Gilbert Youmans
Department of English
University of Missouri-Columbia
Columbia, MO 65211

A New Tool for Discourse Analysis: The Vocabulary-Management Profile[1]

Abstract

A Turbo Pascal program is used to generate vocabulary-management profiles (VMPs). The program counts the new vocabulary words introduced into a text over successive thirty-five-word intervals, and these numbers are then plotted at the midpoints of their intervals. The resulting VMPs show clearcut peaks and valleys that demarcate constituents in discourse. New episodes in narratives and new topics in essays show up as sharp rises preceded by deep valleys in the curve. This correlation between new vocabulary and new topics suggests that is plausible to interpret VMPs as information-management, as well as vocabulary-management, profiles for a discourse. The VMPs for passages from James Joyce and George Orwell show surprisingly regular alternations between new and repeated vocabulary, reflecting two competing principles that underlie the structure of all discourse--innovation and coherence.

1.0 Introduction

This article proposes a new quantitative method for analyzing the distribution of vocabulary in discourse. My approach is statistical; nevertheless, I presuppose a generative theoretical framework, even though generative grammarians have long been skeptical about statistical studies of vocabulary, as evidenced in early criticisms by Chomsky (1958) and Halle (1957 and 1958). Such criticisms notwithstanding, I argue that statistical data can be brought to bear upon issues of linguistic competence as well as style. Specifically, type-token curves (discussed in section 2) correlate with vocabulary size; and a new measure, the vocabulary-management profile (section 3), correlates surprisingly well with constituent boundaries and with information flow in discourse.

In the polemical atmosphere of contemporary linguistics, subdisciplines such as generative grammar, cognitive linguistics, sociolinguistics, computational linguistics, and discourse analysis are often portrayed as opposing camps. I assume, however, that these different approaches are supplementary rather than contradictory. For example, discourse analysts frequently criticize generative grammarians for focusing upon sentences to the exclusion of higher levels of discourse, and Grimes (1975:3) notes that this limitation forces Katz and Fodor 1963 ' . . . to adopt the fiction that in order to make a semantic interpretation of a text, all the sentences of the text have to be conjoined into a single supersentence, which is then amenable to

interpretation by projection rules'. Grimes rejects this procedure as 'a theoretical blind alley' (28). Nevertheless, his own approach is surprisingly similar to Katz and Fodor's, since he believes that ' . . . the grammatical trees that characterize sentences can be extended upward to groups of sentences, without essential discontinuity . . . ' (20). If Grimes is correct, then extended discourses have a constituent structure that is formally equivalent to Katz and Fodor's 'supersentences'.

For the purposes of this article, I assume that some variant of the supersentence approach is correct. In the simplest case, end-stop punctuation is a stylistic variant for the conjunction and. In more complicated cases, for example in conversation, explicit performatives may be necessary to recast a discourse such as (1) into a single sentence (2):

(1) "What time is it?"
"Ten o'clock."

(2) The first speaker asks, "What time is it?" and the second speaker answers, "Ten o'clock."

Using implicit conjunctions and performatives such as these, we can paraphrase any continuous discourse as a single sentence, even narratives as long as War and Peace. Such supersentences can be represented by tree diagrams, with all the usual coordinate and subordinate constituents. In claiming this, I do not mean to minimize the importance of higher-level constituents of discourse such as paragraphs, episodes, and chapters; rather, I mean to suggest that hierarchical groupings such as these can exist within, as well as across, sentence boundaries (at least in theory).

Conceptually, then, the problems raised by sentence structure and discourse structure seem to differ in degree rather than in kind. This is not to say, however, that discourse analysis is identical with syntactic analysis, as sentences such as (3) and (4) illustrate:

(3) After the house caught fire, it burnt to the ground.

(4) The house burnt to the ground after it caught fire.

The embedded adverbial clause precedes the matrix clause in (3) and follows it in (4); hence, the order of clauses is syntactically 'marked' in (3) and 'unmarked' in (4). From the viewpoint of narrative discourse, however, the opposite is true: the order of clauses is narratively unmarked in (3), because the cause precedes the effect, and marked in (4), because the cause follows the effect.

Contrasts such as these illustrate that the rule systems for syntax and for discourse are at least partly independent. Discourses are composed of syntactic constituents (as well as morphological and phonological ones); nevertheless, to classify linguistic elements as constituents of discourse is theoretically

distinct from classifying them as syntactic constituents. To cite one example, syntactically subordinate clauses may be perceived as superordinate (rather than subordinate) constituents of discourse. Indeed, it is a common aesthetic device (especially in literary narrative) to bury key elements of a story deep within subordinate clauses, as for example in some forms of irony and in Agatha Christie-style detective stories.

To put the matter another way, syntactic tree diagrams for supersentences need not be isomorphic with tree diagrams for the same supersentences viewed as discourse. My position on this point differs from the one taken in Polanyi 1985:19, which explicitly assumes such isomorphism in her list of rules for constructing an adequate paraphrase from a story text: 'List as Main Story Line Events only those <u>main clauses</u> which fulfill all event criteria . . . ' [emphasis added]. Polanyi's procedure works well for conversational narratives where clarity and sincerity are primary motivating factors, but it is less successful for literary narratives that rely upon subtlety, irony, and suspense. In the extreme case, the most salient event of a story is sometimes implied rather than stated overtly; that is, it does not appear explicitly in any clause, subordinate or superordinate. Faulkner's frequently anthologized short story 'A Rose for Emily' is one example. After attending Miss Emily's funeral, the townspeople break into an upstairs bedroom, where they discover a skeleton lying on the bed. The story ends as follows:

> (5) Then we noticed that in the second pillow was the indentation of a head. One of us lifted something from it, and leaning forward, that faint and invisible dust dry and acrid in the nostrils, we saw a long strand of iron-gray hair.

