Measuring Lexical Style and Competence:  
The Type-Token Vocabulary Curve

Abstract
A personal computer is used to analyze samples from literary texts by thirteen different authors, including passages from Genesis, Hemingway, Longfellow, Jane Austen, Henry James, George Eliot, James Joyce, and Basic English (created by C. K. Ogden). The total number of words (tokens) and the number of distinct vocabulary words (types) are computed for each sample. The number of types are then plotted against the number of tokens for eight of the texts. From these type-token curves, inferences are drawn about both lexical style (vocabulary use) and lexical competence (vocabulary size). For example, the curves for "Big Two-Hearted River" and for a summary of Macbeth in Basic English nearly coincide for their first 1100 tokens, after which they gradually diverge. This graphical pattern corresponds with intuitions that Hemingway's prose reads much like Basic English but that it draws upon a larger total vocabulary. The curve for Joyce's Ulysses, by contrast, rises much more rapidly than that for a late passage from A Portrait of the Artist as a Young Man; however, after 800 tokens, the two curves begin to converge. This suggests that the difference between Ulysses and Portrait is largely one of lexical style rather than competence. The highest type-token curve for the samples tested was that for Finnegans Wake; the lowest curve was for Genesis. Comparison with type-token statistics gathered by Kucera and Francis suggests that the curves for the Wake and Genesis are near the maxima and minima for English literature.

One of the earliest claims made by generative grammarians, now widely accepted, is that we acquire a nearly complete grammar of our native language by about the age of six or eight. But insofar as a generative grammar includes a lexicon, acquisition is never complete: throughout our lives, we continue to add new words to our vocabularies and to alter our use of old ones. Our lexicon--our vocabulary--is one facet of linguistic knowledge (competence), just as our choice of words in specific cases is one facet of linguistic style (performance).

Most generative grammarians have assumed that statistics are useful in describing style but not competence. As a rule of thumb, if a linguistics article includes diagrams (flow charts and tree diagrams) it is usually written by a generative grammarian, whereas if it includes graphs, it is usually written by a sociolinguist. But computational linguists such as Herdan (1960) and Carroll (1968, 1971) have argued persuasively that statistics can be used to estimate not only the frequency of use of specific words, but also to estimate the size of the vocabulary from which they are drawn; hence, statistical evidence is relevant to both style and competence. In this article, I attempt to demonstrate the usefulness (and the limitations) of one statistical measure--a type-token vocabulary curve--in describing both the style and the competence of a variety of English and American authors.

For this study, I wrote a TurboPascal program which reads text files from a Macintosh SE computer, counts each word (token) in a text, records each new vocabulary word (type) as it is encountered and computes the total number of tokens and the total types accumulated at that point in the text. I then applied the program to twenty passages written by thirteen different authors and compared their type-token statistics. The results are summarized in the table in (10) below.
Certain decisions--and compromises--must be made in any computerized study of vocabulary: for example, does time-worn count as one word (token) or two? are cat and cats different words or different forms of the same word (type)? For the purposes of computerized analysis of texts, the simplest (though not the most accurate) definition for printed word tokens is the one used in Kucera and Francis 1967 and in Francis and Kucera 1982 (p. 3): "Graphic word: a string of contiguous alphanumeric characters with space on either side; may include hyphens and apostrophes but no other punctuation marks." By this definition, can't, cannot, and babe-in-arms count as single words, whereas can not and can opener are two-word phrases. Such a definition can introduce statistical errors because the choice between, say, tear jerker, tear- jerker, and tearjerker reflects editorial conventions rather than genuine differences in vocabulary. Fortunately, the passages examined for this article contain few variants like these--with the notable exception of Finnegans Wake: from a statistical point of view, the size of the vocabulary in the Wake is inflated by Joyce's playful invention of compounds such as upturmpikepointandplace, which would have been written as phrases by nearly any other English author.

Again in order to simplify computer analysis, "A 'distinct word' (type) can also be simply defined as a set of identical individual words" (Kucera and Francis, p. xxi). That is, Kucera and Francis count all and only identical alphanumeric strings as being the same graphic word (type). Differences between upper and lower case are ignored, resulting in occasional errors; for example, Brown (proper noun) and brown (adjective) count as the same word, as do Polish and polish. Similarly, bear (noun, 'mammal') and bear (verb, 'carry') count as the same word, while can't, cannot, and can not count as different words, as do inflectional forms such as bear, bears, and bear's. The authors admit that this sort of computerized count gives somewhat inaccurate results, since "distinct graphic word" is not identical with what most people mean by "distinct word." In practice, however, the two concepts are closely correlated, and in comparative studies such as this one, the precise definition used for "distinct word" is less important than ensuring that the definition is explicit and that it is applied consistently.

