# A NEW TOOL FOR DISCOURSE ANALYSIS: THE VOCABULARY-MANAGEMENT PROFILE

GILBERT YOUMANS

*University of Missouri-Columbia*

A computer is used to count the number of new vocabulary words introduced into a text over a moving interval thirty-five words long. The number of new words in each successive interval is plotted at the midpoint of the interval, generating a curve called the Vocabulary-Management Profile (VMP). Analysis of VMPs for passages from James Joyce and George Orwell illustrates that clearcut peaks and valleys on VMPs correlate closely with constituent boundaries and information flow in discourse. VMPs for these authors show surprisingly regular alternations between peaks and valleys (that is, between new and repeated vocabulary), reflecting two competing principles that necessarily underlie all normal discourse: innovation and coherence.*

**1.** INTRODUCTION. This article proposes a new quantitative method for analyzing the distribution of vocabulary in discourse. It does not attempt to develop a comprehensive linguistic theory of discourse nor to offer evidence in favor of one competing theory over another one. Rather, it describes a statistical tool that is compatible with a wide variety of linguistic approaches to discourse. This first illustrative article focuses on relatively uncontroversial examples—carefully-edited texts by authors such as James Joyce and George Orwell, in which discourse boundaries are often marked by punctuation, paragraph breaks, and the like. When identifying discourse constituents, I have been guided by such punctuation and paragraphing as well as by principles of discourse analysis developed in Labov 1972, Chafe 1974, 1987, Grimes 1975, Longacre 1983, Polanyi 1985, and elsewhere. Whenever possible, I have tried to validate constituent boundaries by joint evidence: paragraph breaks AND the principles in Chafe AND Longacre AND Polanyi, not paragraphs OR Chafe OR Longacre OR Polanyi. The intent is to make identification of constituents as uncontroversial as possible—to place boundaries where trained readers of English literature are most likely to perceive them.

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 (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 §2) correlate with the size of a speaker's lexicon; and a new measure, the Vocabulary-Management Profile (§4), correlates surprisingly well with constituent boundaries and with information flow in discourse.

In contemporary linguistics, subdisciplines such as generative grammar, cognitive linguistics, sociolinguistics, computational linguistics, and discourse

* I am grateful for Paul L. Speckman's comments on the statistical arguments presented in this essay. Any mistakes are, of course, my own responsibility.

analysis are often portrayed as opposing camps. I assume, however, that these different approaches are compatible 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) complains that this limitation forces Katz & 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 & 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 & Fodor's 'supersentences'.

For the purposes of this article, I assume that some form of the supersentence approach is correct. In the simplest case, end-stop punctuation is a stylistic variant for a conjunction *and, and then, and then therefore,* or the like. In more complicated cases, explicit speech tags may be necessary to recast a discourse into a single sentence:

(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 explicit conjunctions and speech tags 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 coördinate 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.

**1.1.** QUANTITATIVE STUDIES OF VOCABULARY. Earlier quantitative studies (Herdan 1960, Kučera & Francis 1967, Carroll 1968, Carroll et al. 1971, Francis & 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 the 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 & Hasan 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 subjects 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 4 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 by Francis & 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 & 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, a 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 if they do, one of them can be respelled: for example, *Po-lish/polish*.

Francis & Kučera 1982 goes beyond a purely graphical 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*, and *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, 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 §4, 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 §4 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, for most purposes, this further refinement of VMPs seems to be unnecessary when analyzing English discourse.

**2.** THE TYPE-TOKEN CURVE. Youmans 1990 analyzes type-token curves for twenty different texts by thirteen different authors, including the curve in Fig-

ure 1 for 'Macbeth', which is T. Takata's paraphrase into Basic English[1] of a passage from Charles Lamb's *Stories from Shakespeare* (Ogden 1934:286–98). The first sentence of this passage will serve to illustrate how the curve in Fig. 1 is derived:

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

In 3 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'.