Everyone who reads 'A Rose for Emily' infers the same thing from (5): after poisoning her unfaithful lover, Miss Emily slept beside his corpse for many years, until she was a gray-haired woman. These inferred actions are so shocking that readers are likely to select them as the most salient events in the story, even though they are never explicitly mentioned by Faulkner's narrator.

Faulkner's story illustrates that discourse structure, like syntactic structure, exists on at least two levels, which I will call the explicit and the implied structures (in place of the deep structure, surface structure metaphor used in Langacker 1983 but now avoided in syntactic theory). Part of the aesthetic appeal of authors such as Faulkner lies in their deliberate manipulation of contrasts between these explicit and implied structures. Events that would normally occupy the foreground ('Did you hear about Miss Emily? She was sleeping with Homer

Barron's corpse') are relegated to the background, and events normally in the background are promoted to the foreground ('We saw a long strand of iron-gray hair').

This example from Faulkner illustrates that Polanyi's rules for paraphrasing stories generate skeletal versions of explicit, rather than implied, structures of discourse. Conceptually, her rules are similar to the Reduction Rules for music described in Lerdahl and Jackendoff 1983, which delete musically 'subordinate' notes in successive stages, thereby generating increasingly skeletal (but still recognizable) versions of musical phrases. Narrative (and musical) transformations do not enter into analyses such as these, which focus exclusively on explicit structure. Similarly, any analytic method which counts vocabulary, such as the one I describe in this article, is limited to measuring explicit, as opposed to implied, constituents of discourse. Granted, explicit constituent structure is only part of the story, but it is an important part, and any new quantitative measure that correlates significantly with explicit constituent structure is likely to be useful in more comprehensive theories of discourse as well.

1.1 Quantitative Studies of Vocabulary

Earlier quantitative studies (Herdan 1960, Kučera and Francis 1961, Carroll 1968, Carroll et al 1971, Francis and Kučera 1982) focus on the relation between the number of types and tokens in texts. Youmans 1990 expresses this relationship through type-token vocabulary curves, which are constructed in the following way: as a discourse unfolds, the total vocabulary (the number of types) is plotted against the total number of words (tokens) that have been used to that point in the text. For the first few words of a normal discourse, every new token is also a new vocabulary word; initially, then, the number of types equals the number of tokens. After the first repeated word, however, the number of tokens exceeds the number of types, and this difference increases with each repetition. Theoretically, if a discourse were long enough, the speaker's total vocabulary would be exhausted, and no new types could be added. Consequently, the type-token curve approaches a maximum limit that is determined by the size of the speaker's active vocabulary. Because of this characteristic, type-token curves can be used to estimate the size of the vocabulary from which discourses are drawn, although rather complicated statistical calculations are required to do so (for discussion, see Carroll 1968 and Carroll et al 1971).

In addition to being the basis for estimates of vocabulary-size, type-token curves might also be expected to correspond with patterns of information management in discourse.

For example, it seems plausible to predict that new topics in essays, new episodes in stories, and the like, should coincide with bursts of new vocabulary (showing up as hills on the type-token curve). Conversely, repetitions in vocabulary (plateaus on the curve) should signal a continuation, rather than a change, in topic. In actual practice, however, these hills and plateaus turn out to be barely visible, and direct inspection of type-token curves reveals little about the management of information in discourse.

Fortunately, the visibility of hills and plateaus on type-token curves can be enhanced with aid of more sophisticated analytic techniques. Borrowing a concept from differential calculus, I wrote a computer program to plot the number of new vocabulary words introduced in a 'moving' interval (usually thirty-five words long). The curves generated by this procedure show well-defined peaks and valleys, which can be interpreted as follows: an upturn in the curve signals an increase in new vocabulary at the end of the interval, whereas a downturn signals an increase in repetitions. The peaks and valleys on these curves prove to be surprisingly successful in signaling the ebb and flow of information in texts.

In more than 100 English narratives, essays, and transcripts examined so far, several clear tendencies have emerged. New vocabulary is introduced less often in the first part than in the second part of clauses and sentences: less often in subjects than in predicates, less often in topics than in comments, less often in given than in new information (Chafe 1974), less often in themes than in rhemes (Halliday and Hassan 1976). Furthermore, higher-level constituents of discourse tend to coincide with major peaks and valleys in this new derivative of the type-token curve. Sharp upturns after deep valleys in the curve signal shifts to new subject in essays, new episodes in stories, and so on.

The data for this article are derived primarily from written stories and essays; however, Tannen 1984:38 points out that in her study of conversation at a Thanksgiving dinner, ' . . . the most useful unit of study turned out to be the episode, bounded by changes of topic or activity, rather than, for example, the adjacency pair or the speech act.' Hence, it is plausible to suppose that this new derivative of the type-token curve generates model information-management profiles for conversation as well as for written texts. However, because new vocabulary and new information are only correlated rather than directly related, I will refer to these new curves as vocabulary-management (rather than information-management) profiles (VMPs). Section (3) illustrates the typical characteristics of these profiles in selected discourses.

