Machine-Actionable Style: Corpus Linguistics Using Treebanked Textual Data to Compare the Style of Latin Authors

Beginning in fall 2011, I have been collaborating with other researchers within the Perseids Project – a offshoot of the Perseus Digital Library – on an effort to facilitate and capture the intellectual work of students and instructors within a new type of digital commentary: the syntactic treebank. I have explored the pedagogical effectiveness of this methodology in prior papers, noting that – based on observation of the work of over eighty students ranging from the third semester to the most advanced courses – the online treebanking GUI (graphical user interface) allows students and instructors to collaboratively engage an ancient text at a granular level, where every word, clause, and grammatical structure within a sentence made can be fully identified morphologically and syntactically: https://perseids-project.github.io/harrington_trees/.

In the course of this work, I have developed a set of tags (short codes that indicate the morphological, syntactic, and, to a degree, the semantic qualities of an element of a sentence): e.g. A-DO indicates *Accusative Direct Object* or NOM-INDQUES indicates an *Indirect Question* – a nominal clause type. While these data are presented as a visually-accessible commentary in the form of a branching syntactic tree where each word or structure hangs from the specific word or structure that it modifies, each completed text is actually digitally encoded in XML (Extensible Markup Language) code, where each word is specifically identified and connected to its linguistic head numerically. It is the digital nature of these data coupled with the expanding library of treebanked texts that now enables the next phase of philological research using numerical algorithms to model, study, and compare the stylistic and compositional practices of authors. In short, the tags and the connections between the sentence elements and their heads can be subjected to digital analysis where the entire *corpus* is queried to reveal, for instance,

every usage and modification of the word *vis* (i.e. what proportion of instances constituted an *Instrumental Ablative* usage and does the noun show a tendency to modify only certain verbal forms), or every case or construction modifying words derived from *iubere* (i.e. what is the range of syntactic usages attracted by that verb in a specific text or author). The annotated text is thus machine-actionable and open to quantitative study of usage at the scale of the entire *corpus*.

Instead of just comparing the lexical usage of authors or word frequency difference between texts, however, the existence of bodies of digitally annotated text (linguistic *corpora*) allows additional levels of quantitative comparison of usage in terms of proportion and frequency of syntactic usages and constructions: e.g. apposition, asyndeton, hyperbaton, use of clauses as subjects of impersonal verbs, variability in the use of subordinating particles with specific types of clause, etc. In addition, the relative complexity of coordination and subordination becomes open to quantitative analysis and comparison. At this point, the object of this paper can now be engaged: I am seeking to develop an array of quantitative analyses of structure and usage that can be used to explore a range of philological questions without resorting to judgements of taste. It is generally accepted, for example, that the *De Bello Alexandrino* was not written by Caesar, but an analysis along the lines discussed above should be able to ultimately demonstrate an array of structural differences in style that would point the way to other comparisons. It is well known that elegiac couplets tend to be much more syntactically restricted than hexameter verse, but within the elegiac genre is there a recognizable difference in the compositional practices of Ovid and Horace? Is there a quantifiable difference in the use of hexameter by Juvenal versus Lucan (across genre) or between Lucan and Ovid (across period)?

I will argue that two potential objections to this methodology – ambiguity and interpreter effects – will be neutralized by the scale of the *corpora* under consideration. There was,

certainly, a degree in interpretation required of the ancient intended audience – interpretation driven by their education and familiarity with other literary texts and, to some degree, their control of the spoken language; for that reason, as with any commentary or translation, there must be some input from the editor of a treebank to complete the meaning, but that degree of interpretation is highly circumscribed by the sequence and semantic implications of the words in context. Disagreements between scholars about the exact characterization of a single sentence element will occasionally occur, but it is the scale of the data set that ameliorates the inevitable noise in the identification of any single element. Comparison on the level of individual sentences would be unlikely to be reliable, but the methodologies of *corpus* linguistics as applied to these questions should prove effective at the scale proposed and open a new mode of textual study.