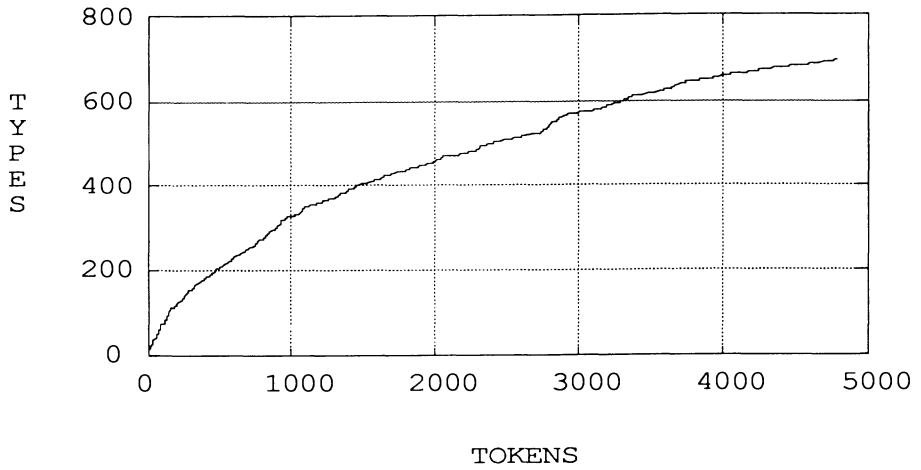


FIGURE 1. Plot of types versus tokens for 'Macbeth' in Basic English.

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 Fig. 1. 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

[1] Basic English was invented by Ogden and his associates to serve as an international language. Unlike other such failed attempts, Basic English is not an artificial language but a subset of a natural one. Its vocabulary and, to a lesser extent, its grammar are limited, but all of its sentences are English. Because discourses in Basic English draw upon a known, unusually small, vocabulary, they are useful statistical resources. A narrative such as Takata's, for example, establishes a plausible lower limit for type-token curves for normal English discourses written by and for adult native speakers.

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, Figure 2 (taken from Youmans 1990) shows the type-token curves for four texts: (a) the first 3000 words of Longfellow's *Evangeline*, (b) the first part of Hemingway's 'Big Two-Hearted River', (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.)

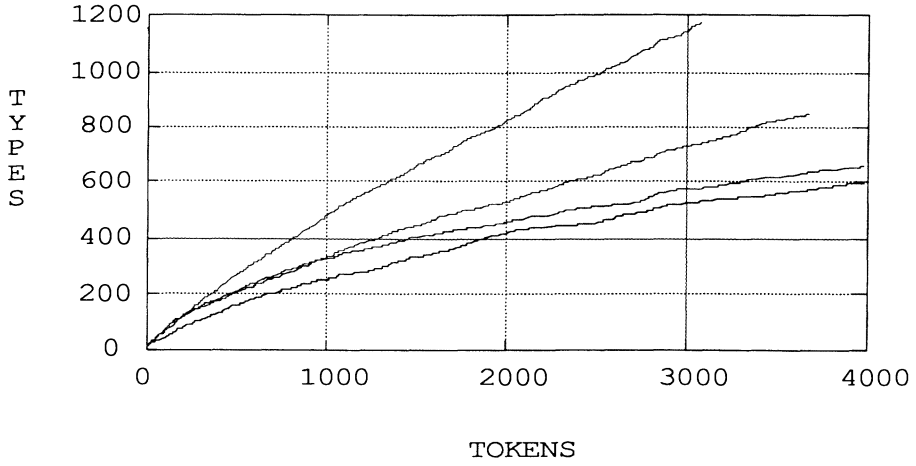


FIGURE 2. Type-token curves for (a) *Evangeline*, by Longfellow (highest curve), (b) "Big Two-Hearted River', by Hemingway (next highest curve), (c) 'Macbeth' in Basic English (next highest curve), and (d) Genesis 2 and ff. (lowest curve).

