SunoikisisDC Summer 2021 Session 7

Reading and Learning ancient Greek with Diorisis and the Scaife Viewer

The Diorisis Ancient Greek Corpus and Diorisis Search

Alessandro Vatri
University of Oxford

The Diorisis Ancient Greek corpus

- 820 texts (from Homer to Nonnus)
- Automatically lemmatized, PoS-tagged, with possible morphological parses
- Sources:
  - the Perseus Canonical Greek Lit repository

    752 texts, https://github.com/PerseusDL/canonical-greekLit
  - “The Little Sailing” digital library

    8 texts, http://www.mikrosapoplous.gr/en/texts1en.html
  - the Bibliotheca Augustana digital library

    60 texts, http://www.hs-augsburg.de/~harsch/augustana.html#gr

The Diorisis corpus: size

	8BC	7BC	5BC	4BC	3BC	2BC	1BC	1AD	2AD	3AD	4AD	5AD	Total
Comedy			78,949	15,805									**94,754**
Essays				3,827				475,169	263,270	361,213	23,965		**1,127,444**
Letters				9,566	1,333					9,654	164,388		**184,941**
Narrative			334,791	208,921	311,307		661,459	961,707	482,568	411,061	98,382		**3,470,196**
Oratory			57,542	529,374				184,783	295,583	2,855	55,683		**1,125,820**
Philosophy				895,412					112,846	213,493			**1,221,751**
Poetry	199,050	16,025	21,120		80,783	3,158	7,341		22,752	17,714	60,107	126,892	**554,942**
Religion		15,788			131,895	463,115		133,864	44,919		17,884		**807,465**
Technical			104,091	326,746	15,409		385,503	24,056	394,012	157,947	3,886		**1,411,650**
Tragedy			207,458										**207,458**
Total	**199,050**	**31,813**	**803,951**	**1,989,651**	**540,727**	**466,273**	**1,054,303**	**1,779,579**	**1,615,950**	**1,173,937**	**424,295**	**126,892**	**10,206,421**

The Diorisis corpus: disambiguation

- Roughly 1 out of 10 Ancient Greek word-forms may be analysed as a form of more than one lemma.
- E.g. *praxeis*:
  - lemma: *prassô*; PoS: verb; morphology: second person singular, active future indicative;
  - lemma: *praxis*; PoS: noun; morphology: nominative or accusative plural.
- PoS automatic tagger as a support for disambiguation.
  - TreeTagger (http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/)
- Limitations, e.g. *basileiôn*:
  - genitive plural of *basíleia* (‘queen’);
  - genitive plural of *basiléia* (‘kingdom’);
  - masculine or neuter nominative singular of the present participle of *basileiaô*.
- Confidence : 100% for output ‘verb’, 50% for output ‘noun’.

The Diorisis corpus: accuracy

- Accuracy (v. 1.0): 90% for Homer, 97% for Lysias (Vatri and McGillivray 2020)
- Version 1.5:
  - Manual corrections suggested by Diorisis Search users
  - Use of dialect information in dictionary
  - Frequency: higher-frequency lemmas preferred
    - This punishes indeclinable forms that happen to be homophonous to inflected forms
  - Regular updates on individual documents in JSON format

Diorisis Search

- Diorisis Search is a cross-platform desktop application built on the Electron framework
- The main purposes of Diorisis Search are:
  - To provide a user-friendly reader for the Diorisis Corpus texts
  - To enable users with no IT training to create and run complex linguistic queries without learning any language