Macro-Etymological Textual Analysis

Jonathan Reeve

New York University

The English language has continually borrowed from foreign languages—close to 30% of modern English words are loanwords from French, and another 30% are borrowed from Latin. These words are often concentrated in semantic frames associated with their origin languages—legal vocabulary contains a preponderance of words of French origin, and the vocabulary of the natural sciences contains many words of Latin and Greek origin. The etymology of words in a text, therefore, may be suggestive of its context or its level of discourse. Should a writer choose the Latinate term "masticate" over the Anglo-Saxon term "chew," for instance, one might assume a scientific context or a high level of discursive formality. By computing the proportion of origin languages for all the words of a given text, we may quantify stylistic properties that are potentially revealing about the text and its rhetorical modes.

The Macro-Etymological Analyzer is a computer program that I wrote for this purpose. Written in PHP on a LAMP stack, it is a web app accessible at http://jonreeve.com/etym, and is freely available for all to use, modify, and distribute under the GPLv3. It accepts as input a user-uploaded text file, and looks up each word in Gerard de Melo's Etymological Wordnet database. These words are then counted by language of origin using two generations of language ancestry, and then categorized by language family. The results are displayed as a pie chart made with the Google Data Visualization API, along with a CSV log file which can be used for comparative analyses. Currently, the program accepts only English texts, but the database supports queries from any source language, and plans are in place to make the program fully multilingual.

Figure 1 shows the proportions of Latinate words—words descended from Latin or romance languages—for each of the 15 genres in the Brown Corpus. Learned texts and government documents show the highest proportions of Latinate words, whereas romance and adventure stories show the lowest. The same textual categories sorted by proportion of Hellenic words (words of ancient Greek origin) show changes in certain categories—religious language exhibits a higher rank, and that of mystery stories is ranked lower than in the Latinate scale. These data suggest that a high proportion of Hellenic words is correlated with religious language, among other genres, and that a high proportion of Latinate words is correlated with learned language. Once literary works are analyzed with this method, these hypothetical correlations become potentially useful as literary critical tools.

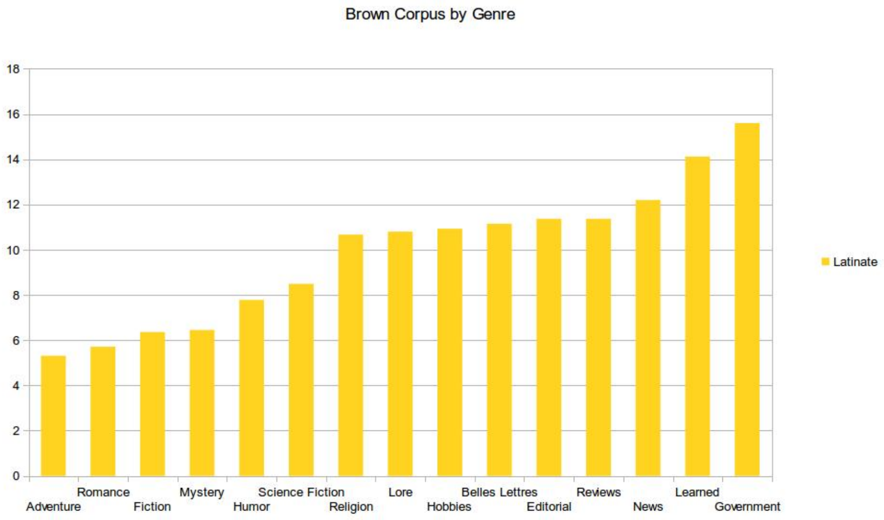

Figure 1: Brown Corpus Genres

In one such analysis, the chapters of *A Portrait of the Artist as a Young Man* were run through the Macro-Etymological Analyzer. This novel, James Joyce's Bildungsroman, is known for its style—one that mimics each progressive age of its protagonist Stephen Dedalus. Early chapters, when he is young, are written with infantile language; later chapters are written with more elevated language. The program's results quantify this stylistic mode, to some degree—Chapters 1 and 2 show low proportions of Latinate words, whereas later chapters show higher proportions, as shown here in Figure 2. The fact that the proportion of Latinate words begins to plateau starting with Chapter 3 might be used to argue that Stephen has at this young age already reached a precocious maturity of vocabulary, which may reflect his study of Latin.

