paper, I discuss two major challenges to implementing this kind of search, and our project's approach to these challenges. The first is to integrate meaning-based queries over large-scale text databases. One technique would be to overlay 'reverse dictionary' translation on regular expression matching, generating batteries of word-form queries at the front-end of searches. But this would be prohibitively inefficient (especially as match-sets would increase exponentially for multi-term queries) as well as inexact (as translation dictionaries tend not to make fine-grained sense distinctions). Our search engine will instead incorporate the Latin WordNet, a lexico-semantic knowledge-base created originally as part of the Fondazione Bruno Kessler's MultiWordNet Project (see Minozzi 2008). Besides enabling performant querying of specific senses of words, this will permit searches based on common semantic and lexical relations (antonymy, hyperonymy, hyponymy, derivation, and so on) as well as in any language covered by the MultiWordNet (in particular, English and

Italian). The second challenge is to determine the correct syntactic structure of texts and to devise a morpho-syntactic annotation scheme capable of integrating with potentially very different grammars (including both constituent and dependency grammars) at an appropriate level of linguisitic description. In this respect, our engine will largely rely on annotated corpora such as the LASLA Latin databank and, where available, treebank data from the Greek and Latin Dependency Treebank, the Index Thomisticus, and the PROIEL Treebank, which will be converted to a unified format under the 'ANNotation of Information Structure' (ANNIS) architecture (Krause and Zeldes 2016).

Our system will thus be able to offer several tiers of functionality. The 'gold standard' will make use of fully syntactically parsed treebank data, though presently this represents an extremely circumscribed case (probably fewer than 100,000 words). A second tier will utilize partial morphological annotations, for an again restricted but somewhat larger corpus (around 1,600,000 words). In either instance, comprehensive and efficient meaning- and syntax-based queries can be delivered by pre-processing texts through a mark-up pipeline. A third tier will be available for 'raw' text input, using corpora provided by the Classical Language ToolKit or from the user, which will be parsed on the fly to provide some degree of syntactic and semantic awareness (for a potentially unlimited token set). In whatever case, we believe the TExtual engine represents a generational shift that will empower new kinds of linguistic, literary, and cultural study of Greek and Latin texts, as well as offering novel ways of exploring ancient literature. Rough proof-of-concept will be demonstrated.