Big words: English speakers' difficulty with the Latinate wordstock in English

Department of linguistics, Grinnell College

May 16, 2011

For author details please contact comment@chuckmasterson.com

Abstract: *Over half of the lexicon of English has been borrowed into the language from Latin or Romance languages, or built using word roots that come from those languages, rather than being created from word roots that have come to the modern lexicon straight from Old English. However, words from Old English are still the most used words in English, and there are several morphological reasons to expect that English speakers may be able to analyze them to figure out their meanings more easily than they can for Latinate words. This study uses a test to investigate whether this is the case. The test asks participants to break down very rare but analyzable words from the native English and Latinate wordstocks into their component morphemes and give what they think are the meanings of these words. It is concluded that speakers find Latinate words harder to analyze. The implications of this finding for education and academic writing are discussed. The concept of "reminding" words is proposed to explain that Latinate words are also likely harder for speakers to remember even after the first encounter, since their surface forms less often remind speakers of their meanings, as opposed to the more transparent native English words.*

Introduction

This study investigates the differences between English speakers' understanding of English words that derive from Latin and those that have entered the present lexicon from Old English. Specifically, it addresses the question of whether English speakers find Latinate words harder to understand than those from Old English. When a person hears or reads an unfamiliar word, or even one that has simply passed out of ready memory, that person must necessarily rely on clues from context and the form of the word to decide what to make of the word. In this study I focus on the form of the word. There are several important reasons, detailed further on, to expect that it might be more difficult for English speakers to analyze the forms of Latinate words than of words from Old English. The goal of this study was to discover whether, and to what extent, Latinate words present a stumbling block for those reading (or speaking) English. The conclusion I have reached is that Latinate words are indeed more difficult than native words for English speakers to understand.

Historical background

Beginnings of the language. English is a member of the Germanic family of languages, which also includes German, Icelandic, Norwegian, and others. All these languages developed from the unattested but reconstructed proto-Germanic language (Algeo 2006:62). The core vocabulary of English, its commonest words, are still those inherited from proto-Germanic

through Old English (also known as Anglo-Saxon), or built more recently with Old English roots. Most of the basic English function words are such words (*the*, *what*, *with*), and so is a sizeable amount of our vocabulary (for example, *work*, *understand*, *download*). However, English, owing to historical circumstances, has adopted an unusually large number of words from Latin and the languages that developed from it (mainly French). Finkenstaedt & Wolff (1973) estimated the total proportion of Latinate words in English at 56.56%.[1] In comparison, they estimated the proportion of native English words at 21.66% (excluding those from other Germanic languages). However, a study by Jones & Wepman (1961, cited in Hughes 2000:35) of natural conversation found that native words accounted very heavily for the most commonly used words: "Of the 200 most commonly used words, 83.5 per cent were Anglo-Saxon, 4.5 per cent were Old Norse (the closest Germanic relative), 10 per cent were from Latin via Old French and the remaining 1.5 per cent were from post-medieval Latin borrowings."

History of borrowing words. The English language has borrowed words from Latin and its daughter languages since before its speakers first inhabited the British isles, and even before it split off from Proto-Germanic (Baugh & Cable 2002:§57). The rate of borrowing first sped up around the year 597. At this point the Romans, who had settled Britain in the early years A.D., had left Britain (in 410) and the Germanic tribes who would become the British people had settled there (in 449). The year 597 marked when returning Romans first began to Christianize those living in Britain (*ibidem* §§60–62). Some of these very old Latin borrowings are *ounce*, *street*, and *pear*, and, from the domain of religion, *priest* and *shrine* (*ibidem*). However, the largest English absorption of Latinate vocabulary did not begin until a Norman French duke, relative of the British royal family, conquered England in 1066 at the Battle of Hastings to become King William I. Under his and subsequent reigns, French became the language of prestige and the upper classes in England. English first began to gain French-derived vocabulary in matters of the court and the church, domains of this upper class; hence, *clergy*, *court*. Later, French words began to enter English much more generally. From the 1200s to the 1400s, English monarchs became first bilingual in English and French and finally monolingual in English, but the French vocabulary remained, especially among the upper classes.

The use of French vocabulary was a class-based and education-based phenomenon in its beginnings; Corson (1995) argues that it still is. Corson points to the upper-class practice in the Middle Ages, and continuing well into the modern ages, of teaching children Latin. Latin served for a long time as the common language of academic writing, even after no one spoke it natively. High-class British parents had their children educated in Latin so they would be able to read and write in this academic language. Indeed, from around 1300 to 1500, many schools insisted that children speak *only* Latin, to the exclusion of the vernacular, and taught children composition only in Latin (Corson 1995:62–64). Thus Latin continued in its use as the language of academia. When writing in the vernacular finally did become accepted and common, those writing academically in English were initially still those who had been educated in Latin as children, and as Corson surmises (*ibidem* 69–70), they were likely to write

[1] This proportion combines the following labels that Finkenstaedt & Wolff used from the *Shorter OED*: Latin, French, Old French, and "Anglo-French," which Finkenstaedt & Wolff found to be ill-defined by the *Shorter OED* editors but most comparable to Old French.

with and create Latinate words—adapting them into English on the model of French-derived words that had already been assimilated and used for a long time by then (see Jespersen 1923:§114)—rather than to translate the Latinate vocabulary of their mental lexicons into native Germanic equivalents.

The Saxonists and antisaxonists. The bringing in of all this foreign vocabulary did not go unnoticed, and through the centuries the Latinate lexicon has had its supporters and detractors. Baron (1982) details several centuries of back-and-forth in public dialog on rhetoric and composition between what he calls the Saxonists—those who wished to limit the foreign incursions into the English vocabulary—and the antisaxonists, their gainsayers.

The Saxonists began to emerge in the 1500s in reaction to the new prevalence of "inkhorn" words, ones that were invented in writing rather than arising from the vernacular. In 1557 for example, John Cheke, a Cambridge scholar, wrote, "I am of this opinion that our own tung shold be written cleane and pure, vnmixt and vnmangeled with borrowing of other tunges" (*ibidem* 9–10). Some of the Saxonists' thoughts on the unsuitability of Lte[2] words coincide with explanations that I will offer in the next section for the separateness of the Anglo-Saxon and Latinate wordstocks. Baron quotes philosopher Herbert Spencer writing in 1873 that "the earliest learnt and oftenest used words [that is to say, the heavily Anglo-Saxon vocabulary of children], will, other things equal, call up images with less loss of time and energy than their later learnt synonyms" (*ibidem* 48). Richard Grant White wrote in 1870:

> Petroleum means merely rock oil. In it the two corresponding Latin words, *petra* and *oleum*, are only put together; and we, most of us, use the compound without knowing what it means. Now, there is no good reason, or semblance of one, why we should use a pure Latin compound of four syllables to express that which is better expressed in an English one of two. . . . The power to form [compound] words is an element of wealth and strength in a language; and every word got up for the occasion out of the Latin or the Greek lexicon, when a possible English compound would serve the same purpose, is a standing but unjust reproach to the language—a false imputation of both weakness and inflexibility. (Baron 1982:49)

In questioning the (aesthetic or symbolic) value of words that most cannot analyze, White prefigures this study's idea of remindingness, which will be discussed further on. The Saxonist impulse, incidentally, is still alive today, as evidenced by the "Anglish Moot" wiki ("What is Anglish?"), which invites anyone to try their hand at writing English without foreign-derived words.

The viewpoints of the antisaxonists, meanwhile, seem generally to come to the idea that English vocabulary before the addition of French was a poor one, incapable of much breadth of expression, and that the added words give the language as a whole a new clarity. One commentator, James Champlin Fernald, wrote (in 1919): "We have done better to borrow. . . . Instead of painfully piling home-grown syllables upon each other or jamming words together under hydraulic pressure of thought, we may simply reach out and raid the universe of speech" (Baron 1982:53). Another, Thomas DeQuincey (in 1839), thought the Anglo-Saxon

[2] In this study the following abbreviations will be used: L—Latin; Lte—Latinate; AS—Anglo-Saxon; *OED*—*Oxford English Dictionary*; *W3*—*Webster's Third New International Dictionary of the English Language*; COCA—Corpus of Contemporary American English. The < sign denotes "from" in etymologies.

vocabulary to have only six to eight hundred words, "most of which express some idea in close relation to the state of war," useful only to express "simple narration, and a pathos resting upon artless circumstances,—elementary feelings,—homely and household affectations" (*ibidem* 25), implying that more complicated times required a more complicated language.