Francis and Kucera 1982 takes their analysis a step farther. Graphic words are grouped into lemmas, defined as follows: "Lemma: a set of grammatical words having the same stem and/or meaning and belonging to the same major word class, differing only in inflection and/or spelling" (p. 3). Hence, the lemma be subsumes the inflectional forms been and being, the suppletive forms am, is, was, were, and even spelling and dialect variants such as are/ah, were/wuh. Most people, when speaking of an author's "total vocabulary," probably mean something like 'total number of lemmas' rather than 'total number of graphic/phonetic words'. For example, cat, cats, and cat's are not usually thought of as "different words," but as singular, plural, and possessive forms of the "same word." Thus, Thorndike and Lorge (1944, p. ix) count inflected forms of nouns, verbs, and adjectives "under the main word," and their list of "30,000 words" actually subsumes thousands of additional graphic words.

Thorndike and Lorge do list suppletive forms separately, with separate entries for words such as am/be/is/was/were. In this they differ from Ogden (1934, p. 3), who describes Basic English as "a careful and systematic selection of 850 English words which will cover those needs of everyday life for which a vocabulary of 20,000 words is frequently employed." Ogden's claim borders on false advertisement, since his copyrighted list of 850 words is not only intended to subsume hundreds of additional inflected and derived words but also suppletive forms--the most spectacular example being he, which is listed as the lemma for twelve distinct graphic words: he, him, his, she, her, hers, it, its, they, them, their, and theirs. (This is the generic he run rampant!) Consequently, the 850-word vocabulary of Basic English actually expands to thousands of distinct graphic words, although these remain a small subset of the total vocabulary.
of a typical native speaker of English.

These examples from Ogden, Thorndike and Lorge, and Kucera and Francis illustrate that there is considerable variation in what counts as the "same word" (type) in statistical studies of vocabulary; hence, it is difficult to compare the results from different studies. For this article, I have used the definition of distinct word (type) which simplifies computer analysis—the one used in Kucera and Francis 1967: a word (type) is any distinctive, continuous string of alphanumeric characters (including hyphens and apostrophes but excluding other punctuation) that is preceded and followed by a space. Where possible, I have also selected samples of 2000 words (tokens) or more—the size of the samples in Kucera and Francis—so that my results for individual authors can be compared directly with theirs.

Statistical studies of vocabulary almost ritualistically report the ratio between types and tokens for a given sample of text. However, this statistic turns out to be nearly useless, as an analysis of one set of type-token ratios will illustrate. In (1), using the MYSTAT software package, I have plotted the type-token ratios against the number of tokens for "Macbeth," which is a translation into Basic English by T. Takata of a passage from Charles Lamb’s Stories from Shakespeare (Ogden, pp. 286-298).

(1) Type-Token Ratios for "Macbeth" in Basic English

The first sentence of the passage in Basic English can be used to illustrate how this curve is derived:

(2) At the time when Duncan the Kind was King of Scotland, there was a great lord, named Macbeth.

The first type-token ratio (for the word at) is 1/1 = 1.0; the second type-token ratio (for the first two words at the) is 2/2 = 1.0, and so on until the first repetition—the sixth word the, where the type-token ratio falls below 1.0 (5 types / 6 tokens = 0.833) and it
remains below 1.0 thereafter. The next repetition is the thirteenth word was, where the type-token ratio is 11/13 = 0.846. As the text unfolds, more repetitions occur, and the type-token ratio continues to fall—rapidly at first, and then more slowly. Eventually, if the passage were long enough, the entire vocabulary of Basic English would be exhausted, and no new words could be added. (This picture is complicated by the fact that proper nouns such as Macbeth and Duncan can be introduced freely into a Basic English text, but presumably our stock of proper nouns is finite, too, and therefore subject to exhaustion.) Because the number of types remains finite, the ratio between types and tokens necessarily approaches zero as the number of tokens approaches infinity (n/8 = 0). Consequently, the curve in (1) begins at 1.0 and declines toward zero. Indeed, for any natural language text, the curve for type-token ratios has this same general shape—apart from special cases such as concordances and other lists, in which the type-token ratio remains fixed at 1.0.

Type-token ratios are reported in studies of everything from freshman compositions to childhood acquisition of language, as well as in sophisticated statistical analyses such as those by Kucera and Francis. Nevertheless, a glance at the curve in (1) makes it clear that type-token ratios are insignificant statistics by themselves, since the ratio for any text (provided that it is sufficiently long) varies from a maximum of 1.0 to a theoretical minimum of zero. Hence, this ratio cannot distinguish any text (or any author) from any other. It is not type-token ratios that are significant, but only the rate at which they decline. In (1), for example, the type-token ratio after 2000 words (tokens) is just 0.23, whereas after 2000 words of Shakespeare's Sonnets, the ratio is significantly higher: 0.39. Indeed, the passage plotted in (1) is nearly 4800 words long, but it draws upon a vocabulary of only 697 words, whereas Shakespeare's vocabulary reaches 780 after just 2000 words of the Sonnets. (The obvious—and correct—conclusion is that Shakespeare's total vocabulary is larger than that of Basic English.)