In Fig. 2 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 2 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 to the translators of Genesis was smaller than that of Basic English. Fig. 2 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 Fig. 2 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 *Evangeline* is drawn from a larger theoretical pool than that of any other work plotted in Fig. 2. 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 Fig. 2 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 Figure 3 illustrate. The middle curves in Fig. 3 (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* (*FW*) 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

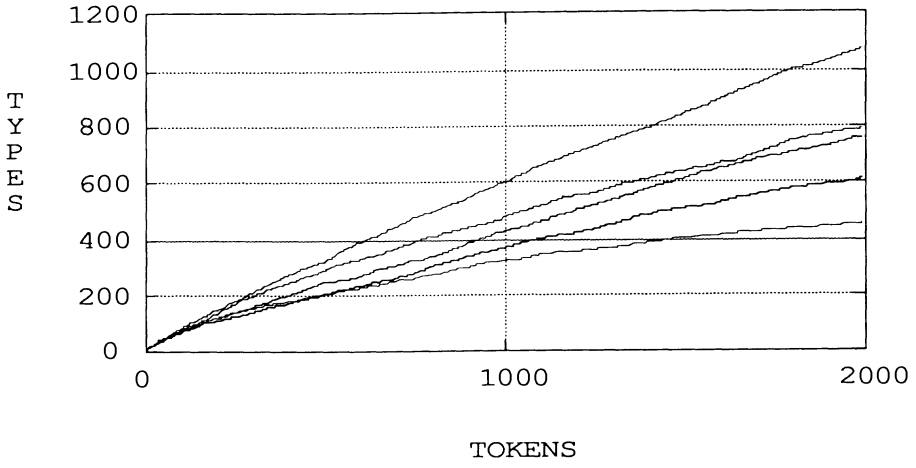


FIGURE 3. Type-token curves for James Joyce and Basic English: (a) *Finnegans Wake* (highest curve), (b) *Ulysses* (next highest curve), (c) a late passage from *A Portrait of the Artist* (next highest curve), (d) an early passage from *Portrait* (next highest curve), (e) 'Macbeth' in Basic English (lowest curve).

size of Joyce's normal literary vocabulary, we might want to imitate Olympic diving judges, throwing out his highest and lowest scores in Fig. 3 and averaging the middle two.

**3.** NEW VOCABULARY AND NEW INFORMATION. Prince 1981 identifies three different levels of 'givenness' arising from (a) predictability/recoverability, (b) saliency, and (c) shared knowledge. Chafe 1987 also subdivides the given-new dichotomy into three degrees of informativity: given, accessible, and new information, which Chafe equates with active, semi-active, and inactive concepts. Chafe contends that active knowledge resides in short-term memory, which is limited to about two seconds of speech—approximately five to nine words and one intonation unit. He suggests that pauses between intonation units reflect the processing time required to activate semi-active and inactive concepts. Activating semi-active concepts requires little cognitive effort; activating inactive concepts is more difficult. Because of this greater difficulty, Chafe proposes a universal cognitive constraint on spoken discourse (1987:31): '... a single intonation unit can convey no more than one previously inactive, or new concept.'

Chafe's three categories of informativity can be extended by analogy to vocabulary management in discourse. Verbal innovation is cognitively more difficult than verbal repetition, and function words are more accessible than content words. Hence, repeated words might be classified as 'given vocabulary' (active words), new function words as 'accessible vocabulary' (semi-active words), and new content words as 'new vocabulary' (inactive words). The celebrated cognitive difficulty of recovering le mot juste, which writers since Flaubert have often complained about, is normally confined to content words. Similarly, when we struggle to recall a word, it is almost always an infrequently-used content word (such as, say, *accismus*) rather than a function word (such as *hence*). A fourth category of words is suggested by *FW*: 'invented vocabulary' (neologisms). Neologisms are cognitively the most demanding of all; Joyce took thirteen years to write *FW*, and he expected his readers to take about that long to read it.