1.2 Counting Words with a Computer

Certain decisions--and compromises--must be made in any computerized study of vocabulary such as this one. A complete lexical analysis of a text would divide words into their component morphemes. However, in a preliminary quantitative study, the simplest statistic to obtain is the number of graphic words, as defined in Francis and Kučera 1982:3: 'Graphic word: a string of contiguous alphanumeric characters with space on either side; may include hyphens and apostrophes but no other punctuation marks.' Again in order to simplify computer analysis, 'A "distinct word" (type) can also be simply defined as a set of identical individual words' (Kučera and Francis 1967:xxi). That is, all and only identical alphanumeric strings count as 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 one word. Homographs such as these are more troublesome in theory than in practice, since contrasting pairs such as Polish/polish rarely occur in the same discourse, and when they do, one of them can be respelled: Po-lish/polish.

Francis and Kučera 1982 goes beyond a purely mechanical definition of word, grouping graphic words into lemmas such as be, which subsumes the inflectional forms been and being, the suppletive forms am, is, was, were, and even spelling and dialect variants such as are/ah, and were/wuh. Presumably, new topics in discourse are correlated more closely with new lemmas than with new graphic words; the change from Gandhi to Gandhi's, for instance, is not likely to be interpreted as a change in topic. It might be best to ignore derivational affixes, too, grouping words such as transport and transportation under the same topic. Synonyms such as unmarried and single also might be grouped together.

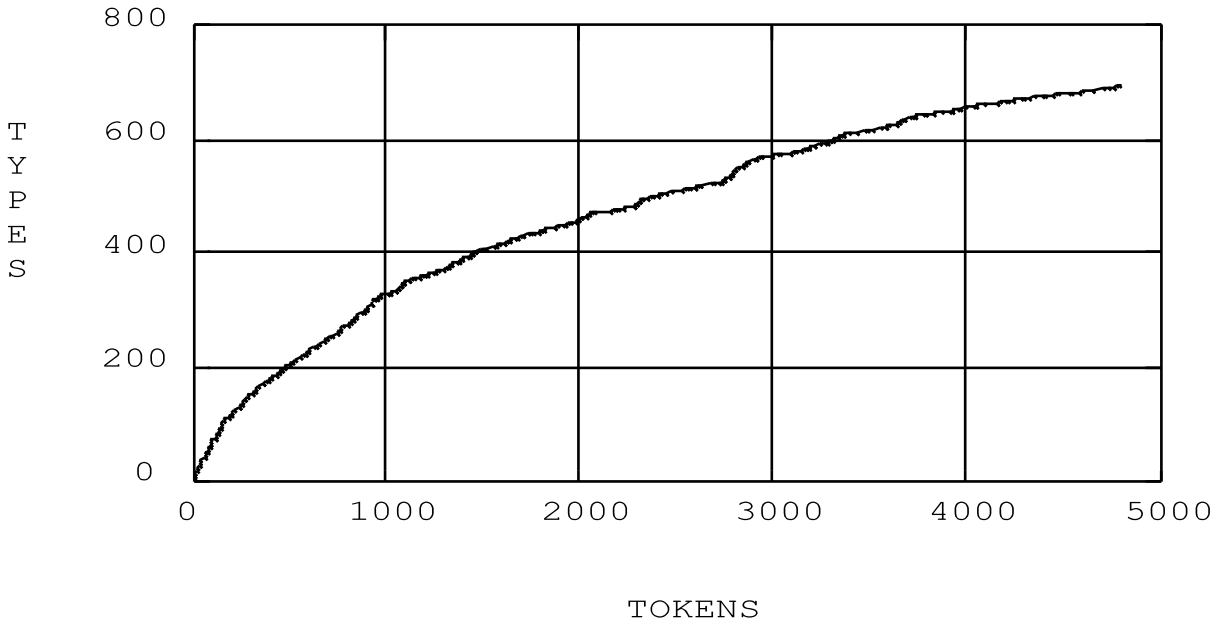
In this article, however, I begin by using the definition that simplifies computer analysis: a word (type) is any distinctive string of alphanumeric characters (including hyphens and apostrophes but excluding other punctuation) that is preceded and followed by a space. It turns out that this computer-friendly definition is surprisingly successful in signaling changes in topic. Later, in section (3), I compare the vocabulary-management profiles generated by this definition with those in which a single symbol x is substituted for all syntactic function words. This refinement has a significant effect on VMPs for about the first 500 words of text, but little effect thereafter. In section (3), I also test the effect of replacing all semantic content words with their lemmas and the effect of

conflating all synonyms. These changes have little visible effect on the VMP: the overall curve is slightly lower, but its shape remains nearly the same. Hence, this further refinement of VMPs seems to unnecessary for most purposes.