Present state of new words. Despite the Saxonists' spirited opposition to the process, English has gone on borrowing Lte and Greek words. Currently AS words are still a definite minority among new coinages, as a sampling of them shows. I randomly selected 100 words from the Addenda to *W3*, which consists of words added to the English lexicon (or at least newly discovered by Webster's) between the 1961 publication of *W3* and the printing of the 1993 edition. I excluded listings indicated as new senses of words that were already listed in the main body of the dictionary on the grounds that they did not represent new coinages, but otherwise I counted all words, including open compounds. I tallied the origins of these words; for words that had roots from more than one language, I counted fractions. Among these words, 40.5 were Latinate (33 directly from Latin, 7.5 by way of Romance languages), 14.8 were Greek, 23 were Anglo-Saxon, 8 came from proper names, and 13.7 came from various other languages.[3] Moreover, the AS words tended to come from popular culture (*ragtop, far-out, hot tub*), whereas the Lte words (especially those directly from Latin) and Greek words tended to express academic concepts (Lte *matrix sentence, supercontinent, conditional probability*; Greek *tracheoesophageal, xenogeneic*)—although this was by no means universally true (AS *black hole, grow plug*; L *sensitivity training, proactive*; Romance *bustier, adobo*).

As Corson (1995:102–104) shows, the vocabulary of academic writing today is indeed especially Latinate, although some domains (especially the natural sciences) also use a great deal of vocabulary derived from Greek (*ibidem* 67). By this point the reasons for coining new academic words with Latin or Greek roots may be largely a matter of convention, scarcely questioned. That is, those who coin new academic words coin them from Latin or Greek roots simply because that is how so many other academic words have been coined, and it seems right to coin words that fit in with the already existing vocabulary. This is analogous with the process of importing English musical terms wholesale from Italian, or ballet terms from French, on the basis of historical precedent. Certainly there is something of this going on in the case of words that are formed according to definite, pre-existing patterns: for example, forming the names of new branches of study using a Greek root plus the Greek form *-ology*, or likewise creating names of new clinical fears with Greek *-phobia*.[4] The process may be influenced by other factors as well, such as, perhaps, the explicit Latin of the binomial species classification in biology.

Ultimately, it will be the staying power of these systems of wordbuilding that will determine the direction of the English lexicon, and the wishes of those who believe the English vocabulary should be reformed will probably have a rather small influence, if any. In the past, efforts to Saxonize English have met mainly with failure, outside a few successes like

[3] Arabic, Cantonese, German, Hindi, Mandarin, Old Norse, Scottish Gaelic, Tamil, Yiddish, unknown sources, and the ahistorically made-up morphemes of the International Scientific Vocabulary.

[4] I am discussing Greek roots here under the supposition that they likely behave much in the same ways as Latinate roots, owing to similar historical circumstances.

the introduction of the words *foreword* for *preface* and *handbook* for *manual* (the latter one a revival of a word that had long been out of fashion) (Baron 1982:59). Baron points to the contrived nature of Saxonists' suggested replacement words to explain some of their failure. Looking to the Saxonists' suggestions, we do find such strange- or archaic-sounding forms as *statespellman* 'ambassador' (Baron 1982:35) and *throwfaresom* 'penetrable' (*ibidem* 20). But perhaps the main reason that new AS words have had difficulty gaining ground over established Lte words is precisely that the Lte words are established in their niches (even while AS words are more established overall), and it is hard to change entrenched patterns of language use, especially for a single person or small group trying to change all speakers' habits, no matter the reason and the motivation. Thus we can account for the failure of AS replacement words that seem less strange and more practical, such as *wheelman* 'cyclist' or *self-working* 'automatic' (*ibidem* 47).

Latinate wordstock as separate from native wordstock

Conceptions of the separation. Because the bulk of Latinate words came as comparatively recent grafts onto the native English wordstock that goes back to before written history, and because they have historically been used mainly in the domain of academic thought, it seems quite possible that these words would be harder for most English speakers to understand. Several popular and academic authors (Bryson 1990:75, Jespersen 1923:§§128–150, Orwell 1999 (1950), Strunk & White 2000:76–77) have written of the differences between Lte and AS words, saying variously that Latinate words are generally longer, harder to understand, and less "homely-sounding" (Corson 1995:1, citing Quirk 1974) than Anglo-Saxon words. It seems as though Latinate vocabulary, after so long in English, could have become a naturalized, unnoticed part of the language. However, these writers all thought it peculiar that English should have so Latinate a vocabulary, and thought of the Lte wordstock as a distinct-feeling section of English. Maylath (1996) notes that "English speakers commonly refer to Greco-Latinate words as 'hard words' or 'big words' (even in cases where they consist of as few as two syllables)." The ongoing seclusion of the Latinate vocabulary in specialist systems of meaning, as detailed above, is likely one factor in the otherness that continues to be associated with Latinate words.

Another likely factor is that the Latinate words came into English later and have not had as much time to become thoroughly established. Corson connects this idea to Luigi Meneghello's notion of Italian words that are "*le ferite antiche che rimarginandosi hanno fatto queste croste delle parole in dialetto*" (ancient wounds that, in healing, have left these scars that are words from the dialect)—as opposed to recent loans, which are "*le ferite superficiali*" (flesh wounds) (Corson 1995:56, citing Meneghello 1963). What this metaphor means in a more technical sense will be explored below.

Formation of Latinate words. We can actually see that some of the earliest Lte loanwords, what Meneghello might call older wounds—like *ounce*, *street*, *pear* mentioned earlier, and also including *mill*, *pit*, *wall*, and *butter*—are short (mostly one-syllable) and similar in syllable structure to native English words (compare *butter* and AS *water*, or *street* and AS *strong*;

contrast the L words these came from, *butyrum*[5] and *strata* (*OED*: butter, *n.*; street, *n.*)). These words also do not take Latinate affixes; in some cases, new words have been built with them using native English affixes: *waller*, *buttery*. In contrast, newer Lte words in English bring with them a whole distinct paradigm of forming new words. To take just one example, the *-ion* noun-making suffix attaches most readily to Lte verbs ending in a verbal ending such as *-(a)te* (*relation*, *creation*, *completion*), and combines with AS verbs only rarely and in an adapted form *-ation*: *starvation*, *flirtation* (*OED:* -ation, *suffix*). Cutler (1981) finds (in England) that English speakers tend to prefer word affixes that do not change the pronunciation of the base word. Suffixes such as *-ity* and *-ify*—both Lte—change the stress pattern of words attached to them that do not have last-syllable stress, and Cutler found that English speakers judged neologisms with these suffixes less acceptable than suffixes that do not affect stress, like *-ness* (AS) or *-ment* (Lte). This may help to account for the great success of the stress-neutral, Lte suffix *-able*, although its hard-to-deny usefulness and the lack of an AS equivalent must also be considered.

Taken as a whole, Lte affixes tend to blend with and change root words more than AS affixes. For an extreme example, consider the Lte prefix *ad-* 'to,' which assimilates, depending on the beginning of the word to which it is added, to any of the following: *a-*(scend), *ac-*(cumulate), *af-*(fix), *ag-*(grieve), *al-*(lude), *an-*(nounce), *ap-*(pend), *ar-*(range), *as-*(semble), *at-*(tend) (*OED:* ad-, *prefix*). Similarly *con-* assimilates to *col-*(lapse), *com-*(pare), *cor-*(relate). The suffix *-ion* commonly alters the last phoneme of its base word: *create–creation*, *collide–collision*. Latin roots are also commonly reduced to their stem when a combining form is added to them: *local* < L *loc-us* + *-al*. This is very uncommon among AS word roots, which have no inflectional endings in the dictionary (singular) form.

Meanwhile, AS suffixes such as *-y*, *-ly*, and *-ness*, and prefixes like *fore-* and *over-*, tend to leave their base words unchanged (see Bar-Ilan & Berman 2007). The suffix *-y* can be added to a wide variety of shapes of words: *folksy*, *clayey*, *watery*. Regularly conjugated English verbs (of the "weak" *-ed* type) never change their stress from the infinitive. Regularly conjugated verbs in Spanish (descended from Latin) always do. This should be noted in relation to the fact that in English, vowels are very often reduced to schwas when they are found in unstressed syllables, whereas in Spanish all vowels are given their full value regardless of the stress. Thus one would expect that a stress shift in Spanish, or another language without reduced vowels, would not be as drastic a shift as in English, where not only the stress but the phonemes are changed (see Jespersen 1923:§26).

