A Macro-Etymological Analysis of *The Canterbury Tales*

Jonathan Reeve

Chaucer's *Canterbury Tales* exhibits one of the richest vocabularies of Middle English literature, a vocabulary that reveals influences from a number of native and foreign languages: Old English, French, Latin, Greek, and Hebrew, among others. While some of this foreign influence may be attributed simply to the history of the English language—the Roman Empire, the Norman conquest of 1066, and several waves of Scandinavian immigrants are each responsible for the addition of new linguistic layers—some of it is attributable to Chaucer's own knowledge of languages: as a well-traveled member of a quasi-courtier class, he had a good knowledge of European languages. Regardless of whether the poet's word choices were consciously foreign at the time of writing in the fourteenth century, however, there are etymological resonances in each word that evoke levels of formality, social class, and genre. A commonly-discussed group of words that illustrates this phenomenon is the triplet *kingly, royal*, and *regal*. The first is of Old English origin; the second, of French, and the third, of Latin. Each possesses a set of associations with their etymological registers that make these words imperfect synonyms. The following experiment will attempt to detect these registers computationally, by quantifying the etymologies of all the words of the *Canterbury Tales*.

Pre-Digital Macro-Etymology

Macro-Etymological textual analysis is nothing new. Even before what is now known as the Digital Humanities, and before its precursor, Humanities Computing, philologists have painstakingly quantified Chaucer's words and their etymological origins. One of the most recent studies, relatively speaking, is Joseph Mersand's 1939 *Chaucer's Romance Vocabulary*. Using, presumably, nothing more technically complicated than sweat and graph paper, he concludes that Chaucer's working vocabulary is 8,430 words, of which 51.8% are of "Romance sources," and 49% of Germanic sources (1939, 43). Many, if not most, of those writing about Chaucer's vocabulary after Mersand have opinions about this method. Christopher Cannon, for instance, argues that while Mersand is not factually incorrect, he would be incorrect to draw conclusions about Chaucer's personal "borrowings" from this data alone (2003, 238). Simon Horobin contends that Mersand's analysis is invalid, in its reliance on the *Oxford English Dictionary* instead of the *Middle English Dictionary*

(2007, 79), and J.D. Burnley dismisses Mersand's findings as "of little use in assessing Chaucerian style, and indeed simply ignore the crucial factor of the contemporary perception of the status of the words" (1983, 135). These dismissals, although annoyingly grouchy in tone, all make fair points, which I hope to address in my algorithmic design below.

Mersand is hardly an innovator in Chaucerian macro-etymological analysis, however, as he himself readily admits. In fact, a full chapter of his book is devoted to early quantitative analyses of Chaucer's vocabulary and its etymologies. Among these are a study by George Marsh in 1859, Alexander Ellis in 1869, and John Wisse in 1878. Mersand even expresses surprise that no "no definite numerical investigation was made before 1850," despite allusions to Chaucer's etymologies as early as 1400 (1939, 21). Each of these previous analyses, however, have methodological problems that Marsand hopes to correct, as I hope to use computational methods to correct and expand upon his.

Methods

The Text

Since this analysis relies on tools of natural language processing that are best suited to modern English, I used an edition that regularizes and modernizes Chaucer's spelling: Project Gutenberg's e-text of D. Laing Purves's 1870s edition, made "for popular perusal" (Chaucer 2000). It is neither a complete translation, which would change the etymologies of many words, nor even a complete spelling modernization. As Purves describes it, "where the old spelling or form seemed essential to metre, to rhyme, or meaning, no change has been attempted. But, wherever its preservation was not essential, the spelling of the monkish transcribers—for the most ardent purist must now despair of getting at the spelling of Chaucer himself—has been discarded for that of the reader's own day." I manually divided this text into prologues, tales, and epilogues, and then parsed the lines programmatically, removing glosses, footnotes, and other artifacts.

The Algorithm

The tool used for this analysis, MacroEtym, is a Python script I had written for an earlier analysis, but one which I expanded and customized greatly for this project. It uses an opinionated algorithm for determining the etymology of a word, and one that deserves a brief description. It begins by tokenizing a text using Penn Treebank conventions, then removes stopwords (common functional words like "a" and "the" that don't contribute much to the analysis), infers their parts of speech, and lemmatizes the tagged results, regularizing plurals to their singular forms, and verbs to their bare

infinitives. Then, it searches for the word in the Etymological Wordnet, a database created from parsing Wiktionary etymological data (Melo 2014).