2.0 The Type-Token Curve

Youmans 1990 analyzes type-token curves for twenty different texts by thirteen different authors, including the following curve for 'Macbeth', which is a translation into Basic English by T. Takata of a passage from Charles Lamb's Stories from Shakespeare (Ogden, 286-298):

(6) Plot of Types Versus Tokens for 'Macbeth' in Basic English



The first sentence of this passage will serve to illustrate how the curve in (6) is derived:

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

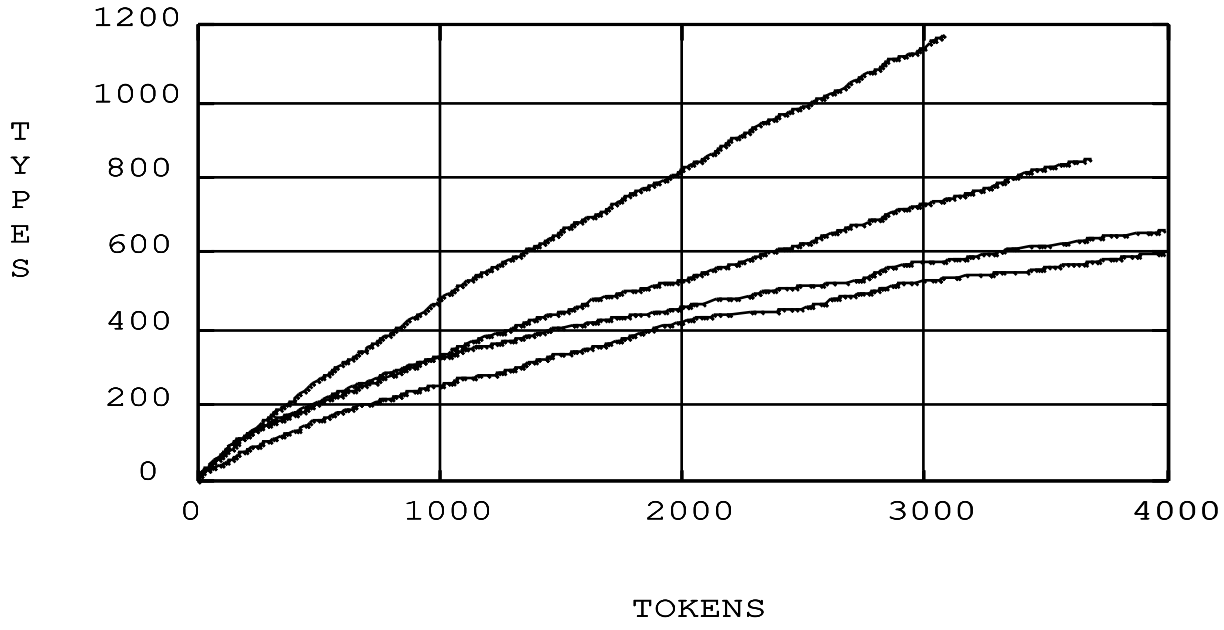
In (7), the number of types equals the number of tokens until the first repetition: the sixth word the. Hence, after six tokens, the number of types equals five. The next repetition is the thirteenth word was; thus, after thirteen tokens, the number of types equals eleven, and so on for the remainder of 'Macbeth'.

The curve for a concordance or any other vocabulary list is

a straight line with types equaling tokens at every point. However, the type-token curves for all normal discourses resemble the one in (6). They begin as a straight line, with types equaling 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. Thus, as the number of tokens approaches infinity, the number of types approaches the total active vocabulary of the author.

Although impossibly long passages would be needed to exhaust the vocabulary of an adult native speaker of English, plausible claims about the relative size of authors' vocabularies can be based upon even short passages. For example, Youmans 1990 compares 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) 4000 words of the King James translation of the Bible beginning with Genesis 2. (The curves (a)-(d) are listed from highest to lowest.)

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



(a) Evangeline, Longfellow (highest curve)
(b) 'Big Two-Hearted River', Hemingway
(c) 'Macbeth', Basic English
(d) Genesis 2 and ff. (lowest curve)

In (8) the middle two curves (for Hemingway and Basic English) are nearly identical for the first 1100 words, after which they gradually diverge. The early similarity of the two curves corresponds with readers' intuitions that, over the short term, Hemingway's prose reads much like Basic English; however, the later divergence between the two curves is graphical evidence that Hemingway's vocabulary is larger than that of Basic English.

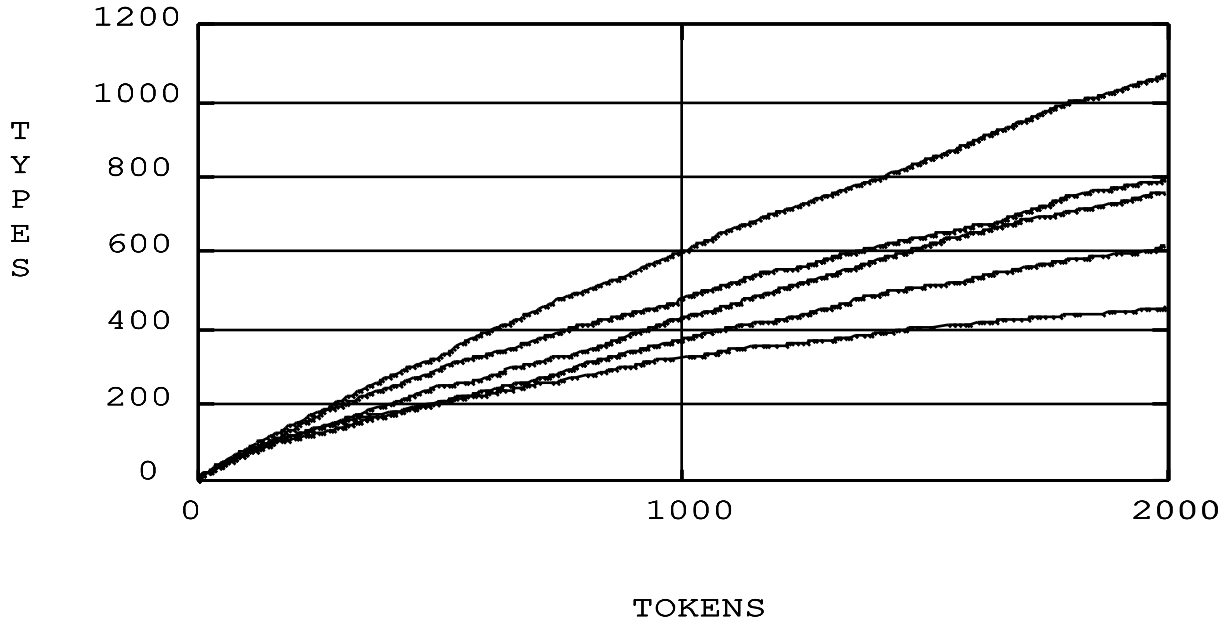
Genesis II and following--partly because of its simple, restricted vocabulary, and partly because of its repetitious, oral-formulaic style--has a lower type-token curve even than that for 'Macbeth' in Basic English. Hence, we might be tempted to conclude that the total vocabulary available for use to the