More than just the affixes associated with Lte words, however, it seems reasonable to conjecture that the word roots themselves could be less understandable to speakers of English, because while Lte roots form the bases of many words, if these same roots exist as free morphemes, these free forms are liable to be rarer than the AS equivalent. Their relative rarity would be expected to come as a result of the morpheme's late addition to the language, in two ways. The following discussion of these ways gives a more detailed idea of what, I believe, is meant by Meneghelli's "*ferite antiche*," and also by Jespersen's notion of "those invisible threads that knit words together in the human mind" (*ibidem*:§143).

[5] From an earlier Greek word *βούτυρον boútyron*.

Absence of Latinate morphemes in root form. First, since Old English was a living language, one would expect that it had words for talking about most things important to its speakers before it began to absorb more words from French and Latin. Thus, the Lte words would be borrowed instead to cover (more or less concrete) new concepts arriving with the new society's meaning system (*confession*, *cleric*, *jury*) or more abstract concepts that had not so far been lexicalized (*essence*, *existence*—see Boas 66). The words used to express the abstract concepts would, because of the greater complexity of the ideas presented, be more likely to be built of several morphemes, and the base morpheme of any of these words would be less likely to enter English on its own. For instance, *essence* has in it the L word *esse* 'to be'; however, this root exists in English only in this word, in a few others such as *essential*, and in rare phrases like *in esse* (usually thought of not as English but as quoted L phrases, as evidenced by the italicization of most of its citations in the *OED* (*OED*: esse, *n.*)). Where a base form was adopted, it would compete with an already established AS word if such a word existed, and possibly be relegated to a more specific, less common usage. So, for example, we see that, though we have the Lte words *describe* (lit. 'write about'), *ascribe* (lit. 'write to') and some other relatives, the word *scribe* has certainly not replaced the common, all-purpose *write*; instead it has gained a much more specific meaning, generally that of 'one who hand-copies a document.' Of course it is not universally true that the Lte word remains the more uncommon in such a pairing. The Lte word *place*, for example, has more or less completely taken the place of the AS *stead*, and Lte *move* is so thoroughly engrained in English that no AS equivalent remains.

Less interconnectedness through recency of addition. Second, Latinate words may be felt as foreign by speakers simply because they have not had as much time to make connections with other words in the language and to become established in common phrases. This factor is harder to precisely describe and quantify than the previous one, but a brief study can be done through an examination of Partridge's *Dictionary of Clichés.* Such an examination suggests that the vocabulary used in our most common phrases has more AS words than Lte words in it. I analyzed the vocabulary of this dictionary by looking at the last cliché on every other page and tallying the word origins of each word involved (excluding phrases borrowed wholesale from other languages such as *cherchez la femme!* and *sub rosa*). Chart 1 shows the counts for the various kinds of words.

Chart 1: Origins of words in Partridge's Dictionary of Clichés

	AS	Lte	Greek	other[a]	non-lexical[b]
count	182	83	5	12	153
%	41.8	19.1	1.1	2.8	35.2
% of lexical	64.5	29.4	1.8	4.3	—

a Including Norse-derived words and proper names.

b I counted non-lexical words separately, defining them as prepositions, pronouns, modal verbs, and the copula *to be*—cf. Berman 2007:15.

These percentages are remarkably different from the proportions of different word origins present in the English lexicon, as set out by Finkenstaedt & Wolff (1973): 21.66% AS (and 2.90% other Germanic), 56.56% Lte, 5.3% Greek, and 13.57% other. This discrepancy suggests that the words that English has had longest have had the most chance to cement themselves in turns of phrase and become well known by those who speak the language.

To look at these factors more concretely, I set out in this study to use dictionary-accepted English words of these two backgrounds to find out whether English speakers do in fact find Latinate words harder to understand than Anglo-Saxon words.

Literature Review

This project is situated among linguistic studies of English in use, some of which approach the matter in question from the standpoint of educational theory. The study also draws heavily on historical linguistics of English, much of which was reviewed in the previous section.

There has been a fair amount of scholarship dealing with the interplay between Lte and AS words. Much of it looks at the issue from the perspective of language acquisition by children who are learning English as either a first or a second language. Corson (1997), especially, studied the issue of "academic" words from this language acquisition perspective. Corson works from the finding that children with less education are less likely to use Greco-Latin words in their writing and speech, and in the end concludes that this is because their background does not expose them to these sorts of words enough for them to absorb them. He also finds that children who read more have a broader vocabulary, especially with Greco-Latin words, because of the much greater frequency of these words in print. (He cites Hayes and Ahrens (1988), who found that "even children's books contained 50% more rare words than either adult prime-time television or the conversations of university graduates.") This research agrees more or less with the previous discussion of the mechanisms that could make Lte words less understandable. Corson is proposing that the rarity of Lte words makes them more difficult. I propose a similar mechanism, but on the level of morphemes instead of words, explaining why English speakers might find Lte more daunting than AS words even if they have seen neither before. Though certain Lte morphemes may be quite common, I propose that these morphemes too might be more difficult by virtue of their rarity *as free morphemes*.

Speakers' morphological knowledge. At this point, some background is needed on speakers' morphological knowledge: how much do people actually break down words when they read them or hear them? Most of the research in this area has been with children. Anglin (1993) details an especially applicable study. He did a study on children's ability to define words randomly chosen from an unabridged dictionary, and found that they are able to provide reasonable definitions for many words that are composed of other, known morphemes, such as *unbribable*, *readmission*, *cardinal flower* (a very red flower) and *western saddle* (the kind cowboys use). The children had presumably never encountered these words before, and in some cases they openly stated so. Thus they must have arrived at the definitions solely by analyzing the

words. Anglin calls their process of guessing the definition of a word *morphological problem solving*. The children's ability to use this skill results in an effective passive vocabulary that is higher than would otherwise be expected and includes words that are completely novel to them but explicable because of their morphology.

White, Power, & White found similar results in their study of whether children can analyze words that have affixes, and also used it to estimate what proportion of new words children could guess through morphological problem solving. They estimated that, without help from context, third-graders could guess the meaning of 12% of unfamiliar words and fourth-graders could guess 27%. They attributed the large amount of vocabulary growth during the school years to children's ability to figure words out this way; so did Anglin. Both studies found that children's ability with morphological problem solving improves over time, and Anglin suggests the improvement is drastic between third and fifth grade.

Studies of the divide between the wordstocks. A study that deals more specifically with the AS–Lte split, and also with adults, was done by Bergman, Hudson, & Eling (1988). They used several words in a *lexical decision* task: a person is shown a string of letters and reacts as quickly as possible to decide whether it is a word. In this way they can study how long it takes people to activate words in their mental lexicons based on how many milliseconds the subject's judgement takes. The study was run on a Dutch-speaking population with Dutch words. Dutch, another Germanic language, has somewhat the same interaction between Germanic and Latinate words that English does, although the researchers implied that Latinate roots in Dutch are (perhaps much) less productive than they are in English, describing the Latin roots generally as "moribund." They found that Lte words had higher reaction times than those based on the native Germanic wordstock. They imputed this difference to the rarity and unproductiveness of Latinate morphemes in Dutch. One would expect the dynamic between the two kinds of words to be somewhat different in English, where Lte words are constantly being coined and added. Differently from this study, Bergman, Hudson, & Eling studied only whether subjects *recognized* words, and did not deal with whether the subjects knew the words' meanings. Their research thus bears more on the immersiveness and speed of reading than on readers' understanding of what they are reading, the area of my interest. Other research in the area of how people break up compound words has also involved measuring in various ways the time it takes people to process words (*e.g.* Frisson *et al.* 2008), but focuses more on the neurological aspects of the issue, which I do not treat here.

Readers' understanding of Lte words is particularly interesting because of the effect that hard-to-understand words would have on certain varieties of English. As mentioned before, academic English is more concentratedly Latinate than any other kinds of English. Bar-Ilan & Berman (2007) studied the AS–Lte divide in more detail. They had children and adults in four different age groups produce texts in four different modes: spoken or written, and expository or narrative. These modes were ranked according to the level of the register that the researchers took them to represent—with registers being defined as "the linguistic difference that correlates with different occasions of use" (*ibidem*, citing Ferguson 1994:16). From lowest to highest they were: spoken narrative, spoken expository, written narrative, written expository. As they expected, they found that the proportion of Lte words rose as the

register level rose. The prominence of Lte words in higher, more educated registers, and their relative absence in lower registers, means that even beyond the other difficulties of learning a higher register—for example, more structuring of language, and learning what are and are not suitable subjects (see Irvine 1979)—English speakers also face a very different vocabulary that may seem, quite literally, foreign to them.

