A Wealth of Variables: Using Syntactic Stylometry to Distinguish Signature Constructions in Herodotus and Thucydides

Computational analyses can reveal much about the style of a text, including author attribution, that was invisible to traditional methods. We have applied an algorithm based on syntactic stylometry to the more than 500,000 tokens of ancient Greek prose and poetry that are analyzed in the Ancient Greek and Latin Dependency Treebanks (AGLDT). Initially, our purpose was to determine authorship, something that the computer does correctly on 1,000 word chunks at a rate of 99.91% (52 mistakes in 59,000 guesses).

This paper delves into the specific data derived from this analysis in order to sketch in real terms some of the more prominent characteristics distinguishing the "syntactic fingerprints" of Herodotus and Thucydides. That is, it extracts the most defining features of Thucydidean and Herodotean style as compared to each other: what constructions does the one favor and the other avoid to such a degree that it becomes statistically relevant for distinguishing the identity of that author? Surprisingly enough, the telling constructions are often quite mundane and inconspicuous.

Most algorithms widely used to identify the distinctive features of texts (in any language) take lexical words as their input. The AGLDT data can, with relatively simple pre-processing, provide what we call "syntax words," deep linguistic structures that are treated as words when they are subjected to stylistic analysis programs. In this paper, we have chosen to represent every word in the texts by a sequence illuminating its own dependency relationship and part-of-speech, followed by that of its parent word. No lexical information is included. Thus, the first word of Thuc. 1 appears not as Θουκυδίδης, but as *sbj-noun-pred-verb* (a noun acting as subject of the main verb).

The particular algorithm applied to the data here is the widely-used "zeta" analysis (Burrows 2007). The algorithm is implemented in the "Stylo" software package for the statistical program R (Eder, Rybicki, and Kestemont 2016). Essentially, the algorithm compares two groups of texts by dividing the target and the comparison into chunks of a set number of words. It then identifies items (here syntax words) relatively favored by the target author by noting which items occur in many chunks of the target, but few chunks of the comparison texts. The method is sensitive to small variations between texts, on the principle that "a wealth of variables, many of which may be weak discriminators, almost always offer more tenable results than a smaller number of strong ones" (Burrows 2002).

The computer has determined the 110 most distinguishing structures derived from book 1 of Herodotus (32,879 tokens) and books 1 and 3.1-40 of Thucydides (32,306 tokens). An examination of these results reveals certain unexpected and subtle preferences, because the computer looks for not only the frequency of use of a particular construction, but also the consistency of use over large pieces of prose.

For example, Thucydides' 3rd, 4th, and 5th most distinguishing structures demonstrate that he has a predilection for using prepositional phrases as attributes modifying nouns. The particular construction *atr-pronoun-auxp-preposition* (an attributive pronoun as the object of a preposition; e.g., τὰ πρὸ αὐτῆς) is his 3rd most distinguishing feature, occurring 35 times in our sample text as opposed to 10 times in Herodotus 1. However, when the object of the preposition is a noun, it is not listed as a good discriminator, though it is a far more common construction (201 occurrences). Likewise, these attributive prepositions as dependent on nouns or articles that act as objects are significant [#4-5], but when they depend on subject nouns, they lose their import as discriminators.

Thus, syntax can be a far better discriminator than vocabulary for determining authorship of ancient Greek prose. Readily-available software packages enable us to identify specific signature constructions of individual authors that are imperceptible to the casual reader. If we use the zeta analysis to determine the favored distinguishing features of Herodotus and Thucydides, we discover that many constructions that appear to be mundane, when taken together *en masse*, are sufficient to decide authorship with a high level of correctness.

Bibliography

Burrows, John. 2002. " 'Delta': a Measure of Stylistic Difference and a Guide to Likely Authorship." *Literary and Linguistic Computing,* 17.3 pp. 267-287.

-------. 2007. "All the Way Through: Testing for Authorship in Different Frequency Strata." Literary and Linguistic Computing, 22.1, pp. 27-47.

Eder, Maciej. 2015. "Does Size Matter? Authorship Attribution, Short Samples, Big Problem." *Digital Scholarship in the Humanities* 30.167–182.

Eder, M., Rybicki, J. and Kestemont, M. (2016). "Stylometry with R: a package for computational text analysis." *R Journal*, 8(1).107-121.