translators of Genesis was smaller than that of Basic English. The graph in (8) provides visible evidence that this conclusion is mistaken: the curve for Genesis remains below that for Basic English, but the two curves gradually converge rather than diverge. Presumably, if the samples were longer, the vocabulary of Genesis would eventually surpass that of Basic English.

The highest curve in (8) is that for Longfellow's Evangeline. This curve is not only higher than those for Hemingway, Basic English, and Genesis, but it also diverges from them, implying that the vocabulary in Evageline is drawn from a larger theoretical pool than that of any other work in (8). This does not mean that Longfellow's total vocabulary was necessarily larger than Hemingway's (although additional evidence suggests that it probably was); rather, the curves in (8) imply that Longfellow writing on this subject, in this genre, and for this audience drew upon a larger potential vocabulary than Hemingway did when writing 'Big Two-Hearted River'. An accurate estimate of an author's total vocabulary would require representative samples of speech and writing on different subjects, in different genres, and for different occasions.

James Joyce's prose is an excellent illustration of the danger of trying to estimate an author's total vocabulary from a single sample, as the type-token curves in (9) illustrate;

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



(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

The middle curves in (9) (for Ulysses and a late passage from Portrait) are roughly parallel, even convergent; hence, these two passages seem to be representative samples of Joyce's normal literary vocabulary. By contrast, the variation between Joyce's highest and lowest curves is extraordinary--ranging from the 1078 types introduced in the first 2000 tokens of Finnegans Wake to just 615 in the early passage from Portrait. This wide variation results from Joyce's deliberate manipulation of the 'implied lexical competence' of his narrators. The early sections in Portrait suggest the consciousness (and the limited vocabulary) of a young boy, whereas FW, with its polyglot puns and invented vocabulary, suggests a universal dream language with an almost unlimited lexicon. Consequently, when estimating the size of Joyce's normal literary vocabulary, we might want to imitate Olympic diving judges, throwing out his highest and lowest scores in (9) and averaging the middle two.

Used cautiously, type-token curves can be the bases for

plausible judgments about relative vocabulary-size, but they tell us almost nothing about information management. The ebb and flow of new vocabulary is very difficult to detect in (6), (8), and (9); some minor bumps and hills are visible, but their boundaries are too imprecise to provide clearcut evidence of shifts in topic, much less the relative magnitude of these shifts.

3.0 Vocabulary-Management Profiles (VMPs)

Section (2) illustrates that type-token curves are too smooth to signal changes in topics clearly; hence, a more sensitive quantitative indicator is needed. In differential calculus, the instantaneous rate of change of a function (its 'velocity') is given by its first derivative, dy/dx, but differentiation is impossible for type-token curves because we do not know the differentiable equations (if any) that define the curves. Furthermore, type-token curves are derived from discontinuous rather than continuous data (because discourses are composed of separate words). Hence, any attempt to compute the 'instantaneous' rate of change of type-token curves over 'infinitesimal' intervals would be pointless. The relevant statistic is not dy/dx, but rather $\Delta y/\Delta x$, the rate of change over a finite interval (where Δy equals the number of new types, and Δx equals the number of new tokens, in the interval).

The ratio $\Delta y/\Delta x$ can vary from a maximum of 1.0 (if all of the tokens in the interval are new types) to a minimum of 0.0 (if no tokens in the interval are new types). In vocabulary studies, the smallest possible interval is a single word, with $\Delta x = 1$. If the token in that interval is a new type, then $\Delta y = 1$, and $\Delta y/\Delta x = 1/1 = 1$. If the token is a repeated word, then $\Delta y = 0$, and $\Delta y/\Delta x = 0/1 = 0$. Thus, for single-word intervals, $\Delta y/\Delta x$ equals either one or zero, which is to say that this ratio merely tells us what we know already--that a given token is or is not a new type. Consequently, intervals longer than one word are needed if the ratio $\Delta y/\Delta x$ is to yield any new information.