Corson, to illustrate a different possible situation, asserts that for Greek speakers in ancient times the "simplicity [of the words used in learned discourse] had been the hinge of Hellenic intellectual progress, since for the Greeks the vocabulary of learned discourse was the vocabulary of ordinary language reapplied. The words were available for use by all the Greeks, yielding direct popular access to learned discourse with no hint of a 'lexical bar' like that for many users of English" (1997:66). Though this claim perhaps attributes more importance to vocabulary than merited, one wonders whether English speakers might feel differently about higher education were academic vocabulary as straightforward in English as Corson claims it was in ancient Greek. Such a vocabulary is certainly not a linguistic impossibility, and indeed some modern languages abstain from borrowing foreign words to describe new or complex concepts. Icelandic has historically formed new words out of native elements, regardless of the form of the word in other languages, leading to constructions such as *sjónvarp* 'television' < *sjón* 'sight' + *varp* 'throw, project,' or *ljóstillífun* 'photosynthesis' < *ljós* 'light' + *til* 'to' + *lífun* 'life, survival.' As we have seen, over the history of English, Saxonists have made many attempts to rid the language of foreign influences and make English more like these languages. But for all their enthusiasm, these attempts have not gained appreciable ground against the widely accepted, conventionalized Lte vocabulary, and must generally be taken as hypothetical only.

Place of this research. I have not found studies that mix the two foregoing fairly distinct lines of research—investigation of speakers' morphological knowledge, and of the AS–Lte split in English. Either studies have looked at each word as a whole (a whole that is either Lte or AS), or they have looked at how speakers use morphological problem solving on English words of any type. However, it seems important to look at how English speakers might differently analyze, for example, *foretell* and *predict*. These are the elements that this study brings together, by investigating how people break down words of each of these two kinds.

Moreover, this study focuses on adults rather than on children still developing linguistic competence. Adults are who most language is directed to, and who will be most likely to encounter and interpret the obscurest of English's Lte vocabulary. However, Bergman, Hudson, & Eling's (1989) and Bar-Ilan & Berman's (2007) are the only two studies to have dealt with adults that I found, besides Maylath's (1996), which is a bit different from the rest in that it deals with readers' opinions about different lexicons rather than understanding of them. Following Corson's (1995, 1997) concentration on education, I do consider level (and kind) of education an important factor in whether a speaker can effectively interpret Lte words, but I look not at education during childhood but rather at the degree to which people have been initiated, through a college education, into the specialized, academic meaning systems to which the Lte words are endemic.

Corson's various studies (1995, 1997) are based in educational theory, and he offers a great deal of advice for the teaching of English. Bergman, Hudson, & Eling were studying

how complex words are encoded in the mental lexicon; White, Power, & White focused on the ways in which vocabulary can be broadened by speakers making connections among words. All of these are issues that this study is relevant to, but the latter is one of the more prominent.

Anglin's study (1993) was the most methodologically similar to mine. In one part, children were asked verbally what they thought a word meant and pressed for clarification until they got the definition right or made it clear that they could not. This is similar to the open-ended answer form that I have used with my survey form, and in analyzing my results I have used some of the same methods Anglin used. I drew some methodological ideas from his study, as detailed in the methods section. One major influence from Anglin's study that I will deal with here is that it led me to draw words from a dictionary rather than from a corpus of language in use. I originally was concerned with the kinds of words that English speakers might encounter in texts they would be likely to read, but I ended up concluding with Anglin that it would be desirable to investigate speakers' competence with the language *as a whole.* Using words from an unabridged dictionary was the most satisfactory way to find a sample that represented the entire language.

From the foregoing, what we know based on the literature is that people are able to do morphological problem solving on unfamiliar words, and that Lte words form a class of words that are used in higher, less familiar registers. These are the factors that lead to the question of how difficult it may be for English speakers to analyze and understand Lte words when they are used.

Methods

Drawing the sample of words. To obtain a list of words on which to test the study participants, I drew from *W3*. This dictionary has been chosen by previous researchers (Anglin 1993, Dupuy 1974, Goulden *et al.* 1990) and data are available on the kinds of words it contains (Anglin 1993) and the number of words (Anglin 1993, Dupuy 1974). One of the main reasons cited by Anglin for his choice of *W3* is that it is the largest unabridged, synchronic (that is, non-historical) dictionary of the English language. This means that a random sample drawn from it will be composed wholly or mostly of words that are encountered in the sorts of present-day texts that the survey respondents are likely to have encountered.

To avoid biasing the results toward any particular type of word, I used a random sampling method to draw words from the dictionary. Adapting the method used by Anglin (1993), I recorded the fifth-to-last entry on every other page of the dictionary, with an "entry" being any word flush to the left margin. This gave me a raw sample of 1,359 words.

The kinds of words represented in this sample are important. Firstly, I focused on rare words that would likely be novel to the participants, so they would have to rely only on the shape of the word, and not on previous knowledge. I distinguished between several different kinds of words in order to focus specifically on two kinds: multimorphemic words that came from Latinate roots, and multimorphemic words from Old English roots. Because of the

nature of my study, the definition of *morpheme* that I used had to be somewhat different from the one used by Anglin (1993) in his study of morphological problem solving. Anglin only counted a word as being multimorphemic if the word's constituent elements were listed as words or as combining forms in the dictionary. Instead of looking for entries for the parts of each word, I used the etymological information provided by the dictionary (and supplemented, where needed, by information from the *OED*). Any word whose etymology showed that it was made at some point in time from two or more morphemes, whether Latinate or Old English, was counted as multimorphemic. This allows the use of such words as, for example, *niveous* (roughly synonymous with *snowy* or *snow-white*), which is coined from the Latin root *niv-* 'snow' and the Latin-derived adjectival suffix *-(e)ous*; the word is allowed even though *niv-* is not listed as a form in *W3.* This method allows the possibility of investigating how well English speakers understand morphemes that are bound and no longer have (or never had) explicit English meanings independent of the words they are bound in.

Types of words excluded. To make the conclusions clearer and eliminate confounding variables, many kinds of words were eliminated from consideration for inclusion in the test. Perhaps most basically, simplex words (those with only one morpheme) were eliminated, since they are unanalyzable and thus Latinate and Anglo-Saxon simplex words should stand on equal ground in the need to be memorized by rote. Similarly, simplex words with only inflectional affixes were eliminated, since these affixes do not add any particular meaning to a word, and so a simplex word that has one is still in meaning a simplex word. That is to say, the difference between what is denoted by *stretch* and by *stretched* is only a difference of tense; the meaning of the underlying lexeme is the same (Matthews 1991:49–54). Previous research has been carried out on children's acquisition of inflectional morphemes (*e.g.* White, Power, & White 1989), and has found that these morphemes—parts of basic linguistic competence—are acquired quite early. For this reason I did not test for them. Following Aronoff & Fuhrhop (2002), I considered the adverbial *-ly* suffix to be inflectional rather than derivational. (The rarer adjectival *-ly*, however, as in *friendly*, remains derivational.)

Because this study focuses only on Latin and Old English, any words that were composed of morphemes that came from languages besides Latin and English were also eliminated. This included a sizeable number of words that came to English either straight from Greek roots or from Greek roots filtered through Latin, as well as some words from various other world languages, many of which were the names of plants or animals peculiar to certain language-areas. Previous studies have tended to combine (Corson 1995, 1997) or even conflate (Bar-Ilan & Berman 2007) Latinate and Greek words. Although, while it was alive, Latin borrowed many words from Greek, the two wordstocks are distinguishable and quite different. I felt it was important to focus on words that came specifically from the Latin wordstock, rather than simply from a *different* wordstock, in order to keep the words comparable to each other. Words with morphemes mixed from more than one language (such as *nationless* < Lte *nation* + AS *-less*) were also eliminated, on the grounds that it would be impossible to draw any conclusion about the nature of specifically Latinate or specifically Old English words from such hybrids.