Type-token ratios are meaningless, then, unless we also specify the number of tokens used in computing them: 0.39 at 2000 words (Shakespeare), 0.23 at 2000 words (Basic English), and so on. But this makes it pointless to compute a ratio at all, since this ratio provides no more information than the raw data do: 780 types at 2000 tokens (Shakespeare), 460 types at 2000 tokens (Basic English), and so on. That is, we can compare the number of types directly (780 versus 460) rather than the type-token ratios (0.39 versus 0.23) and the ratio between these two pairs of statistics is necessarily the same (780/460 = 1.7; 0.39/0.23 = 1.7). Hence, despite the frequency with which they occur in studies of natural language texts, type-token ratios are both meaningless and pointless statistics, and it is preferable to plot the number of types in a passage directly against the number of tokens, as in (3), rather than type-token ratios as in (1):
Hereafter, I will refer to any graph such as (3) as a type-token vocabulary curve, or more briefly as a type-token curve. The curve for a concordance or any other vocabulary list is a straight line with types = tokens at every point. Type-token curves for other texts, however, resemble the one in (3). They begin as a straight line, with types = tokens, until the first repeated word. Thereafter, the number of tokens exceeds the number of types, and this margin grows larger with every additional repetition. Consequently, type-token curves rise rapidly at first, then begin to lose momentum as repetitions become more frequent and the author's vocabulary is used up. The number of types reaches its maximum when the author's vocabulary is completely exhausted. Hence, as the number of tokens approaches infinity, the number of types approaches the total vocabulary of the author. Because of this characteristic, the type-token curve can be used to estimate the size of the vocabulary from which a text is drawn, although rather complicated statistical calculations are required to do so (see Carroll 1968 and 1971).

The curve in (3) clearly shows the early effects of exhaustion of the artificially constrained vocabulary of Basic English, although its total vocabulary of graphic words is several times as large as the 697 used in this passage. This example illustrates that very long passages--impossibly long ones--would be needed to exhaust the vocabulary of an adult native speaker of English, partly because we use many words very rarely (if at all); for example, a very long a sample of our writing and/or speech would be needed to turn up a word such as prestidigate, much less all of its inflections and derivatives (prestidigitates, prestidigitated, prestidigation, and so on). None of these words appears in the million-word sample recorded in Kucera and Francis 1967 or in the five-million-word sample of Carroll, Davies, and Richman, although prestidigitator is included in both. Indeed, Carroll (p. xxvii) notes that American Heritage's five-million-word sample yielded a total of just 86,741 graphic types, but he estimates that this
sample is drawn from a total theoretical vocabulary of 606,906. Many millions of tokens more would be needed to turn up even half of these 606,906 types.

While it is theoretically possible, then, for a type-token curve such as (3) to reach a stable plateau (allowing us to read an author's total vocabulary directly from his type-token curve), in practice we must be satisfied with statistical extrapolations based upon small samples. Insofar as these extrapolations are valid, even short passages can provide significant evidence about the relative size of authors' vocabularies. For example, (4) shows the type-token curves for four texts: (a) the first 3000+ words of Evangeline (Longfellow), (b) the first part of "Big Two-Hearted River" (Hemingway), (c) "Macbeth" in Basic English, and (d) 2000 words of the Bible beginning with Genesis 2. (The curves a-d are listed from highest to lowest.)

(4) Type-Token Curves for Longfellow (highest curve), Hemingway, Basic English, and Genesis (lowest curve)

One of the most striking features of this graph is that the type-token curves for Hemingway and Basic English are virtually identical for the first 1100 words, after which they gradually diverge. This divergence is graphical evidence that Hemingway's vocabulary is (of course) larger than that of Basic English. But the remarkable coincidence of the two curves for the first 1100 words corresponds with our intuition when reading "Big Two-Hearted River" that its style is very tightly constrained, lexically as well as syntactically. For the first 1100 words at least, "Big Two-Hearted River" reads much like a story written in Basic English, in the sense that new vocabulary is introduced as slowly, and the syntax is as simple. The vocabulary in "Big Two-Hearted River" is nearly as familiar as that in Basic English, too, although occasionally Hemingway uses uncommon, specialized words which show an insider's knowledge of
his subject and demonstrate that his lexical restraint is self-imposed rather than forced upon him by a small vocabulary; for example, he refers to the tumpline of Nick Adams' pack and the sile of a dock (sile is listed in Webster's Third International Dictionary, but not in the latest Random House Unabridged Dictionary).