In another analysis, the extracted monologues of the seven narrators of Virginia Woolf's novel *The Waves* were computed with this program. As shown in Figure 3, the two university-educated characters, Bernard and Neville, show the highest proportions of Latinate words, while the housewife Susan shows the lowest. In fact, the male characters rank higher for Latinate words than the female characters—this would be an interesting starting-point for a discussion of gender in *The Waves*, especially framed by Woolf's much-discussed writings on gender politics.

The Macro-Etymological Analyzer was also used to chart variations between editions of a text. The seven revisions of Whitman's *Leaves of Grass* made

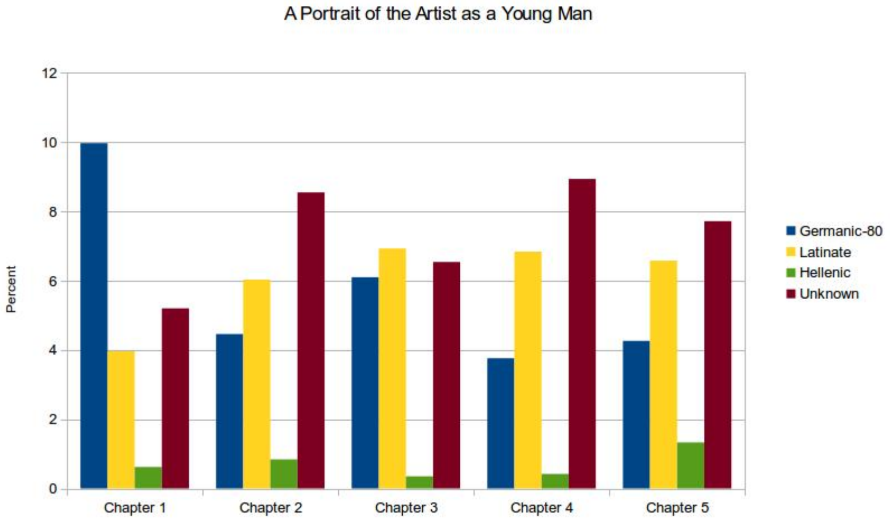

Figure 2: *A Portrait of the Artist as a Young Man*

available by the Whitman Archive were analyzed with this program. The results show a gradual increase in Latinate words from the 1855 edition to that of 1891-2. This might be used to argue that Whitman inflated his style with each revision, introduced foreign loanwords as he gained a more international reputation, or used a greater breadth of words as his vocabulary increased.

These experiments were not without their surprises, of course. An early test of selected books of the King James Bible seemed promising, as it revealed the gospels Matthew, Mark, Luke, and John to have much higher proportions of Hellenic words than other books (see Figure 4). Unlike the books of the Old Testament, which were mostly written in Hebrew, these books were translated from the Greek—a fact which might seem to explain the presence of Hellenic words. Upon closer examination, however, the program was discovered to be counting the etymology of frequently-mentioned names like "Jesus" among words of Hellenic origin, and it was these names that accounted for most of the Hellenic words. Although the language of the source text did not prove to be the determinant here, this discovery may yet be valuable for other reasons—the synoptic gospels of Matthew, Mark, and Luke show similar portions of Hellenic words, whereas that of John is 100% greater. This would seem to support the hypothesis that the synoptic gospels were adapted from a common source text, whereas that of John had an independent source.