Technical terms were eliminated because the words in the test were meant to reflect words that *any* reader might have come in contact with and, moreover, that any reader might reasonably be able to construe the meanings of given the meanings of their constituents. The typical reader cannot be expected to construe the meaning of a word such as *butaldehyde* that is part of, in this case, the special knowledge-area of chemistry, without a background in chemistry. Including specialized words such as these would likely lead to unequal performance among subjects based on the primary field of their educational background, rather than only on their knowledge of morphemes. Among the categories of words counted as specialized words were all common and taxonomical names for species of organisms, such as *notchwing* and *olacaceae*, because each of these supposes a familiarity with the species in question. That is to say, a *notchwing* is not simply any animal with notched wings, but specifically a European moth *Rhacodia caudana*. For the same reasons, foods and musical terms were removed.

Another type of words that was eliminated was inspired by Anglin's (1993) discussion of idiomatic words, that is, words whose meanings (though nontechnical) cannot be construed from their parts. An example would be *surefire*, which basically means 'foolproof, infallible,' but in its derivation would perhaps suggest 'something that is certain to ignite.' The other categories of words that were removed are words derived from proper names (as *Ockhamistic*), words mutated unrecognizably from their root morphemes (as *matax* < *mattock* + *ax* 'a combination mattock and ax'), and dialect words (as *gaedown* 'a drinking bout (Scots)').

This left 253 Lte and 164 AS words, totaling 417. These words were ranked according to rarity. Initial ranking was carried out using the most compendious applicable scientifically built corpus of language-in-use available, the Corpus of Contemporary American English (410 million words) (Davies, 2008–). Because many of these words were too uncommon to show up in COCA, a second method had to be devised for ranking the rarity of these rarest words (which were of greatest interest to the study). I used the Google Books search engine (http://books.google.com), which scans an enormous corpus (about 15 million books, or 12% of the books ever printed (Michel *et al.* 2002)), restricting the search to the period from 1900 to the present to avoid thoroughly obsolete words. All of the rare words appeared in this corpus, with the rarest having frequencies between 10 and 100. I recorded Google Books frequencies for all words with frequencies up to 100,000. With the words ranked by frequency, I made two parallel lists, one AS and one Lte, and chose random pairs of words matched by frequency (assessed using the logarithm of the frequency count). Nine pairs of words were chosen and used in the final survey.

These eighteen words were supplemented with six nonrandom pairs of synonyms that had parallel etymologies but came from either side of the AS–Lte split. Synonym pairs were only accepted if their logarithmic frequency indexes were of comparable magnitudes. The words in these pairs were much more common than the other 18, none of them unheard-of in typical speech. One synonym from each pair was used in either of the two forms of the survey: three Lte and three AS on either form. (Appendix A lists all the words chosen.)

The survey form contained these 24 words in a randomized order (though with the paired synonyms at the beginning). For each word, the participant was asked first to break

down the word into the smallest "meaningful parts" (morphemes) he or she could identify, then to define each of these parts, and finally to define the entire word. Definitions in the form of either dictionary-like descriptions or sentences showing knowledge of the word were accepted. Participants were also asked to rate the word on how hard they found this process of analysis, from 1 to 5. (Appendix B shows a sample question from the test, with a response filled in.) Participants were allowed 30 minutes to finish the test. The time limit was meant to encourage participants to give their first impressions of a word, rather than to take more time analyzing the word than they would be likely to take if they came across it while reading. Most participants had no trouble finishing the entire test in 30 minutes. Some finished in as little as 20 minutes; however, some did not reach every word by 30 minutes, leaving the rest blank.

The participants were college students from a small Iowa college and residents of the town in which the college is located, as well as a few working- and middle-class participants from Ohio. They were not given information about the two different kinds of words until after the study.

Responses were judged by the author; hard-to-judge cases were resolved by consultation with another native English speaker to whom the meanings of the words were explained.

Analysis

The first and simplest finding from the data was that the subjects scored significantly higher with the AS words than with the Lte words. On average, across all education levels and considering only the unpaired words—that is, ignoring the synonym pairs, which were everyday words with well-known definitions—participants got 18.7% more AS words right than they did Lte words: 51.2% Lte right, 32.5% AS right (Student's paired t: $n = 28$, $t = 6.05$, $p < .0005$).

Among the (everyday) paired synonyms, as expected, there was no significant difference between AS and Lte in the subjects' ability to define them: 91% of the AS words were defined right, and 93% of the Lte words (Student's paired t: $n = 6$, $t = -1.02$, $p = .355$). However, there was a significant difference in their ability to define the *morphemes* of these words. On average, subjects correctly defined 79% of the morphemes in any given AS member of a synonym pair, but only 53% of the morphemes in a Lte member (Student's paired t: $n = 6$, $t = 4.18$, $p = .009$). (Morphemes were scored as correct if the meaning that the subject provided matched the meaning of a morpheme, even if the subject had not correctly isolated the form of the morpheme. For example, a subject defining the *frat* in *fraternity* as 'brother' received credit for the meaning even though the Latin root for 'brother,' *fratern-us* or *frater*, includes the *er(n)* as well.) Among the unpaired words, the same finding held, though it was not quite as robust: 69.7% AS morphemes right, 51.6% Lte morphemes right (Student's paired t: $n = 9$, $t = 2.47$, $p = .026$).

Subjects' own ratings of the difficulty of each word also showed Lte words as the harder ones. On a scale from 1 to 5 (with 1 being the easiest), the average rating of a Lte word was

3.03, while the average AS word was rated 2.69 (Student's t: >300 d.f., $t = 2.78$, $p = .0029$), a difference of 0.34 in a sample with a pooled standard deviation of 1.30. Among only the unpaired words, the tendency held: Lte words averaged a 3.29 rating, and AS words averaged a 3.00 rating (Student's t: >200 d.f., $t = 2.44$, $p = .008$). Among only the synonym pairs, which the subjects overwhelmingly could define as whole words, the rated difficulty of analyzing Lte words over AS words was yet more pronounced: the AS words were rated 2.33, and the Lte words rated 3.01 (Student's t: >80 d.f., $t = 4.03$, $p < .0001$).

In this sample, a college education had no significant effect on subjects' ability to define words of either kind. Defining low college education as two semesters or fewer, I found that those with low college educations (for short, "LC" subjects) tended to correctly define unpaired words of either kind less often than those with 3 semesters or more of college ("HC" subjects), but not at a level that was statistically significant in this sample: For Lte words, 50.9% HC against 41.7% LC (difference 9.3%; Student's t: 12 d.f., $t = 1.17$, $p = .264$); for AS, 65.7% against 51.7% (difference 14.1%; Student's t: 12 d.f., $t = 1.90$, $p = .079$). (See Fig. 1.)

Subjects who had studied Romance languages ("R" subjects) did not, on average, do significantly better at defining Lte words than subjects who had not ("NR" subjects), defining 52.2% of unpaired Latinate words right on average, as against 42.3% for NR subjects, a difference of 9.9% (Student's t: 25 d.f., $t = 1.54$, $p = .137$). Unexpectedly, R subjects did score significantly better on the *Anglo-Saxon* words than the NR subjects did: 67.2% as against 53.2%, a difference of 14.0% (Student's t: 25 d.f., $t = 2.16$, $p = .041$). (See Fig. 2.) This seems to lead to the counterintuitive conclusion that studying a Romance language gives one an edge at defining AS words. However, the two percentage differences, 9.9% and 14.0%, are on the same order of magnitude, and I find it likely that they actually show either that studying a Romance language gives one an edge in analyzing words of *any* kind, or, perhaps more likely, that the subjects in my sample who had studied Romance languages happened to be better at analyzing words in general. My belief in this latter possibility is strengthened by the fact that 11 of the 15 R subjects noted on their survey form that they could not speak fluently in the Romance language they had studied (they wrote, for example, "studied only" or "not fluent"), suggesting that their studies might have had a relatively minimal impact on their skill with Lte word roots. It is possible that a larger study (with, perhaps, more thoroughly bilingual participants) would reveal a tendency for speakers of Romance languages to succeed more often at analyzing Lte words.