The famous Hemingway style becomes a mannerism in later works such as To Have and Have Not, but in "Big Two-Hearted River" it is a successful example of expressive form. Each perception, each action, is broken down into its basic elements, as though the whole were too much for any human to comprehend, or to bear. At one level, this disconnected, atomistic portrayal of so uncomplicated an experience as a fishing trip reflects a neurotic sensibility, suggesting that Nick Adams, a returning veteran, is using his trip partly as therapy, to help him recover from the shock of the First World War. Since Hemingway's death, it is difficult to avoid reading all of his work as an extended suicide note, but in this story, his lingering focus on small details goes beyond neurosis; it becomes ritualistic, turning a back-packing and fishing trip into a kind of religious ceremony.

The ritualistic quality of "Big Two-Hearted River" is reinforced by its stylistic association with the most famous example of "basic English" of all, the King James version of the Bible. The first chapter of Genesis has the lowest type-token curve of any text I examined for this study. Genesis 1 is 796 words long, but it uses a vocabulary of just 150 words (types)--considerably less than the 276-word vocabulary in the first 800 words of "Macbeth" in Basic English. This extraordinarily low figure results not only from the simple, restricted vocabulary used in Genesis, but also from its liturgical repetition:

11 And God said, Let the earth bring forth grass, the herb yielding seed, and the fruit tree yielding fruit after his kind, whose seed is in itself, upon the earth; and it was so.
12 And the earth brought forth grass, and herb yielding seed after his kind, and the tree yielding fruit, whose seed was in itself, after his kind: and God saw that it was good.

Obviously, this sort of repetition depresses the type-token curve for Genesis 1. We associate such recurrence with litany, song, and oral formulaic verse rather than with ordinary prose.

Subsequent chapters of Genesis are not quite so repetitious; consequently, they introduce vocabulary more rapidly--218 distinct words after 800 tokens (as opposed to just 150 for Genesis 1). Nevertheless, the type-token curve for the next 4800 words of Genesis is below that for "Macbeth" in Basic English. The difference between the two passages would be even greater if proper names were deleted from both texts, since the sample from Genesis includes a few of those famous biblical genealogies ("Irad begat Mehujael: and Mehujael begat Methusael: and Methusael begat Lamech," and so on). Hence, we might conclude (erroneously) from the type-token curves in (4) that the total vocabulary available for use to the translators of Genesis was smaller than that of Basic English. The graph in (4) provides visible evidence that this conclusion is mistaken: the curve for Genesis is lower than that for Basic English, but the two curves gradually converge rather than diverge, and if the samples were longer, the vocabulary of Genesis would eventually surpass that of Basic English. Nevertheless, the lexicon of Genesis is narrowly constrained, and in this respect it reads much like Basic English. Basic English in turn has a biblical flavor, as do many of Hemingway's stories (or so I would claim).

In their lexical parsimony, Genesis, Basic English, and "Big Two-Hearted River" are well outside the mainstream of English literature, the upper range for which is
suggested in (4) by the type-token curve for Longfellow's *Evangeline*. *Evangeline* makes use of a larger vocabulary than that in any other work examined in this study except for *Finnegans Wake*—which is perhaps the most lexically exuberant text in English literature, and certainly outside its mainstream! Longfellow's type-token curve in (4) is not only higher than those for Hemingway, Basic English, and Genesis, but it also diverges from them, implying that the vocabulary in *Evangeline* is drawn from a larger theoretical pool than is that of any other work in (4). This does not mean that Longfellow's vocabulary was necessarily larger than Hemingway's (although additional evidence suggests that it probably was); rather, the curves in (4) imply that Longfellow writing on this subject and in this genre drew upon a larger potential vocabulary than Hemingway did when writing "Big Two-Hearted River."

The graph (4) illustrates the wide variations possible for type-token curves in English literature. *Evangeline* is high for ordinary literary texts and Genesis is near the minimum. James Joyce, however, takes advantage of an even wider range of possibilities as (5) indicates:

(5) Type-Token Curves for James Joyce and Basic English

![Graph of Type-Token Curves](image)

(a) *Finnegans Wake* (highest curve)
(b) *Ulysses*
(c) Late Passage from *A Portrait of the Artist*
(d) Early Passage from *Portrait*
(e) "Macbeth" in Basic English

Predictably, the lowest type-token curve in (5) is the one for "Macbeth" in Basic English and the highest curve is for *Finnegans Wake*. The 1078-word vocabulary introduced in the first 2000 tokens of *Finnegans Wake* is thirty-one percent larger than that of its nearest competitor among the passages I examined (*Evangeline*: 820 types after 2000 tokens). This extraordinarily high figure for *Finnegans Wake* results both from Joyce's rapid change of topics (associated with his "stream of consciousness" style) and especially from his playful invention of new vocabulary. On every page, he
introduces dozens of words that have not appeared before or since in English literature. These two hallmarks of *Finnegans Wake*—narrative discontinuity and lexical innovation—make it a notoriously difficult book to read. Undoubtedly, then, the type-token curve for *Finnegans Wake* defines an upper limit for comprehensible English prose. It is an open question whether the theoretical vocabulary base for *Finnegans Wake* is finite. There is no obvious limit to a lexicon that admits compounds such as *upturnpikepointandplace* and hundred-letter anagrams for thunder in half-a-dozen different languages.