If this analogy between new vocabulary and new information in discourse is correct, then the incidence of new content words should be higher in new information than in given and accessible information. Such a correlation does exist, as is illustrated by 4, which is an excerpt from a conversational narrative quoted in Chafe 1987. Brackets enclose the phrases that Chafe identifies as new information, and upper case signifies new content words.

(4) ... it['S FUNNY] though, ... i do THINK that [MAKES a DIFFERENCE] .. but, i can RECALL ... uh- ... a BIG UNDERGRADUATE CLASS that i had, ... where .. everybody [LOVED] the INSTRUCTOR, ... a--nd .. he [was a ... REAL .. uh .. OLD WORLD ... SWISS]-- ... GUY, .. this [was uh .. a BIOLOGY] COURSE, ... a--nd he-- ... [LEFT all of the-- sort of uh-- ... real CONTACT with STUDENTS .. up to] .. his ASSISTANTS.

(... Mhm,)

... A--nd .. he [would COME into class], ... [a--t .. uh-- you KNOW THREE or F] .. [PRECISELY ONE MINUTE after the HOUR], or something like that, ... a--nd he-- .. [wou--ld .. IMMEDIATELY OPEN] his ... NOTES [up], ... [in the FRONT of

the] ROOM, .. and he [ST] and every ... every LECTURE, ... [after the FIRST], .. [STARTED the SAME WAY].

In 4, every phrase that Chafe identifies as new information contains at least one new content word. Furthermore, in new information the ratio between new content words and other words is 0.8 (28/35), whereas in given and accessible information, this ratio drops to 0.3 (11/38)—a positive correlation of .229 with a probability of just .015 that this correlation occurs by chance.

Chafe's universal constraint on spoken discourse—one new concept per intonation unit—implies that new concepts are likely to occur at roughly equal intervals in continuous discourse. Similarly, insofar as new concepts and new vocabulary are correlated, the incidence of new vocabulary tends to rise and fall at rhythmic intervals in concert with intonation groups. Intonation groups, in turn, combine to form higher-level constituents: clauses, sentences, paragraphs, episodes, and the like. It is reasonable to postulate that these higher-level constituents also have new information/new vocabulary peaks, and that these peaks, too, tend to occur at rhythmic intervals.

Unfortunately, type-token curves like those in §2 provide little information about such rhythmic variations in discourse. The ebb and flow of new vocabulary is very difficult to detect in Figs. 1–3; 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. The next section describes a new kind of curve, the VMP, which is designed to compensate for this deficiency in type-token curves.

**4.** VOCABULARY-MANAGEMENT PROFILES (VMPs). Type-token curves such as those in §2 are too smooth to signal changes in topics clearly; a more sensitive quantitative indicator is therefore 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 $\Delta y$ equals the number of new types, and $\Delta x$ 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 $\Delta x$ 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 $\Delta x$ 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, $\Delta y$, introduced over a moving interval, $\Delta x$; then I plotted the values for $\Delta y$ at the midpoint of the intervals $\Delta x$. 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, $\Delta x$ 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 Figure 4, which is the VMP for the first 2000 words of James Joyce's short story 'The Dead'.