Lastly, I analyzed the data for completeness of breakdown: that is, of the possible number of morphemes that could have been isolated from each word, how many did subjects actually isolate as potential meaningful units—regardless of whether they actually knew the meanings of these units? To analyze this, I found, for each response, the number of meaningful units that the subject had isolated in the first step of the breakdown (where I asked what were the "parts of the word"), and divided it by the number of possible morphemes that could have been isolated. "Meaningful units" here are of course not the same as the basic morphemes that make up a word; I also accepted combinations of two or more basic morphemes, so that, for example, someone analyzing *accumulation* as *accumulate* + *-ion* would be credited with

isolating two meaningful units. However, not all combinations were accepted: only those that had meanings of their own or had been meaningful at some point in the word's history (for example in Latin or French). Because compound words are generally built by adding new morphemes onto existing (and thus meaningful) words, the order of wordbuilding typically showed which units could be counted as meaningful. However, a few exceptions were made, notably the *-ation* suffix made of *-ate* + *-ion*, which, as discussed earlier, has become a single morpheme on its own and can be added to words independently of the *-ate* suffix. To clarify with the same example, the analysis *accumul* + *-ation* would only be credited with one meaningful unit (*-ation*), since *accumulation* formed[6] as [[ac-[[cumul]-at]]-ion]; the form *accumul* never existed in this wordbuilding, and never functioned in any form as a meaningful morpheme or word. The form *-ation* would not be counted either were it not for the exception noted above. *Accumulate* + *-ion* would count as a 2/4 breakdown and *accumul* + *-ation* would count as a 1/4 breakdown.

The subjects showed a strong tendency to break down AS words more thoroughly than Lte words. Among the paired synonyms, where one would expect the subjects' breakdowns to be most equivalent if the words were equally analyzable in the subjects' minds, the average breakdown of an AS word was 98% complete, but for a Lte word it was only 68% complete. This difference was statistically significant (Mann–Whitney's U: $z = 5.93$, $p < .0001$) (Conover 1998). The same tendency held among the unpaired words, where an AS word's breakdown was on average 91% complete and a Lte word's was 79% complete (Mann–Whitney: $z = 3.75$, $p = .0001$).

The data showed several instances where subjects were trying to use morphological problem solving but running up against limits to their knowledge of word roots, leading to mistaken solutions to these words' morphological problems. One fairly common manifestation of this tendency was when subjects tried to use the definition of a Latinate morpheme that is used independently in English, even if that morpheme had an independent definition different from the one used in the word. For example, several subjects guessed that the *extra-* in *extravisible* ('outside the range of vision') meant 'more than ordinary', which is the definition of *extra* as a standalone English word. However, in *extravisible*, it actually means *outside of*, as it did in Latin and as it still does in such words as *extraordinary* and *extraterrestrial*. This also happened with the *salut-* in *insalutary* (L 'health,' English 'greet') and the *pro* in *probability* (English 'for, in favor of' but in this word in Latin it was part of *prob-* 'to test').

This last example-word given illustrates subjects' tendency to look for meaning wherever they could find it. In some cases the search for sense overrode the correct division of the words into morphemes. This was particularly visible in the word *sinistrocular* ('relating to or dominant in the left eye,' < *sinister* 'left' + *ocul-us* 'eye' + *-ar* (adjectival suffix)), from which some people isolated the form *sin*, meaning variously 'to do bad,' 'together' (presumably influenced by the Greek prefix *syn-* with that meaning), or 'without' (this response from someone who had studied Spanish, where *sin* indeed means 'without'). However, only two of the subjects who isolated *sin* were able to use it to forge tentative definitions ("Seeing

[6] As we can tell from the morphological rules governing each morpheme: *cumul-us* 'pile' is a noun; *-ate* makes verbs out of nouns; *ad-* (*ac-*) 'to' specifies a direction for verbs; *-ion* makes nouns from verbs.

together? Eyes working together?" and "Without the aid of vision?"), and both these definitions were accompanied by question marks. The tendency to look for meaning in mistaken places was much more pronounced among Lte words. In AS words it only showed up in a few responses that defined, for example, the *-hood* suffix as the noun *hood* 'head covering attached to a shirt.'

Discussion and Conclusions

The results of this study clearly suggest that the Latinate words of the English wordstock are harder for people to understand than the native, Anglo-Saxon words. The implications of this conclusion for the practice of education and academics are significant.

The difficulty of Lte words

Remindingness. This study found that English speakers find it harder to do morphological problem solving on Lte words, implying that Lte words are thus more difficult for people to understand. It may fairly be objected that, once they know the definition of a word, people will not need to perform morphological problem solving on it to determine its meaning every time they encounter it, which would suggest this study's results only bear on readers' first encounters with words—an area of fairly small interest, especially since rare words may be accompanied by an explicit definition on their first mention anyhow. However, although this study was run with words explicitly intended to be novel for the subjects (except for the paired synonyms), its results can be extended far beyond first encounters with words.

It can be supposed that a word that shows its meaning to the speaker in its form—a word that is reminding[7]—will be easier remembered than a word that is opaque to the speaker. In some cases this should be expected to come about, straightforwardly, as a result of the greater frequency in the language (and thus greater reinforcement in the mental lexicon) of the component morphemes; thus the definition of a word like *pollbook* 'a book containing the results of a poll' could be easily recalled from its parts even if the speaker had forgotten the word itself, since *poll* and *book* are fairly common words. However, this same mechanism does not serve in the case of many Lte words, since some of their crucial component morphemes do not exist independently in English, as discussed earlier. In this case, the remindingness of a word is expected to come about through the speaker's knowledge of the root in another language, or through explicit study of Lte word roots. For the reminding effect to happen, however, the word roots must still be more reinforced in the speaker's mental lexicon than

[7] In light of the earlier discussion of academic wordbuilding, it is perhaps worth mentioning that I have chosen this word rather than *mnemonic,* a Greek-derived word that at first seemed more appropriate; I eventually decided that *reminding* conveys the same meaning with greater analyzability. It includes a Lte morpheme, *re-*, but this morpheme is very productive and there is no living AS equivalent, so the word serves to illustrate the kind of situation where a Lte morpheme is clearly more appropriate than some contrived AS replacement.

the word they compose; otherwise the speaker will simply retrieve the definition of the word in question first.

But the results of this study suggest that speakers are not, from any source, getting the knowledge they need to use Lte word roots in such a way ideally, since participants overall correctly defined about 18% fewer Lte word roots than AS roots. So we can conclude that Lte roots are about 18% less useful for jogging one's memory than AS word roots are. This is when the reader can even isolate the Lte root; the analysis of completeness of breakdown shows that this is often not the case. Lte words, it seems, are often entered into the mental lexicon as unanalyzed wholes. The finding from the paired synonyms illustrates this well: the participants quite often found it possible to define each component of the AS member of a pair (79% of morphemes right), but they were rather less often able to define the Lte components (53% right), even when they were able to provide a reasonable definition for the whole word (which they did 93% of the time).

When subjects did know Lte morphemes, their source of knowledge was not obvious. Though in such cases as that of *sinistrocular*—which was defined right by four speakers, two of whom had studied Latin—it was possible to guess that some participants who defined it right did so because of their knowledge of Latin, in other cases it was not at all clear exactly where they got their knowledge, since those who spoke Romance languages did not fare better at defining Lte morphemes than those who did not. It is possible that some speakers have, intentionally or not, memorized Latin roots from their use in species names, etymologies that are sometimes provided with first-use definitions of obscure Lte words, or other sources used in English-language contexts.

Considering remindingness, the difference between speakers' understanding of Lte and AS words becomes an issue of ease of processing. The meaning of an uncommon but reminding word can be remembered through any of three strategies—from context, or by remembering the whole word, or by remembering its parts—or, crucially, through a combination of these, where context or the meaning of one or more of the parts reminds the speaker of the definition of the whole (even if they do not exactly sum up to this whole). The definition of a word that is not reminding, meanwhile, can only be remembered through two of these strategies.

We should then expect that the processing time for a word that a speaker knows will be longer when the word is not reminding. High processing time may be the most meaningful conception, at least in the context of this study, of what makes a word "difficult" (a concept I have been using intuitively thus far), since an unusually long processing time for a word can interfere with the normally smooth processing of words during reading, listening, and speaking, and thus interrupt or even halt the understanding or production of a sentence. Here Bergman, Hudson, & Eling's (1988) study is relevant in its finding that Lte words take longer to process for Dutch speakers; however, their study does not focus on the time required for speakers to *understand* these words, which may have a yet more pronounced effect, though one harder to measure. In this way we can see that the non-reminding nature of Lte words can indeed be expected to make them more difficult. The data from this study, in showing that readers can understand some reminding words even if they have never before encountered their definition, bear this conclusion out. Indeed, the instances of

mistaken analysis found in the data suggest that the shapes of non-reminding words may go beyond being unusable into actually being a hindrance for speakers, as they may attempt to connect the word to some sort of definition but be misled, by the false morphemes superficially present in the word, to assign it an incorrect definition.