If the word is found, but its ancestors are determined to belong to the same language as the original (for instance, modern English words composed by conjoining two other English words), or if the ancestor belongs to a "middle" language variant, like Middle English, MacroEtym searches for another ancestor. For all other words, the first ancestor is used. This logic is intended to foreground meaningful etymological resonances. This means that the few Old English words that are actually of Latin derivation, are labeled as Old English, but this might be considered a feature, since these words tend to be indistinguishable from Germanic words to the modern ear. In addition to ignoring current languages, MacroEtym also ignores affixes by default. Rather than parse a word like "automobile" as, say, 40% Greek (-*auto*) and 60% Latin (-*mobile*), it just considers these prefixes, like stopwords, to be background functions of the language, and incidental to the etymological resonance of the word, which in this case would be labeled as Latin.[1]

If a given word is not found in the Etymological Wordnet, MacroEtym will attempt a custom lemmatization of the word according to morphological patterns of Middle English. If the lemma is still not found, it will search for the word in a secondary etymological wordnet of 19,135 words of Middle English, which I created from parsing all the word forms listed in the Project Gutenberg edition of A.L. Mayhew and Walter Skeat's 1888 *A Concise Dictionary of Middle English*. If not found there, it will finally search for the word in an experimental tertiary wordnet of 38,074 words, awkwardly assembled by parsing the irregular etymological strings from a plain-text edition of the Oxford English Dictionary.[2]

Once a word is found in one of the three wordnets, it is then categorized according to language family, the biggest categories being Germanic, containing words of Old English, German, Dutch, and Scandinavian origin; Latinate, containing words of Latin, French, Italian, and Spanish origin; Hellenic, containing mostly words of Ancient Greek origin; and Semitic, containing words of Hebrew origin.

Results

Prologues and Tales

Figure 1 shows the proportions of Latinate, Germanic, Hellenic, and Semitic language families, organized by tale. Immediately noticeable here is that the scales are quite different: proportions of Germanic words fluctuate between 50 and 75 percent (these numbers are higher without stopwords removed); proportions of Latinate words fall

[1] The code for this algorithm is available in the notebook chaucer-macro-etym on the project GitHub repository.

[2] The code that parses these dictionaries may be found in the notebooks parse-CMED and parse-oed notebooks, respectively.

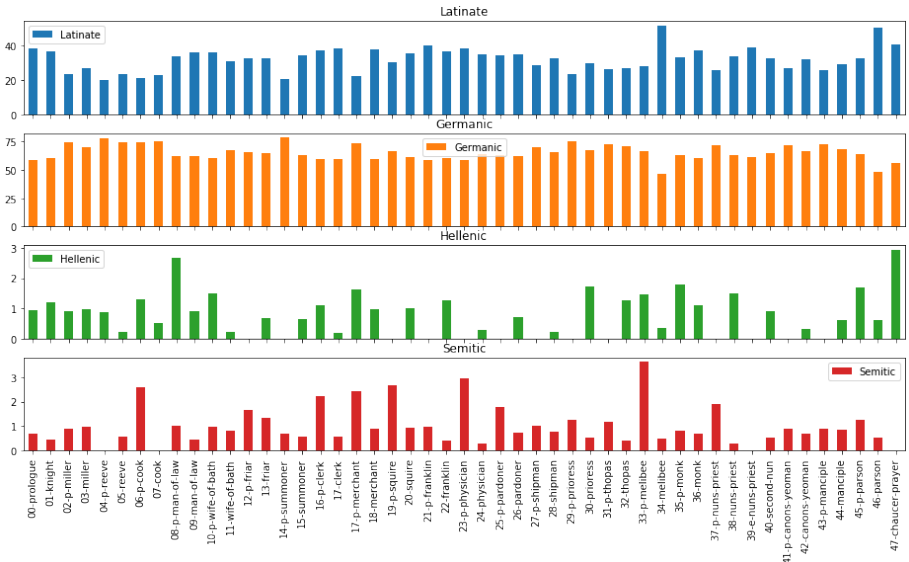

Figure 1: Language Families, by All Prologues and Tales

between 20 and 50 percent, and Hellenic and Semitic words all represent less than three percent of the total. Since some of these texts are very short—Chaucer's final "retraction," for instance, is only 369 words—we should treat the final two language families with some degree of suspicion, since in those cases, the fluctuations here reflect only the difference of about eleven words. It is for this reason that I will be focusing primarily on Latinate words here.