A number of other experiments were also conducted, and are described in this paper. Included among texts analyzed by the Macro-Etymological Analyzer were: selected Canterbury tales (in modern English translation), a series of early

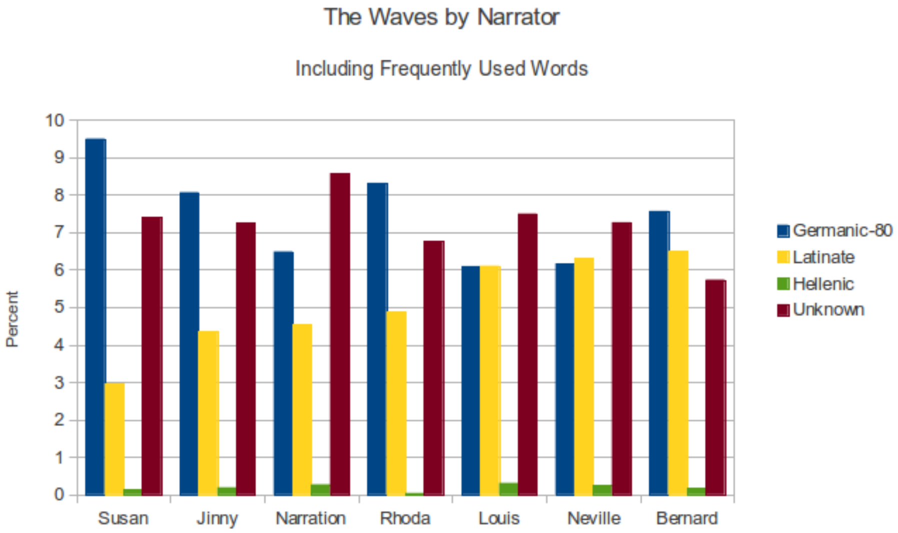

Figure 3: The Waves Narrators

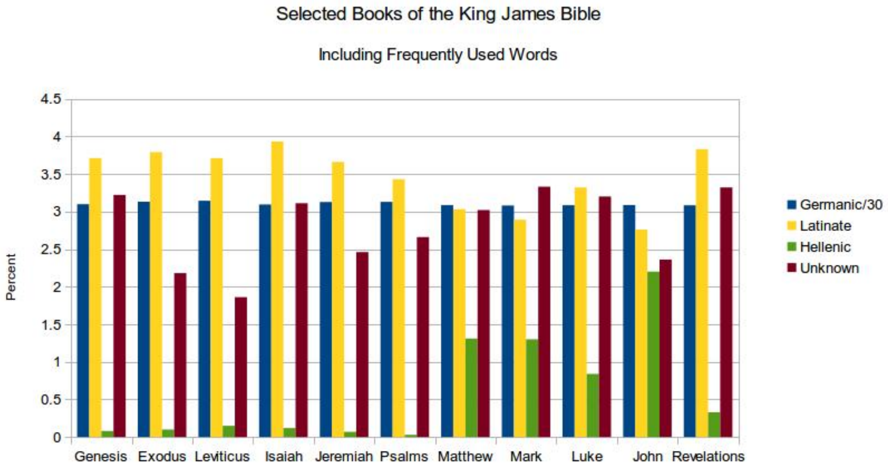

Figure 4: KJV Bible

and late Henry James novels, a collection of Victorian novels compared with a collection of modernist novels, and groups of French and German novels in English translation. Questions to be explored include:

- Do translated works show a larger-than-normal proportion of words with etymological origins in the language of the source text?
- Given a large enough data set, can linguistic trends (such as a general decrease in the use of Latinate words) be detected with this program? Can macro-historical events such as the Scientific Revolution be detected?
- Do male and female writers of the 19th century differ in the origin-types of words they use?
- Can the semantic frames in which these etymological groups of words are concentrated be explained historically, such as through the habits of the French-speaking English aristocracy in the era following the Norman Conquest?

Finally, this paper will discuss how this new tool might contribute to the suite of computational stylistics tools already available, and how macro-etymology might constitute a new metric that could be used towards stylistic fingerprinting or authorial detection.