In academic writing. We can fairly guess that the great prevalence of Latinate words in academic discourse is one factor that makes such discourse difficult and inaccessible to many readers, as claimed by Corson (1993). It is interesting to note how at least one other factor in the difficulty of academic writing that has been identified may actually lead vocabulary to act as an even greater stumbling block. Biber & Gray (2010) detailed several characteristics of the register of academic English, and a prominent one was a tendency for authors to condense meanings that might be expressed with an entire clause or sentence into noun phrases: to use one of their examples, *computation time* instead of a clause such as *the time required to compute something.* Since specialized concepts are what are conveyed in these noun phrases, specialized words will be needed, and as discussed before, the most specialized words in English are Latinate. Expansions of these compressed phrases may not have fewer Lte words—indeed, as the example illustrates, they may have more (*required* has been added here)—but the compressed phrase will typically have very few function words to orient the reader, meaning that the meaning of the noun phrase must be conveyed wholly by the meanings of the words that make it up. The reader must fill in the gaps with previous knowledge of how the words could be related, to arrive at the correct meaning rather than, in this example, an interpretation like *the time of day at which something is computed*. Though context will quite often clarify such compounds, in phrases whose meanings hinge on implied connections between nouns, it will be all the more important to know the meanings of these nouns (especially when they are each made of several morphemes, unlike *time* here), and all the harder to make sense of a passage if a crucial word is not in the reader's vocabulary.

Implications of the difficulty. Though some English speakers may only be inconvenienced by the difficulty of Lte words, needing perhaps to look them up, it seems altogether likely that the prevalence of Lte words in academic registers leads some English speakers to have an unfavorable opinion of academia in general. Poking fun at academic and Latinate language has a long history in English, represented in a host of characters whose hifalutin talk serves as the butt of jokes: Jespersen mentions "Shakespeare's Dogberry and Mrs. Quickly [to which I would add Holofernes], Fielding's Mrs. Slipslop, Smollett's Winifred Jenkins, Sheridan's Mrs. Malaprop, Dickens's Weller senior, Shillaber's Mrs. Partington, and footmen and labourers innumerable" (Jespersen 1923:§143). With such portrayals of academic language representing the popular opinion of Lte words—and perhaps, to some extent, influencing it (see Johnstone 2008:10)—it must be concluded that there are many English speakers who feel that a Lte vocabulary shows pretension or obscurantism. Such popular feelings are echoed by the authors mentioned earlier (Bryson, Jespersen, Orwell, Strunk & White) who identified Lte words as being longer, harder to understand, and less "homely-sounding."

Corson (1995) concludes that the widespread inability to fully engage with Lte words creates a "lexical bar" that keeps less socioeconomically privileged people from fully engaging with academia. Though this study did not find significant differences between those with and

without college educations in the ability to understand Lte words, neither can such a possibility be discounted on the basis of the data. In any case, the effect of the difficulty of Lte words can be expected to apply not only to those who are deciding whether or not to embark into higher education, but also to those presently immersed in it. Those who have most cause to read academic prose are college students and professors, so they stand to be most affected by the difficulty of the Lte words in the texts they read. The effect of the difficulty would be a continual running up against unfamiliar words or words that require significant effort (and a break in concentration) to understand—as well as a struggle to properly employ the vocabulary of one's chosen discipline. These are familiar phenomena to those in academia, though it is beyond the scope of this project to address just how common and how pronounced they are. We can say from the data here that people subjectively judge the breakdown of Lte words harder than for AS words. Although the difference found, 0.34 on a scale from 1 to 5 (standard deviation 1.30), is not particularly striking, it nonetheless shows that readers can be aware on a conscious level of the difficulty of the Lte subset of words, even if they may not recognize the subset they find hard as anything but "big" or "hard" words. Research into the extent of readers' difficulty with Lte words should prove fruitful, perhaps focusing on how much a Lte wordstock can slow a reader down in reading a text, or otherwise dealing with these words in context.

Affected less severely but still affected would be those who, outside of a collegiate context, try to keep abreast of developments in science, or in other academic fields, represented in this study by those with little or no college education. Though science magazines geared toward the general public, such as *Popular Science*, write in a journalistic style aimed at broad understanding, rather than an academic style aimed at deep and specialized understanding, terminology from the academic community bleeds through into these areas of popular discourse. Research from an earlier stage of this project on a recent corpus of *Popular Science* articles shows that words not often found in common speech have made their way into this magazine: *deorbit*, *inerting*, *counterrotating* (and from the AS side, *overboosting*, *unbrand*, *treadless*)—often not explicitly defined (as none of these were, even though they were among the rarest words I found). A difficult, Lte vocabulary in these areas of knowledge means more difficulty in engaging with this knowledge—a handicap that should be borne in mind in thinking, for example, of the controversial state of education in the United States.

Suggestions

Though the results of this study suggest that English would be easier for its speakers if its vocabulary were more Anglo-Saxon in nature, the history of the Saxonists suggests that any sort of effort to rid English of its foreign words is probably rather unlikely to succeed. Even were it possible to institute a total replacement of the Lte words, it is hardly clear that such a thing would be desirable, since all speakers would have to learn a new set of words for concepts that are, presumably, already marginal enough in their mental lexicons.

Instead, given the finding that Lte words are harder for readers, what we may focus on is the use of the different kinds of English words in composition. Writers on style commonly advise against using needlessly long or obscure words, an implicit injunction against Lte words (considering the popular identification of Lte words as long or hard). Orwell

(1999[1950]) explicitly makes the connection to etymology: "Bad writers . . . are nearly always haunted by the notion that Latin or Greek words are grander than Saxon ones." This advice has added force considering the results of this study. It seems fairly clear that, for more understandable writing, AS words should in general be preferred over Lte equivalents. It must be pointed out that to many or most writers the etymology of words will not be clear at a glance. And some Lte words have no AS synonyms or only obscure or obsolete ones, in which case they should not be replaced with an awkward substitution or circumlocution. So perhaps the more broadly useful advice is that it is better to use a plain, reminding word (or phrasing) rather than its more obscure synonym, provided the sense is right. This will often lead to an AS phrasing, and when it does not, the AS phrasing may well be the wrong one for the occasion. Strunk & White (2000) elaborate on this thought:

> Anglo-Saxon is a livelier tongue than Latin, so use Anglo-Saxon words. In this, as in so many matters pertaining to style, one's ear must be one's guide: *gut* is a lustier noun than *intestine*, but the two words are not interchangeable, because *gut* is often inappropriate, being too coarse for the context. Never call a stomach a tummy without good reason.

They go on to stress that the choice is "a question of ear," which is perhaps as specifically as this can be phrased without an in-depth study that would not fit in the confines of this paper.

Another area of interest that this study's results bear on, and one not nearly so widely discussed as writing style, is the coining of words. I have encountered many newly coined words or phrases in academic writing that are Latinate or Greek-derived but could easily have been otherwise. As mentioned earlier, recent word coinages in English, and especially academic ones, distinctly tend to come from Latin or Greek roots, but this need not be so. The idea of *complementary schismogenesis*[8] could have been coined in English with little or no loss of meaning as something like *two-sided splitmaking*. A great many new phrases could likewise be coined from English roots, resulting in words whose meanings could be worked out through morphological problem solving, and which would more easily reside in English speakers' memories than non-reminding agglomerations of Lte roots. If a trend of coining new words from AS roots were to catch on, there is even the possibility that new AS versions of Lte words could naturally supplant the Lte words, perhaps a more likely way for English to get a more AS vocabulary than through the efforts of modern-day Saxonists.

This issue merits further study, especially into the effect of college education on the ability of adults to engage with Lte vocabulary, to find more conclusive answers to the matter than were revealed in this study. The difficulty of Lte words in context remains scarcely studied, and the concept of remindingness might profitably be investigated from a neurological or other approach that allows the presence of context or quantifies the effect in different ways. As discussed above, academic discourse in English is inaccessible often to the point that it is parodied, and an understanding of how to broaden its accessibility could result in greater equality in education of the type that Corson (1995, 1997) advocates, not to mention the possibility of easing some of the difficulty of engaging in academic discourse for those

[8] This describes the divisive situation that arises when two people in a conversation each think the other is being rude because they have different ideas of what is polite—see Johnstone 2008:141.

already involved in it. Though such possibilities may be distant, studying the issue must be the beginning of any progress.

Acknowledgements
Thanks to Tim Arner for his invaluable comments on this paper during its preparation; to Brigittine French for helping out with the formation of the project and for help as a second reader; to Janet Gibson for help on experimental design; to Chris Olsen for helping ensure that the statistics make sense; to Ragnar Þórrison for putting up with all my questions about Icelandic, and not just the ones for this project; and to Adam Lauretig for his "cruel and arbitrary" second opinions on the test responses.