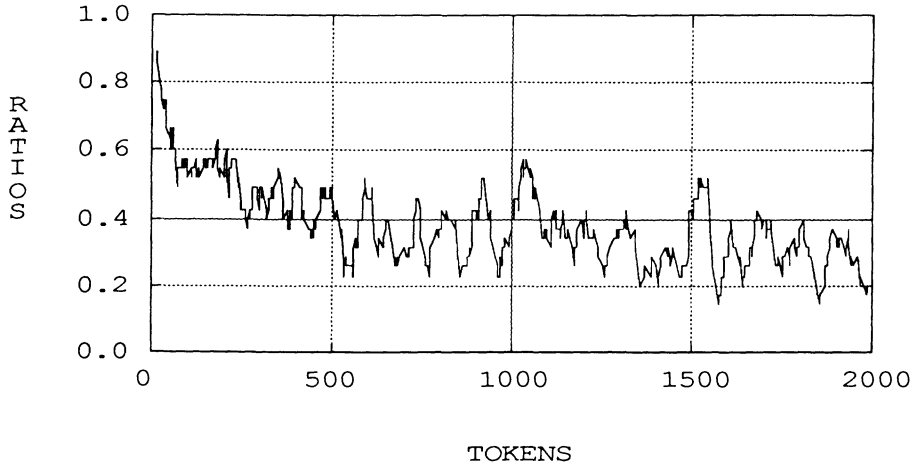


FIGURE 4. The Ratios $\Delta y/\Delta x$ for James Joyce's 'The Dead' ($\Delta x = 35$).

[2] More precisely, I plotted points only for intervals of text that end with a new vocabulary word. This reduces the number of points that need to be plotted by about two thirds while still signaling all turning points in the curve. The program is written in Turbo Pascal for the Macintosh SE. Points are plotted using the MYSTAT statistical package. Further information about the program may be obtained by writing the author.

There is a striking contrast between the jagged appearance of the VMP in Fig. 4 and the relatively smooth type-token curves for the same author in Fig. 3. The curve in Fig. 4 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 Fig. 4 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 Figure 5 illustrates the effect of increasing the interval $\Delta x$ from thirty-five to 101 words. The passage is the same as the one plotted in Fig. 4. The VMP in Fig. 5 loses considerable detail. Most notably, it fails to reveal the regular 100-word alternation between peaks and valleys that is so obvious in Fig. 4. This illustrates a general principle about VMPs: they cannot detect patterns shorter than $\Delta x$. However, Fig. 5 shows even more clearly than Fig. 4 that the peaks occurring shortly after 1000 and 1500 words are especially prominent ones. This illustrates another characteristic of VMPs: longer intervals for $\Delta x$ are more useful for detecting long-term patterns in discourse than shorter intervals are.

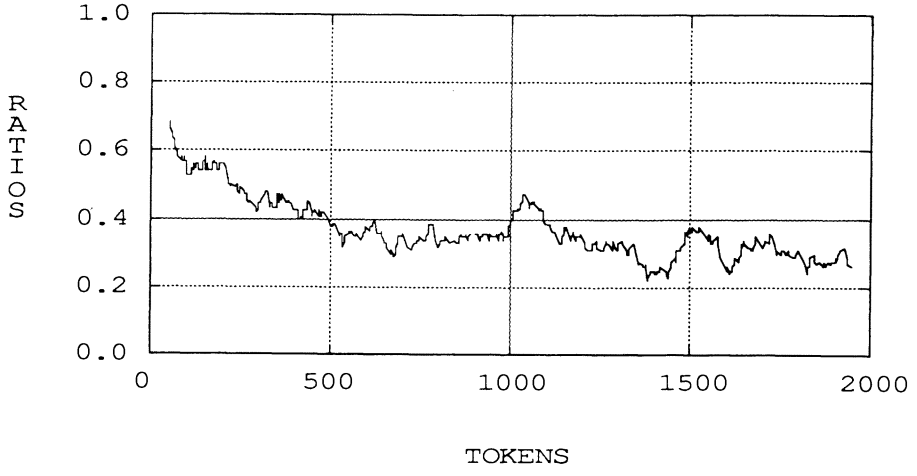


FIGURE 5. The ratios $\Delta y/\Delta x$ for James Joyce's 'The Dead' ($\Delta x = 101$).

The remainder of this article focuses on intermediate-term intervals, with $\Delta x = 35$. Once the interval for $\Delta x$ is fixed, we no longer need to compute the ratio $\Delta y/\Delta x$, 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 $\Delta y/\Delta x$, but $\Delta y$, 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. Con-

sequently, the vertical scale in Figure 6 differs from the one in Fig. 4, but otherwise the two curves are identical.