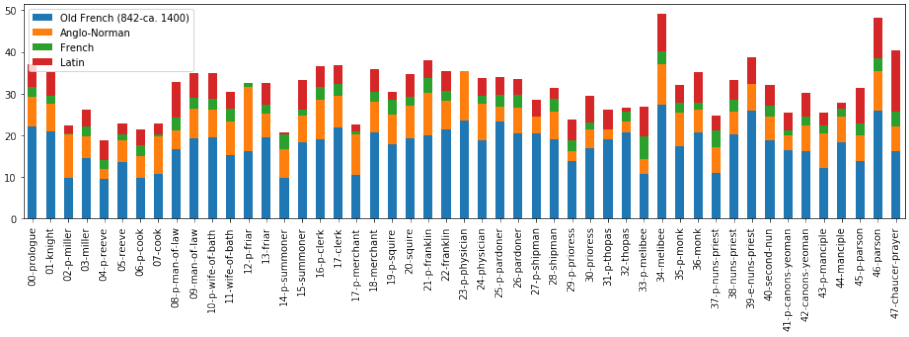

Figure 2: Latinate Words, All Prologues and Tales

Figure 2 shows the proportions of Latinate words per tale, subdivided into individual languages. Broadly speaking, the prologues and tales with the highest proportions of Latinate words are the prose tales: the Chaucer character's Tale of Melibee at 51.6%, and the Parson's Tale, at 50.33%. Chaucer's final retraction, also in prose, is at

41.0%, and the next highest proportions are in Franklin's prologue, at 40.4%, the Nun's Priest's Tale at 38.7%, and the Clerk's, at 38.5%. Yet this macro-etymological analysis is not just detecting a prose signal, as opposed to one of metered poetry, as will later become more apparent. There might be a few reasons why Latinate words appear more often in prose. First, the metrical requirements of Chaucer's line might not as easily allow many contiguous multisyllabic words, as are typical of many Latinate words; they are allowed in prose. Next, the prose modes of The Canterbury Tales are all dramatic departures from their environs, dramatic in both in intensity and theatricality, so it is fitting that their etymological modes are also contrasting.

Conversely, the tales with the lowest proportions of Latinate words are all prologues, and with the exception of the Summoner's, all from Fragment I: the Reeve's at 19.9%, the Cook's at 21.7%, the Summoner's at 21.9%, and the Miller's, at 22.3%. Unlike other prologues in the *Canterbury Tales*, these feature a high incidence of dialogue, which might be expected to have fewer Latinate words than the more constructed modes of many of the tales.

One important facet of this Latinate proportions analysis is its jagged, sawtooth shape across contiguous tales in a given fragment. The first of these patterns appears between the Knight's tale and the Miller's prologue: there is a sharp drop in proportions of Latinate words, indicating a sharp difference in tonality. The inflated tone of the Knight's tale, which ends with a happy wedding, or rather, the Latin *matrimoigne* and French *mariage* (I.3095), is quickly deflated by the Miller, who "for dronken was al pale," and who swears, using words that would be at home in (a Middle English translation of) *Beowulf*, "By armes, and by blood and bones, / I kan a noble tale for the nones, / With which I wol now quite the Knyghtes tale" (I.3125-7). The Miller's tale itself, as most tales do, shows a higher proportion of Latinate words than its prologue, but this elevation is temporary, for the quiting dynamic that the Miller inaugurates is continued with even greater force by the Reeve, whose prologue shows the lowest proportion of Latinate words in the fragment. The etymological trend within these fragments is one of Latinate oratorial floridity answered by raw, punchy Germanic talk.

Another notable quiting exchange, and one which is also observable along this etymological axis, is that between the Friar and the Summoner. There, the quiting dynamic reaches its apotheosis: the Friar's tale explicitly concerns a corrupt Summoner, who is literally a devil in disguise. And lest we believe, after this, that the satirical depictions of fellow pilgrims couldn't get any more profane, the Summoner's Prologue answers this with a scene where "out of the develes ers ther gonne dryve / Twenty thousand freres" (III.1694-5). This contrasts greatly with the Latinate moral lesson with which the Summoner ends his tale, and this contrast is clearly shown in the macro-etymological analysis.

A third pair, and one that exhibits one of the most dramatic shifts in etymological tone, is that between the Clerk's tale and the Merchant's prologue. The Clerk, who has been forewarned by the Host to "Speketh so plain at this time, … / That we may understonde what ye seye" (IV.19-20), nonetheless embarks on a verbose tale, full of verbal flourishes. The tale is set in Italy, and represents a relatively faithful translation

from the Petrarchan story *De obedienta ac fide uxoria mythologie*, a fact which may help to explain the high incidence of Latinate words (Chaucer 2008, 880). It ends with what is announced, in French, to be "Lenvoy de Chaucer," and a flurry of words of French origin: *reverence, eloquence, aventaille*, and *apparaille*, to choose a few. This high French mode is brought crashing down into the Anglo-Saxon world of bones and ale with the "murye" words of the Host, who exclaims, "By Goddes bones, Me were levere than a barel ale / My wyf at hoom had herd this legende ones!" (IV.1213-5). The Merchant echoes the Host's sentiments of marital woe in his short Prologue that follows, one which is light on invocation, but thick with casual Germanic dialogue.