References

Algeo, John, 2010. *The Origins and Development of the English Language,* 6th edn. Boston: Wadsworth.

Anglin, Jeremy M., 1993. *Vocabulary Development: A Morphological Analysis.* Monographs of the Society for Research in Child Development, 58:10.

Aronoff, Mark, and Nanna Fuhrhop, 2002. "Restricting Suffix Combinations in German and English: Closing Suffixes and the Monosuffix Constraint." *Natural Language and Linguistic Theory* 20:3, pp. 451–490.

Baron, Dennis E, 1982. *Going Native: The Regeneration of Saxon English.* Publication of the American Dialect Society No. 69. Tuscaloosa, Ala.: University of Alabama Press.

Bergman, M. W., P. T. W. Hudson, & P. A. T. M. Eling, 1988. "How Simple Complex Words Can Be: Morphological Processing and Word Representations." *Quarterly Journal of Experimental Psychology* 40(A), pp. 41-72.

Biber, Douglas, 1994. "An Analytical Framework for Register Studies." *In* Biber, Douglas, and Edward Finegan, eds., *Sociolinguistic Perspectives on Register*. New York, NY: Oxford University Press.

——, and Bethany Gray, 2010. "Challenging Stereotypes about Academic Writing: Complexity, Elaboration, Explicitness." *Journal of English for Academic Purposes* 9:1, pp. 2–20.

Boas, Franz, 1911. *Handbook of American Indian Languages.* Bureau of American Ethnology, Bulletin 40. Washington, DC: Government Printing Office.

Bryson, Bill, 1990. *The Mother Tongue: English and How It Got That Way.* New York: W. Morrow.

Corson, David, 1995. *Using English Words.* Boston: Kluwer Academic Publishers.

——, 1997. "The Learning and Use of Academic English Words." *Language Learning* 47, 671–718.

Conover, W.J., 1998. *Practical Nonparametric Statistics.* New York, NY: Wiley.

Cutler, Ann., 1981. "Degrees of Transparency in Word Formation." Canadian Journal of Linguistics 26 (1): 73–77.

Davies, Mark, 2008–. *The Corpus of Contemporary American English: 425 Million Words, 1990–present.* Available online at <http://www.americancorpus.org>.

Dupuy, H., 1974. *The Rationale, Development, and Standardization of a Basic Word Vocabulary Test* (DHEW Publication No. HRA74-1334). Washington, DC: U.S. Government Printing Office.

Ferguson, Charles A., 1994. "Dialect, Register, and Genre: Working Assumptions about Conventionalization." *In* Biber, Douglas, and Edward Finegan, eds. *Sociolinguistic Perspectives on Register*. New York, NY: Oxford University Press.

Finkenstaedt, Thomas, & Dieter Wolff, 1973. *Ordered Profusion: Studies in Dictionaries and the English Lexicon.* Heidelberg, Germany: C. Winter.

Frisson, Steven, Elizabeth Niswander-Klement, and Alexander Pollatsek, 2008. "The Role of Semantic Transparency in the Processing of English Compound Words." British Journal of Psychology 99:1, 87–107.

Goulden, R., P. Nation, & J. Read. 1990. "How Large Can a Receptive Vocabulary Be?" *Applied Linguistics* 11, 341–363.

Hayes, C. P., & M. Ahrens, 1988. "Vocabulary Simplification for Children: A Special Case of 'Motherese'?" *Journal of Child Language* 15, pp. 395–410.

Johnstone, Barbara, 2008. *Discourse Analysis.* Malden, Mass.: Blackwell.

Matthews, P. H. *Morphology* (2ed.), 1991. Cambridge, England: Cambridge University Press.

Maylath, Bruce, 1996. "Words Make a Difference: The Effects of Greco-Latinate and Anglo-Saxon Lexical Variation on College Writing Instructors." *Research in the Teaching of English* 30:2, pp. 220–247.

Michel, Jean-Baptiste, Yuan Kui Shen, Aviva Presser Aiden, *et al.*, 2002. "Quantitative Analysis of Culture Using Millions of Digitized Books." *Science* 331, pp. 176–182.

Orwell, George, 1999 (1950). "Politics and the English Language." *In* Hirschberg, Stuart, & Terry Hirschberg, eds., *Reflections on Language*. New York: Oxford University Press.

Partridge, Eric, 1978. *A Dictionary of Clichés.* Boston: Routledge & Kegan Paul.

Shakespeare, William. *Love's Labour's Lost.* Cambridge: Cambridge University Press, 1923.

Strunk, William, and E.B. White. *The Elements of Style.* New York: A.B. Longman, 2000.

Quirk, Randolph, 1974. *The Linguist and the English Language.* London: Edward Arnold.

"What Is Anglish?" n.d. In *The Anglish Moot*. Retrieved April 1, 2011, from <http://anglish.wikia.com/wiki/What_is_Anglish%3F>

White, T. G., M. A. Power, & S. White, 1989. "Morphological Analysis: Implications for Teaching and Understanding Vocabulary Growth." *Reading Research Quarterly* 24, 283-304.

Appendices

A: List of words used in the survey.

Words varied between the two test forms	
Form 1	*Form 2*
brotherhood (AS)	fraternity (Lte)
probability (Lte)	likelihood (AS)
buildup (AS)	accumulation (Lte)
insignificant (Lte)	meaningless (AS)
hindsight (AS)	retrospect (Lte)
discharge (Lte)	unload (AS)
Words common to both test forms	
Lte	*AS*
extravisible	stringsman
sinistrocular	wordster
insalutary	grabhook
disherit	worldful
niveous	breadthways
chainlet	pollbook
subcoastal	tinwork
institutor	spottiness
exhalant	nethermost

B: Sample survey question and response.

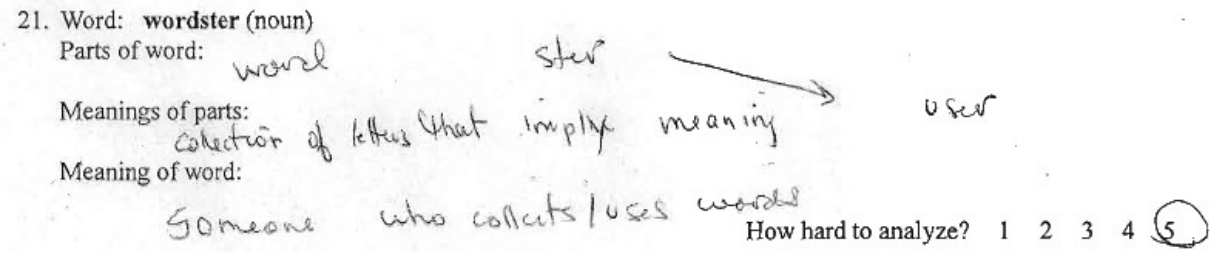

Figure 1: Percentage of unpaired words defined right by subjects with different levels of college education.

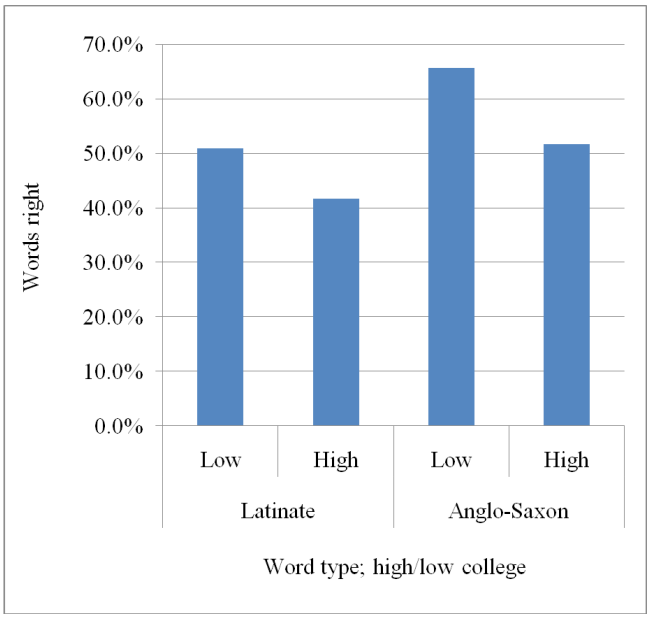

Figure 2. Percentage of unpaired words defined right by subjects who had and had not studied a Romance language.