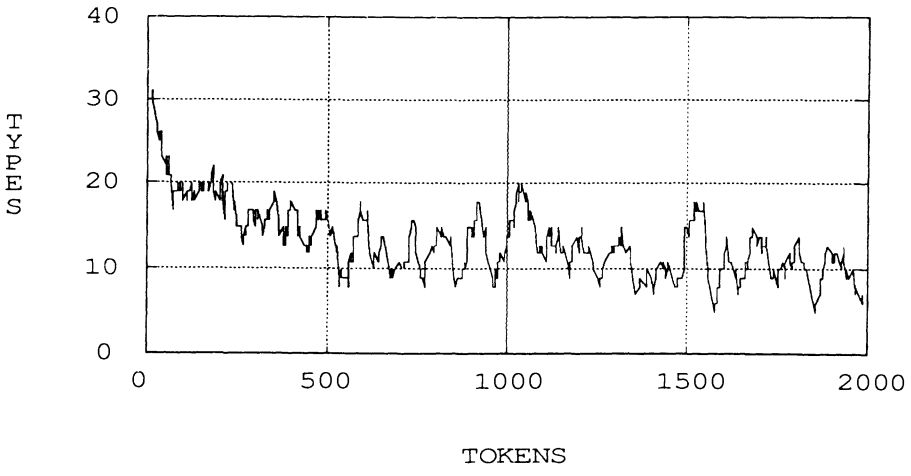


FIGURE 6. The VMP for 'The Dead' ($\Delta x = 35$).

The VMP in Fig. 6 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 new topics, whereas new semantic content words (nouns, main verbs, adjectives, and some adverbs) do denote new 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 Figure 7.

The two curves Figs. 6 and 7 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 viewpoint of information management, the chief advantage of the VMP in Fig. 7 over the one in Fig. 6 is that Fig. 7 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 Fig. 7 is slightly lower than Fig. 6 overall.

The next step is to conflate the inflected and derived forms of the semantic

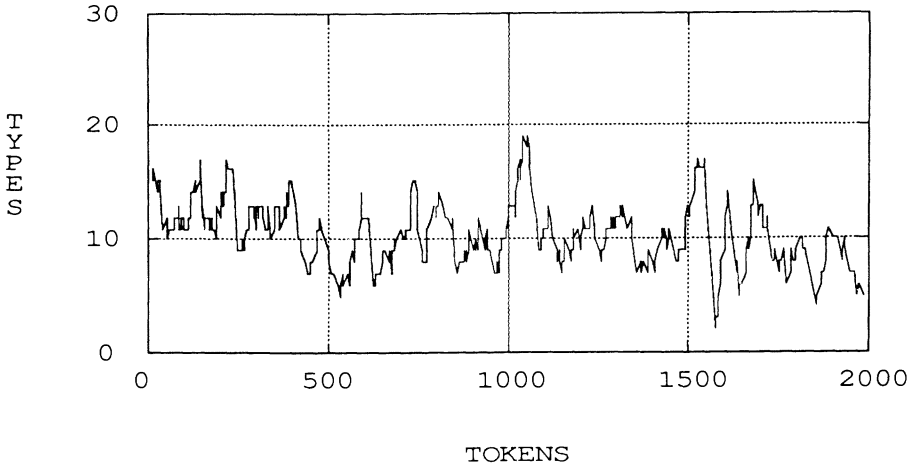


FIGURE 7. The VMP for 'The Dead' with function words replaced by *x*.

content words that remain in the skeletal version of 'The Dead' profiled in Fig. 7. Replacing all inflected words with their stems and deleting selected derivational affixes such as *-ly* reduces the total vocabulary in the topical skeleton only by about 87 additional types, from 572 to 485 words. The VMP for this reduced version of the passage is shown in Figure 8. The overall curve in Fig. 8 is lowered again, but otherwise its VMP is very similar to the one in Fig. 7. The reason for this is that the loss of 87 vocabulary words in Fig. 8 is distributed more or less evenly throughout the text. Consequently, the VMP in Fig. 8 and the one in Fig. 7 give nearly identical signals: 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 Fig. 8. This step reduces the total vocabulary in the