At the opposite end of the spectrum from *Finnegans Wake* is the famous opening section of *A Portrait of the Artist as a Young Man*, in which Joyce's prose suggests the consciousness and the limited vocabulary of Stephen Dedalus as a preschool child:

(6) Once upon a time and a very good time it was there was a moocow coming down along the road and this moocow that was down along the road met a nicens little boy named baby tuckoo . . .

This opening section is 273 tokens long, with a total vocabulary of just 139 types—slightly less than the vocabulary of 144 types used in the first 273 tokens of the Basic English passage graphed in (3). Hence, these statistics confirm our intuition as readers that the vocabulary in the beginning of *Portrait of the Artist* is as narrowly constrained as that for Basic English.

Stephen is a school boy in the second section of *Portrait of the Artist*, and Joyce's prose suggests Stephen's maturer, but still constrained, vocabulary, as in the second paragraph:

(7) Rody Kickham was a decent fellow but Nasty Roche was a stink. Rody Kickham had greaves in his number and a hamper in the refectory. Nasty Roche had big hands. He called the Friday pudding dog-in-the-blanket.

Somewhat surprisingly, the first 273 tokens of this section require even less vocabulary—130 types—than that in the first section of *Portrait of the Artist* (139 types after 273 tokens). The type-token curve for the first 2000 words of the second section of *Portrait* is plotted in (5). For the first 500 tokens, this curve nearly coincides with that for Basic English. After that point, the two curves diverge, slowly but steadily. Consequently, Joyce's passage has the flavor of Basic English, suggesting the limited vocabulary of a school boy, but his prose clearly draws upon a larger theoretical vocabulary than that of Basic English—larger even than that required for Hemingway's "Big Two-Hearted River."

The third type-token curve plotted in (5) is derived from the first 2000 tokens of the next-to-last section of *Portrait of the Artist*. Stephen is a young man with literary aspirations, and Joyce's prose suggests a maturer, literary vocabulary:

(8) What birds were they? He stood on the steps of the library to look at them, leaning wearily on his ashplant. They flew round and round the jutting shoulder of a house in Molesworth Street . . .

This passage includes a self-consciously learned discussion between Stephen and his friend Cranly. To suggest this maturer viewpoint, Joyce frees his vocabulary from the constraints that he imposed at the beginning of *Portrait*. After just 100 words, the type-token curve for this later passage begins to diverge from that of the earlier section of the novel. This divergence implies (accurately) that the later sections of *Portrait* draw upon a larger theoretical vocabulary than the earlier sections do. Certainly, from the reader's point of view, Joyce's unconstrained style has none of the flavor of Basic English.
The fourth type-token curve plotted in (5) is derived from the opening of *Ulysses*:

(9) Stately, plump Buck Mulligan came from the stairhead, bearing a bowl of lather on which a mirror and a razor lay crossed. . . .

This passage includes abrupt shifts: from external narration, to dialogue between Buck Mulligan and Stephen Dedalus, to Stephen's internal monologue. *Ulysses* is perhaps the most radically experimental novel written in English before *Finnegans Wake*, and it foreshadows some of the lexical exuberance of Joyce's last novel. Predictably, then, its type-token curve diverges rapidly from that for the beginning of *Portrait* and for "Macbeth" in Basic English. Indeed, for the first 300 tokens, *Ulysses* nearly keeps pace with *Finnegans Wake* itself; thereafter, their curves diverge. However, after its fast start for the first 800 tokens, the type-token curve for *Ulysses* begins to converge with, rather than diverge from, the curve for the late section of *Portrait of the Artist*. The results might be different for longer samples, but the type-token curves in (5) do not imply that the vocabulary of *Ulysses* is drawn from a larger theoretical base than that for *Portrait*. On the contrary, the total vocabulary implied by the two curves looks to be about the same. In one sense this result might be expected, since the two works, despite their very different feel, were written by the same author, and presumably Joyce's active English vocabulary did not change much between the writing of *Portrait* and *Ulysses*. Hence, these two curves illustrate that as the sizes of samples increase, type-token statistics tend to reflect the total vocabulary of authors rather than their style alone. That is, statistics for longer samples provide evidence for lexical competence as well as lexical performance.