When each tale is broken into eight equal segments, and each segment is etymologically analyzed, the multi-line segment with the lowest proportion of Latinate words is from the Reeve's Prologue. Here, the Reeve tells the Miller that while he may appear old, he is not weak. It shows the quiting theme hard at work, and is thus highly Anglo-Saxon:

> But if I fare as dooth an open-ers—
> That ilke fruyt is ever lenger the wers,
> Til it be roten in mullok or in stree.
> We olde men, I drede, so fare we:
> Til we be roten, kan we nat be rype;
> We hoppen alway whil that the world wol pype.
> For in oure wyl ther stiketh ever a nayl,
> To ahve a hoor heed and a grene tayl,
> As hath a leek; … (I.3871-3879)

The Reeve, in comparing himself to the medlar, uses a colorful colloquial term for the fruit, "open-erse." This is a term of decidedly Germanic origin—it is *arse,* and not the French-derived *derrière* or the Latin-derived *posterior.* The rest of the passage is literally a *pot pourri* of Germanic words, mostly from Old English. The rhetorical effect of this string of staccato Germanic is one that answers the vulgar French with which the Miller ends his tale—"Thurgh fantasie that of his vanytee"; "*par compaignye*" (I.3835, -9)—with a set of food analogies that is, like the rotten fruit itself, down to earth.

In contrast, the multi-line segment with the highest proportion of Latinate words is from Chaucer's retraction:

> Wherefore I biseke yow mekely, for the mercy of God, that ye preye for me that Crist have mercy on me and foryeve me my giltes; / and namely of my translacions and enditynges of worldly vanitees, the whiche I revoke in my retracciouns: (X.1083-4)

However much this prose passage may be an ironic advertisement for Chaucer's other works, a list of which will soon follow, it nonetheless evokes a mood of sacerdotal sincerity and rhetorical flourish, achieved in part by its Latinate vocabulary. There is a legal resonance in *enditynges, revoke* and *retracciouns,* for instance, which all enter English from Latin, by way of Old French.

When each tale is analyzed according to stanza, however, a different picture emerges.

According to this analysis, the multi-line stanza with the lowest Latinate proportion is from the Man of Law's tale, a description of Constance's journey:

> Forth gooth hir ship thrughout the narwe mouth
> Of Jubaltare and Septe, dryvynge ay
> Sometyme west, and sometyme north and south,
> And sometyme est, ful many a wery day,
> Til Christes mooder—blessed be she ay!—
> Hath shapen, thurgh hir endelees goodnesse,
> To amek an ende of al hir hevynesse. (II.946-52)

Here, the Germanic words *ship*, *mouth*, *driving*, *north*, *south*, *east*, and *west* are situated in the anaphoraic sequence "sometime…sometime" that evokes an epic mood, aggrandizing Constance's journey, and performing its "many a wery day" in what one might call a wearysome way. The monosyllabic nature of many of these words allows for the Man of Law's hyponotic meter to further enhance this effect. The Germanic words *blessed*, *endelees*, *goodnesse*, and *heavynesse* have a simple quality appropriate to an innocent and much-maligned saint character as Constance, one that constrasts sharply with the description of the Roman Senator's analogous journey that follows, where the Senator very Latinously "repaireth with victorie / To Romeward, … saillynge ful roially" (967-8).

Another such royal description appears in the multi-line stanza with the highest proportion of Latinate words, the ekphrastic description of Nebuchadnezzar, from the Monk's tale:

> The myghty trone, the precious tresor, The glorious ceptre, and roial magestee That hadde the kyng Nebugodonosor With tonge unnethe may discryved bee. He twyes wan Jerusalem the citee; The vessel of the temple he with hym ladde. At Babiloigne was his sovereyn see, In which his glorie and his delit he hadde. (VII.2143-50)

As with Constance and the Roman senator, there are contrasting Germanic and Latinate modes that coincide with innocence and experience, Christianity and classicism. Compare the passage above with the Monk's Germanic description of prelapsarian Adam as "With Goddes owene fynger wroght ws he, And nat bigeten of mannes sperme unclene" (VII 2008-9). As mentioned earlier, there is a certain gravitas to the French word *royal* that is not as pronounced as in an equivalent word of Old English descent, such as *kingly*, just as the word *gravitas* itself sounds more serious than Constance's *hevynesse*. Chaucer doesn't use these modes unironically, however, for the Monk's historical histrionics is soon negated almost bathotically by the Knight, who, though guilty of excessive verbosity himself, pleads to have "namoore of this," the Monk's cast of characters (VIII.3957).