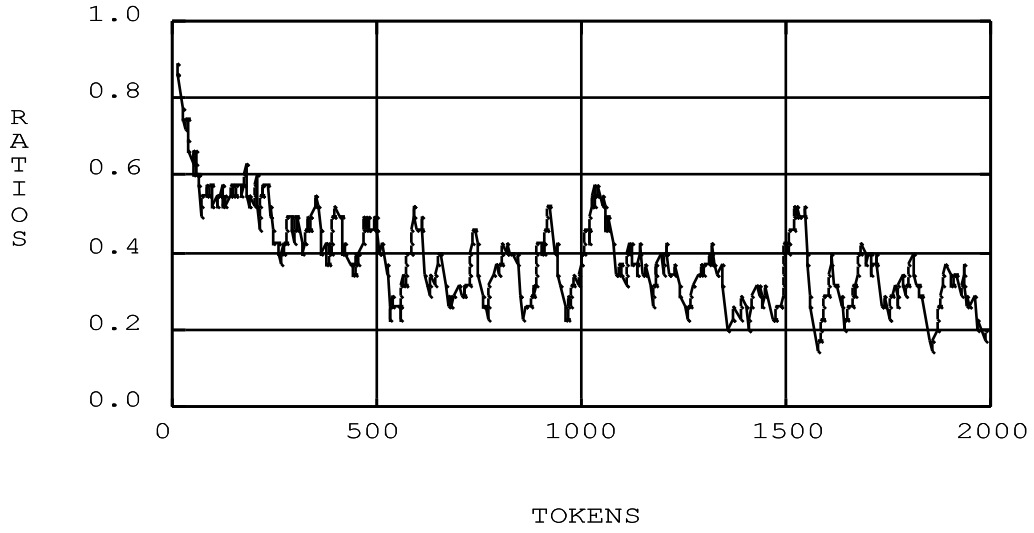
At the opposite extreme, if Δx is extended until it equals the length of the entire text, then $\Delta y/\Delta x$ is just the type-token ratio for the discourse as a whole. Obviously, a single statistic cannot reveal anything about information flow. Hence, in order to be useful, the interval Δx must be greater than one but less than the length of the discourse. For this article, I experimented with five different intervals: 11, 25, 35, 51, and 101 words. I wrote a computer program that counts the number of new types, Δy, introduced over a moving interval, Δx; then I plotted the values for Δy at the midpoint of the intervals Δx. Thus, for $\Delta x = 35$, the number of new types introduced in words 1-35 is plotted at the 17th token; the number of new types introduced in words 2-36 is plotted at the 18th token, and so on

for the remainder of the text.[2]

Longer intervals, such as 101 words, generate 'smoother' VMPs; their peaks are not as high, and their valleys are not as low. Longer intervals are also less sensitive to short-term variations in the rate of introduction of new vocabulary. In this sense, shorter intervals are more 'accurate'. However, as intervals become too short, $\Delta y/\Delta x$ often falls to zero, especially at the ends of texts, where new vocabulary is introduced less frequently. Obviously, when the ratio $\Delta y/\Delta x$ drops to zero, it no longer signals changes in the rate of introduction of new vocabulary. Hence, Δx can be too short as well as too long.

For the texts examined in this article, intervals of thirty-five words proved to be a good compromise, exhibiting most of the virtues and few of the deficiencies of longer and shorter intervals. This is the interval chosen for (10), which is the VMP for the first 2000 words of James Joyce's short story 'The Dead'.

(10) The Ratios $\Delta y/\Delta x$ for James Joyce's 'The Dead' ($\Delta x = 35$)



There is a striking contrast between the jagged appearance of the VMP in (10) and the relatively smooth type-token curves for the same author in (9). The curve in (10) shows a series of clearcut peaks and valleys; later, I will demonstrate that major valleys on VMPs correlate very closely with the boundaries between major constituents of discourse.

Another striking characteristic of the curve in (10) is its surprising regularity; after about 250 tokens, peaks and valleys occur once every hundred words or so. This regularity suggests that there is a rhythmic alternation between new and repeated vocabulary in the typical well-crafted story, an alternation that parallels the periodic ebb and flow of new information in a text, the regular pattern of innovation and elaboration that is necessary to give both forward momentum and coherence to discourse.

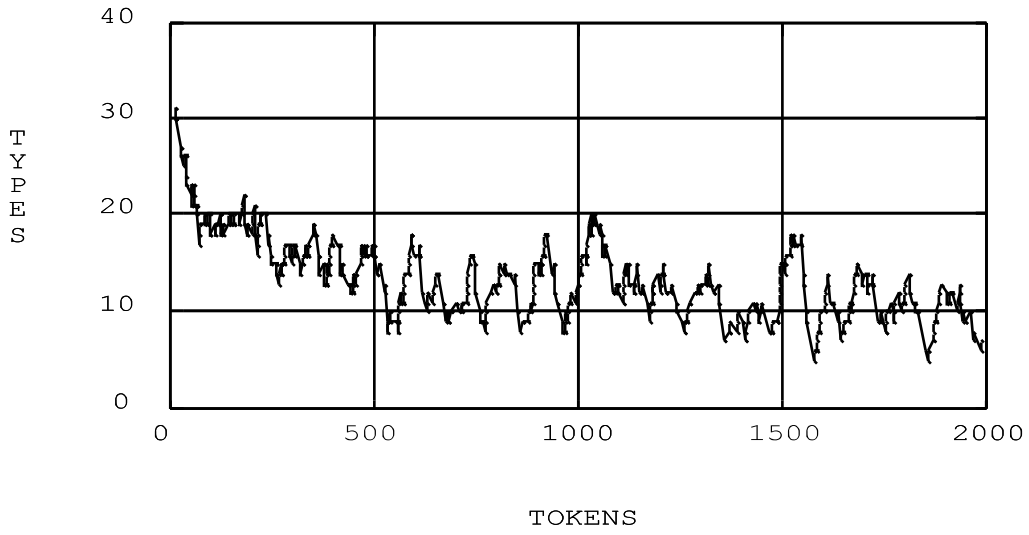
The curve in (11) illustrates the effect of increasing the interval Δx from thirty-five to 101 words. The passage is the same as the one plotted in (10).