The vocabulary in the early part of *Portrait* is artificially constrained, suggesting the restricted lexical competence of the young Stephen Dedalus. The vocabulary in *Finnegans Wake* is artificially enlarged with neologisms, reflecting the self-consciously expanded lexical competence of Joyce's late, experimental prose. By contrast, *Ulysses* and the later parts of *Portrait* draw upon roughly comparable lexicons. With respect to their vocabularies, the difference between these two works seems to be based largely upon differences in lexical style rather than lexical competence. These stylistic differences are readily perceptible to the reader: the more rapidly rising type-token curve for *Ulysses* reflects its more radical "stream of consciousness" style, resulting in more frequent and more striking narrative discontinuities (with quick changes in topic and vocabulary), whereas the later sections of *Portrait* read much more like a conventional, continuous narrative.

From the reader's point of view, short-term differences between type-token statistics are probably perceptible over ranges of only a few hundred words or less. These differences seem to be perceived primarily as contrasts in "semantic density"--differences in the number of topics introduced within a brief space. *Ulysses*, with its greater short-term semantic density, is a more challenging book to read than *Portrait* of the Artist is, even though the two books seem to be drawn from comparably large lexicons. Long-term differences in type-token statistics are probably perceptible primarily in terms of the variety and the difficulty of the vocabulary used. Over the short term, Hemingway reads much like Basic English, and indeed this is one of the hallmarks of his style. The difference over the long term seems to be this: the prose of Basic English is surprisingly pleasing, even quaintly charming at first, but its lack of variety begins to wear thin after several thousand words; whereas, at his best, Hemingway can sustain a reader's interest indefinitely. In addition, Hemingway sprinkles his prose with uncommon words such as tumpline and sile, implying that he holds in reserve a much larger lexicon than Basic English does. Of course, *Finnegans Wake*, with its profusion of new words, implies the largest lexicon of all.
A summary of the type-token statistics for Hemingway and Joyce, along with several other authors, is listed in (10) below. For each text, the number of types is shown at 200-token intervals for the first 400 - 2000 tokens.

(10) Table of Types Per 400 - 2000 Tokens in Twenty Literary Texts

<table>
<thead>
<tr>
<th>Tokens</th>
<th>400</th>
<th>600</th>
<th>800</th>
<th>1000</th>
<th>1200</th>
<th>1400</th>
<th>1600</th>
<th>1800</th>
<th>2000</th>
</tr>
</thead>
<tbody>
<tr>
<td>Genesis 1</td>
<td>93</td>
<td>130</td>
<td>150</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Genesis 2</td>
<td>129</td>
<td>182</td>
<td>218</td>
<td>250</td>
<td>278</td>
<td>317</td>
<td>349</td>
<td>385</td>
<td>420</td>
</tr>
<tr>
<td>Basic English</td>
<td>178</td>
<td>230</td>
<td>276</td>
<td>325</td>
<td>359</td>
<td>389</td>
<td>416</td>
<td>439</td>
<td>460</td>
</tr>
<tr>
<td>&quot;Big Two-Hearted&quot;</td>
<td>172</td>
<td>239</td>
<td>289</td>
<td>330</td>
<td>384</td>
<td>430</td>
<td>469</td>
<td>498</td>
<td>528</td>
</tr>
<tr>
<td>Aristotle's Ethics</td>
<td>163</td>
<td>215</td>
<td>277</td>
<td>331</td>
<td>385</td>
<td>431</td>
<td>473</td>
<td>511</td>
<td>562</td>
</tr>
<tr>
<td>Death in Afternoon</td>
<td>189</td>
<td>243</td>
<td>304</td>
<td>350</td>
<td>404</td>
<td>460</td>
<td>503</td>
<td>556</td>
<td>591</td>
</tr>
<tr>
<td>Portrait of Artist 2</td>
<td>173</td>
<td>236</td>
<td>310</td>
<td>373</td>
<td>425</td>
<td>485</td>
<td>528</td>
<td>580</td>
<td>615</td>
</tr>
<tr>
<td>Emma</td>
<td>218</td>
<td>301</td>
<td>365</td>
<td>427</td>
<td>476</td>
<td>526</td>
<td>576</td>
<td>616</td>
<td>658</td>
</tr>
<tr>
<td>&quot;Dry September&quot;</td>
<td>194</td>
<td>256</td>
<td>331</td>
<td>383</td>
<td>432</td>
<td>496</td>
<td>567</td>
<td>621</td>
<td>675</td>
</tr>
<tr>
<td>Golden Bowl</td>
<td>215</td>
<td>302</td>
<td>377</td>
<td>449</td>
<td>510</td>
<td>563</td>
<td>604</td>
<td>662</td>
<td>701</td>
</tr>
<tr>
<td>Tom Jones</td>
<td>225</td>
<td>300</td>
<td>364</td>
<td>431</td>
<td>494</td>
<td>552</td>
<td>615</td>
<td>669</td>
<td>735</td>
</tr>
<tr>
<td>&quot;The Dead&quot;</td>
<td>215</td>
<td>293</td>
<td>360</td>
<td>427</td>
<td>508</td>
<td>566</td>
<td>626</td>
<td>689</td>
<td>741</td>
</tr>
<tr>
<td>Portrait of Artist 3</td>
<td>201</td>
<td>270</td>
<td>343</td>
<td>431</td>
<td>500</td>
<td>578</td>
<td>650</td>
<td>709</td>
<td>758</td>
</tr>
<tr>
<td>House of Mirth</td>
<td>217</td>
<td>293</td>
<td>378</td>
<td>461</td>
<td>537</td>
<td>598</td>
<td>660</td>
<td>724</td>
<td>767</td>
</tr>
<tr>
<td>Oliver Twist</td>
<td>221</td>
<td>318</td>
<td>394</td>
<td>459</td>
<td>519</td>
<td>592</td>
<td>647</td>
<td>717</td>
<td>776</td>
</tr>
<tr>
<td>Sonnets (Shakes.)</td>
<td>231</td>
<td>320</td>
<td>406</td>
<td>473</td>
<td>534</td>
<td>604</td>
<td>679</td>
<td>726</td>
<td>780</td>
</tr>
<tr>
<td>Ulysses</td>
<td>245</td>
<td>333</td>
<td>414</td>
<td>477</td>
<td>551</td>
<td>613</td>
<td>669</td>
<td>750</td>
<td>792</td>
</tr>
<tr>
<td>Middlemarch</td>
<td>234</td>
<td>330</td>
<td>410</td>
<td>489</td>
<td>558</td>
<td>626</td>
<td>687</td>
<td>746</td>
<td>803</td>
</tr>
<tr>
<td>Evangeline</td>
<td>215</td>
<td>309</td>
<td>391</td>
<td>477</td>
<td>551</td>
<td>618</td>
<td>689</td>
<td>759</td>
<td>820</td>
</tr>
<tr>
<td>Finnegans Wake</td>
<td>276</td>
<td>394</td>
<td>496</td>
<td>601</td>
<td>708</td>
<td>794</td>
<td>893</td>
<td>1001</td>
<td>1078</td>
</tr>
</tbody>
</table>