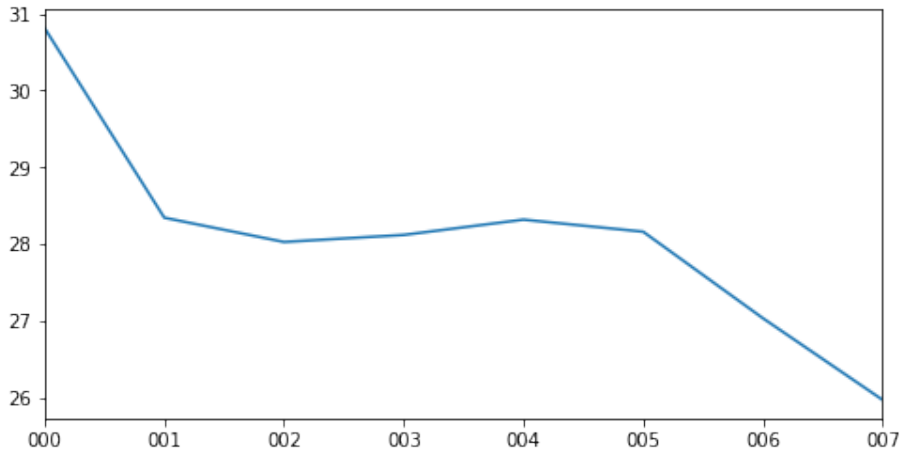

Figure 3: Mean Latinate Proportions Across Segments of Prologues and Tales

Macro-Etymology of the Individual Prologue/Tale

So far we have discussed the macro-etymologies of tales and prologues, but what might we discover about the macro-etymologies of the various parts of individual tales? To answer to this question, I divided each prologue and tale into eight equally-sized parts, and ran MacroEtym across each of them. Somewhat surprisingly, a fairly consistent trend may be seen across the narrative time of each tale, no matter the tale. Figure 3 shows the percentages of Latinate words per segment, averaged across all tales and prologues. On average, the first eighth of each tale shows roughly six percent more Latinate words than the last eighth, with intermediary tales showing a gradual falling gradient. There may be a number of possible explanations for this phenomenon, but my theory is that Latinate words are most likely to appear in descriptions: descriptions of characters, scenes, and ideas, which are most likely to fall at the beginning of a tale. Prayers and invocations, as well, which are high in Latinate words, happen more often, and for longer stretches of time, at the beginnings of tales. Other narrative elements, like dialogue, are statistically much lower in Latinate words, and are more likely to fall in the middle or end of a tale.

Conclusions

While this study is by no means novel, and has been preceded by centuries of analogous macro-etymological analysis, the narrative explanation of these trends, one that situates them among the pilgrims' interpersonal dynamics—departs from prior philological methods that have explained these trends in terms of Chaucer's sources, personal vocabulary, or educational history. To summarize, I have found that:

1. Sudden shifts in etymological register, along the Latinate axis, at least, are coincident with the sudden shifts in tone that accompany the "quiting" interchanges among pilgrims.
2. Prose tales show much higher proportions of Latinate words than verse tales.
3. In general, tales exhibit roughly 50% more Latinate words than prologues.
4. When divided into segments, the average trend across tales is a drop in the use of Latinate words. This is more pronounced in tales than in prologues.

Works Cited

Burnley, J. D. 1983. *A Guide to Chaucer's Language*. London? Macmillan.

Cannon, Christopher. 2003. "Chaucer's Style." In *The Cambridge Companion to Chaucer*, edited by Piero Boitani and Jill Mann, 2nd ed. Cambridge Companions to Literature. Cambridge, U.K. ; New York: Cambridge University Press.

Chaucer, Geoffrey. 2000. *The Canterbury Tales, and Other Poems*. Edited by David Laing Purves. http://www.gutenberg.org/ebooks/2383.

———. 2008. *The Riverside Chaucer*. Edited by Benson. Oxford University Press.

Horobin, Simon. 2007. *Chaucer's Language*. New York: Palgrave Macmillan.

Melo, Gerard de. 2014. "Etymological Wordnet: Tracing the History of Words." In *Proceedings of the 9th Language Resources and Evaluation Conference (LREC 2014)*. Paris, France: ELRA.

Mersand, Joseph E. 1939. *Chaucer's Romance Vocabulary*. New York: Comet Press.