(11) Ratios ∆y/∆x for James Joyce's 'The Dead' (∆x = 101)

The VMP in (11) loses considerable detail. Most notably, it fails to reveal the regular 100-word alternation between peaks and valleys that is so obvious in (10). This illustrates a general principle about VMPs; they cannot detect patterns shorter than ∆x. On the other hand, (11) shows even more clearly than (10) that the peaks occurring shortly after 1000 and 1500 words are especially prominent ones. This illustrates another characteristic of VMPs: longer intervals for ∆x are more useful for detecting long-term patterns in discourse than shorter intervals are.

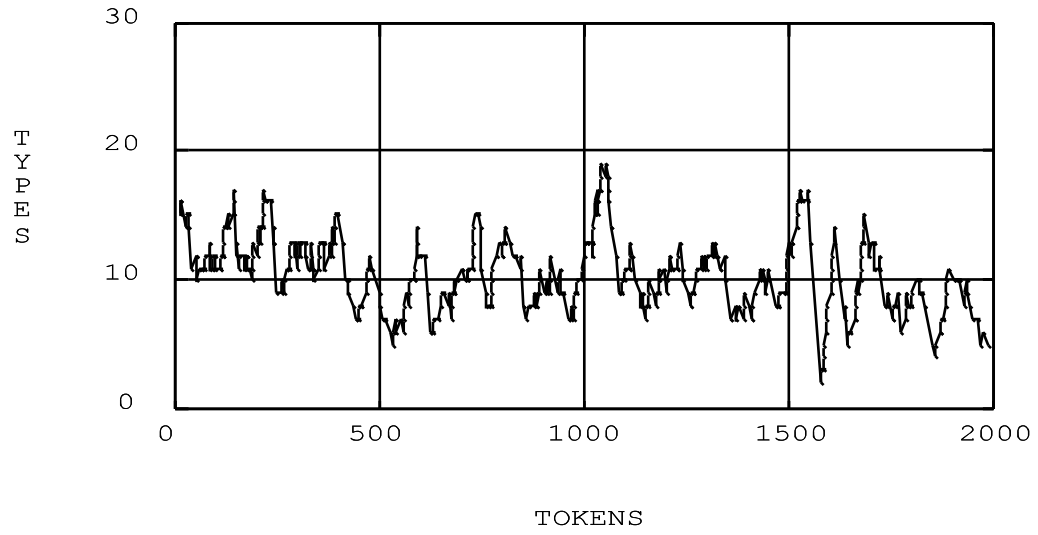
The remainder of this article focuses on intermediate-term intervals, with ∆x = 35. Once the interval for ∆x is fixed, we no longer need to compute the ratio ∆y/∆x, because dividing by a constant affects only the scale, and not the shape, of the VMP. Hence, in the curves below, I do not plot ∆y/∆x, but ∆y, the number of new types introduced over an interval of thirty-five words. This number can vary from a minimum of zero to a maximum of thirty-five. Consequently, the vertical scale in (12) differs from the one in (10), but otherwise the two curves are identical:

(12) The VMP for 'The Dead' (Δx = 35)



The VMP in (12) treats all graphic tokens as equals: for example, the first occurrences of the and Gabriel both count as new types. From the point of view of information management, this egalitarianism is undesirable: Gabriel (the name for the main character in the story) denotes a topic of discourse, whereas the does not. The boundary between topical and nontopical words is fuzzy rather than well-defined; but for experimental purposes, it is convenient to assume that syntactic function words do not denote topics, whereas semantic content words (nouns, main verbs, adjectives, and some adverbs) do denote topics. Given this assumption, we can generate a topical skeleton for a discourse simply by substituting a single symbol such as x for all its function words. This substitution reduces the total vocabulary of the first 2000 tokens of 'The Dead' by about 170 words, from 742 distinct types to about 572 (depending upon which words are designated as function words). The VMP for this skeletal version of the passage is plotted in (13).

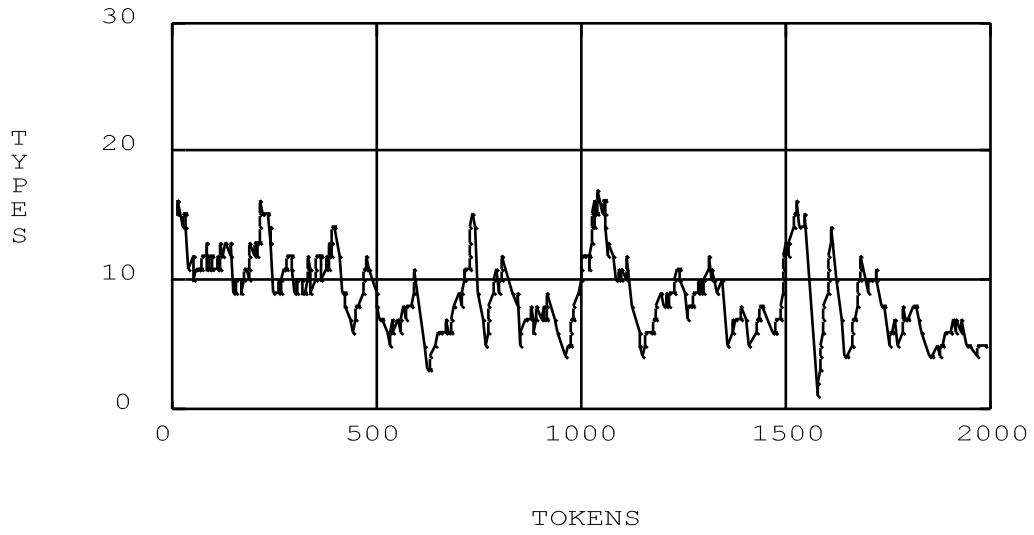
(13) The VMP for 'The Dead' with Function Words Replaced by *x*