This table documents the extraordinary lexical range of English literature, from the simple vocabulary and formulaic style of Genesis to the seemingly limitless lexical inventiveness of Finnegans Wake. At the lower end of the scale, Hemingway's radical lexical restraint in "Big Two-Hearted River" is striking; his vocabulary of 528 words is significantly less than Joyce's 615 in a section from Portrait that imitates the restricted speech of a school boy. Hemingway's lexical parsimony continues even in his more complicated prose in the thirteenth chapter of Death in the Afternoon, which introduces just 591 types in its first 2000 tokens.

More surprising, perhaps, is the restricted vocabulary used in W. D. Ross's English translation of the first book of Aristotle's Ethics, which introduces just 562 types in its first 2000 tokens. (It goes without saying that cross-language comparisons--say, between Basic English and Aristotle's original Greek--would be invalid.) Despite its constrained vocabulary, Ross's translation of Aristotle does not seem biblical or Hemingwaysque to the reader; instead, the Ethics exhibits characteristics of technical and scientific writing, with a restricted, narrowly defined vocabulary applied to a single subject, and no premium placed upon elegant variation. The type-token statistics for Aristotle seem less surprising when compared with those gathered by Kucera and Francis (1967), who list specific data for 51 texts grouped into twelve different categories. Each of these samples is roughly 2000 words long, and it is possible to estimate with reasonable accuracy the number of types after 2000 tokens for each text. Kucera and Francis include eight samples from "Learned and Scientific" writing, which vary from a low of approximately 622 types after 2000 tokens to a high of 737, with a mean of 685. This compares with an overall mean for the 51 passages of 756 types after 2000 tokens. That is, "Learned and Scientific" writing (which would certainly include Aristotle) typically relies upon a smaller vocabulary than writing in general does.
(The mean for Kucera and Francis's samples from "Government and Industry" is even lower: 661 types after 200 tokens.)

The table in (10) shows a significant gap between the "school boy" section from Portrait of the Artist (615 types) and Jane Austen's Emma (658 types). This gap suggests that a minimum for professional writing in English lies somewhere between 615 and 658--except for special cases such as liturgical prose, technical writing, and radically constrained styles such as Hemingway's. Kucera and Francis's statistics support this hypothesis; only five 2000-token samples require a vocabulary of less than 640 words--one government document (606 types), two learned/scientific articles (620 and 622 types), one article on religion (629 types), and one sample of romance fiction (636 types). Hence, 630-640 is a plausible floor for "normal" professional writing in English. Jane Austen's prose is near this floor, which is one reason that it remains so readable after nearly two centuries, whereas more recent authors such as Dickens and George Eliot are more difficult for modern readers.

Next after Emma is William Faulkner's "Dry September," with 675 tokens after 2000 types. This relatively low figure illustrates that the celebrated difficulties of Faulkner's prose are confined mostly to his experimental novels, such as The Sound and the Fury. By contrast, a short story such as "Dry September" owes more to the hard-boiled school of American naturalism than to the poetic prose of James Joyce.

The next author, Henry James, is also celebrated for his difficulty, but his 701 tokens after 2000 types of The Golden Bowl are actually below average for professional prose, confirming that the difficulties in his style are largely syntactic rather than lexical. On the whole, James's vocabulary tends toward the abstract rather than the concrete, and he prefers interpretive conversation and subtle nuance to detailed description and fast-moving action. All these factors tend to dampen his type-token curve. (By contrast, "Adventure and Western" fiction--presumably faster-moving than The Golden Bowl--averages 788 types after 2000 tokens in Kucera and Francis's samples.)

The type-token statistics for the next four passages cluster around the norm for professional prose: the opening of Fielding's Tom Jones (735 types), Joyce's "The Dead" (741 types), the late section from Portrait (758 types), and Wharton's House of Mirth (767 types). Dickens's Oliver Twist (776 types) and Shakespeare's Sonnets (780 types) are significantly above average but well within the normal range.

The next three texts are in the high normal range. One striking feature about the first of them, Ulysses, is that its type-token curve rises more quickly at first than do those for George Eliot's Middlemarch and Longfellow's Evangeline. After just 400 tokens, Ulysses introduces 245 types, thirty more than Evangeline and eleven more than Middlemarch. However, at 1000 tokens, the curves for Ulysses and Evangeline cross, and after 2000 tokens Evangeline (820 types) surpasses Ulysses (792 types). Again, longer samples might yield different results, but the evidence from these curves is that Middlemarch and Evangeline draw upon larger theoretical vocabularies than Ulysses does. Nevertheless, from a reader's point of view, Ulysses gives the impression of greater semantic density (and more rapid changes in topic) partly because its type-token curve rises more steeply during the first 800 tokens. Such short-term statistics tell us more about our short-term responses to a text than do longer-term statistics.

The type-token statistics in (10) and in Kucera and Francis 1967 suggest that the norm for an English text written by a single professional author is about 750 types for a 2000-token sample. Typical variations from this norm extend from a low of about 650 to a high of about 850. Kucera and Francis list only five texts written by a single author
that are above this range: one in the "Belles Lettres" category (861 types), one in "Adventure/Western" (882), one in "Romance" (902), and two in "Popular Lore"--868 and 913, the highest number being for an article by Kenneth Allsop, "The Bootleggers and their Era." Even this unusually high figure falls far short of the 1078 types found in the first 2000 tokens of Finnegans Wake.

To conclude, this discussion of one simple and relatively straightforward statistical measure, the type-token curve, illustrates some of the promise of, and the limitations upon, statistical analyses of literary style and competence. A single author, James Joyce, is capable of writing texts that lie well outside the normal 650 - 850 range. Hence, a statistician might be forgiven for concluding that Finnegans Wake and the first sections of Portrait of the Artist were written by different authors--forgiven because Joyce deliberately manipulates the very categories that usually define lexical style and competence. Indeed, as part of his "stream of consciousness" style, Joyce invents a new literary technique, which requires a new literary-critical term, implied linguistic competence, corresponding with Wayne Booth's implied author. That is, when he dramatizes the consciousness as well as the speech of his characters Stephen Dedalus, Leopold and Molly Bloom, Humphrey Chimpden Earwicker, and Anna Livia Plurabelle, Joyce invites us to infer their linguistic competence--their syntactic patterning and their vocabularies. Insofar as his narrators succeed in submerging themselves in his characters, Joyce manages to disguise his personal style and competence, which must be inferred from samples of all of his prose and speech (insofar as it has been recorded). Other authors are not so protean--Henry James is always Henry James, and smaller samples of his writing will suffice to estimate the size of his vocabulary.

The foregoing analysis suggests that relatively small samples (1000 tokens or so) can provide useful evidence about lexical style in a text, whereas longer samples can give useful evidence about the lexical competence implied by that text (that is the size of the theoretical vocabulary from which it is drawn). Having said this, I might close with a statistic that is certain to interest me more than the reader. Excluding graphs and numerals, the number of types after the first 2000 tokens of this article is 617, just two words more than Stephen Dedalus's as a school boy; but I flatter myself that this puts me in the "Learned and Scientific" category along with Aristotle, rather than with James Joyce's portrait of a child (however artistic).

Note
This article could not have been written without Paul L. Speckman's advice on statistics. Any mistakes that remain are, of course, my own responsibility.

Bibliography