The two curves (12) and (13) contrast significantly for their first 400 tokens; afterward, the VMPs are remarkably similar. The reason for this is that syntactic function words are relatively few in number but high in frequency. Consequently, the first occurrences of function words tend to cluster near the beginning of a discourse and taper off quickly thereafter. The effect of function-word vocabulary on VMPs becomes less and less significant as discourses unfold. Thus, from the point of view of information management, the chief advantage of the VMP in (13) over the one in (12) is that (13) reveals two clear peaks in the first 250 tokens of the passage. Later, after 500 tokens, the two curves give very similar signals: their major peaks and valleys nearly coincide, although (13) is slightly lower than (12) overall.

The next step is to conflate the inflected and derived forms of the semantic content words that remain in the skeletal version of 'The Dead' profiled in (12). Replacing all inflected words with their stems and deleting selected derivational affixes such as -ly reduces the total vocabulary in the topical skeleton by only about 87 additional types, from 572 to 485 words. The VMP for this reduced version of the passage is shown in (14):

(14) VMP for 'The Dead' with Function Words Replaced by *x* and Affixes Deleted



The overall curve in (14) is lowered again, but otherwise its VMP is very similar to the one in (13). The reason for this is that the loss of 87 vocabulary words in (14) is distributed more or less evenly throughout the text. Consequently, the VMP in (14) gives nearly identical signals with the one in (13): the major peaks and valleys on the two curves coincide almost exactly.

The final step is to conflate the synonyms and the near-synonyms that remain in the topical skeleton for (14). This step reduces the total vocabulary in the passage by only about a dozen words. The effect on the VMP is barely visible.

To summarize: deleting affixes and conflating synonyms appears to be an unnecessary refinement in VMPs if their purpose is to provide graphical signals for major shifts in the flow of information in English discourses.[3] On the other hand, distinguishing between function words and content words does have a significant effect, particularly for the first 500 words of a text. Consequently, I revised my original computer program to count the 200 most common function words (listed in Carroll et al 1971) as repeated rather than new vocabulary. This version of the program successfully recognizes 160 of the 170 function words in the first 2000 words from 'The Dead'; consequently, its VMP is

almost identical with the one in (13). I will use this revised program to generate all remaining VMPs. As an instrument for measuring information flow in discourse, this version of the VMP is a bit like a wind sock at an airport; it is surprisingly effective in telling us which way, and even how hard, the wind is blowing.

Bibliography

Carroll, John B. 1968. Word frequency studies and the lognormal distribution. Proceedings of the conference on language and language behavior, ed. by E. M. Zale, 213-235. New York: Appleton-Century-Crofts.

Carroll, John B., Peter Davis, and Barry Richman. 1971. Word frequency book. New York: American Heritage.

Chafe, Wallace L. 1974. Language and consciousness. Language 50:1. 111-133.

Chafe, Wallace L. 1987. Cognitive constraints on information flow. Typological studies in language. Ed. Russell S. Tomlin. Philadelphia: John Benjamins. 22-51.

Chomsky, Noam. 1958. Review of Langage des machines et langage humain. Language 34:1. 99-105.

Francis, W. Nelson, and Henry Kučera. 1982. Frequency analysis of English usage: lexicon and grammar. Boston: Houghton Mifflin.

Halle, Morris. 1957. In defence of the number two. Studies presented to Joshua Whatmough. The Hague: Mouton.

Halle, Morris. 1958. Review of Language as choice and chance. Kratylos 3. 20-28.

Halliday, M. A. K., and Ruqaiya Hasan. 1976. Cohesion in English. London: Longman.

Hendricks, William O. 1977. 'A Rose for Emily': a syntagmatic analysis. PTL 2. 257-295.

Herdan, Gustav. 1960. Type-token mathematics: a textbook of mathematical linguistics. 's-Gravenhage: Mouton.

Herdan, Gustav. 1962. The calculus of linguistic observations. 's-Gravenhage: Mouton.

Kučera, Henry, and W. Nelson Francis. 1967. Computational analysis of present-day English. Providence, Rhode Island: Brown University Press.

Labov, William. 1972. Language in the inner city. Philadelphia: U. of Pensylvania Press.

Labov, William. 1980. The social origins of sound change. Locating language in time and space, ed. by William Labov, 251-265. Orlando, Fla.: Academic Press.

Ogden, C. K. 1934. The System of Basic English. New York: Harcourt, Brace and Co.

Polanyi, Livia. 1985. Telling the American story: a structural

and cultural analysis of conversational storytelling. Norwood, NJ: Ablex.
Tannen, Deborah. 1984. Conversational style: analyzing talk among friends. Norwood, N. J.: Ablex Publishing Corp.
Thorndike, Edward L., and Irving Lorge. 1944. The teacher's word book of 30,000 words. New York: Teachers College, Columbia Univ.
Youmans, Gilbert. 1990. Measuring lexical style and competence: the type-token vocabulary curve. Style 24 584-99.