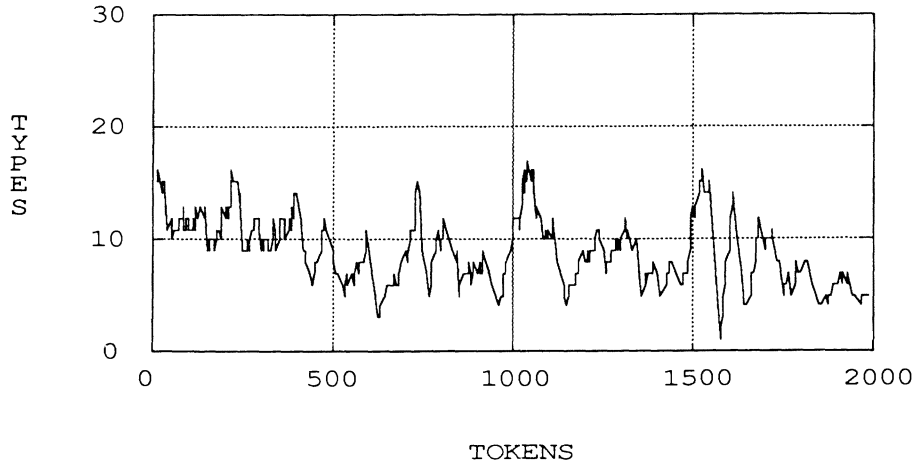


FIGURE 8. The VMP for 'The Dead' with function words replaced by *x* and affixes deleted.

passage only by about a dozen words, and 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'; its VMP is therefore almost identical with the one in Fig. 7. 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, although it typically lags slightly behind (or jumps slightly ahead) of major changes in the weather.

**4.1.** HIERARCHICAL CONSTITUENTS OF VMPS. The VMPs in Figs. 4–8 look much like stock-market curves, with upturns and downturns, with minor peaks and valleys superimposed upon major ones, and so on. Some preliminary definitions will be useful:

(5) VMP PEAK: any point or plateau immediately preceded and followed by lower points on the VMP.

(6) VMP VALLEY: any point or plateau immediately preceded and followed by higher points on the VMP.

(7) VMP CONSTITUENT: any peak bounded by two valleys.

Higher peaks are more salient than lower ones. Hence, three possibilities suggest themselves: VMP constituents can be coördinate, subordinate, or conjunctive. In 8 these different possibilities are represented graphically.

(8) (a) COÖRDINATE (CONJOINED) VMP CONSTITUENTS: /\/\
(b) PRESUBORDINATION: /\ (c) POSTSUBORDINATION: /\
(d) CONJUNCTIVE PEAK: /\/\

The relationships illustrated in 8 can be defined as follows:

(9) COÖRDINATE (CONJOINED) VMP CONSTITUENTS: Successive VMP constituents $C_i$ and $C_{i+1}$ are coördinate if they have the same height and comparable widths.

(10) SUBORDINATE (EMBEDDED) VMP CONSTITUENTS:
(a) PRESUBORDINATION: $C_i$ is subordinate to $C_{i+1}$ if the peak of $C_i$ is lower than the peak of $C_{i+1}$ and the first valley of $C_i$ is lower than the second valley of $C_i$.

[3] Such refinements in the VMP are likely to produce more significant results for discourses in agglutinative languages and highly inflected ones. However, a simpler procedure for such languages would be to insert spaces between affixes and root morphemes; then the computer program would treat affixes much like function words.

(b) POSTSUBORDINATION: $C_i$ is subordinate to $C_{i-1}$ if the peak of $C_{i-1}$ is higher than the peak of $C_i$ and the first valley of $C_i$ is higher than the second valley of $C_i$.

(11) CONJUNCTIVE (TRANSITIONAL) CONSTITUENTS: $C_i$ is transitional if (a) the first and second valleys of $C_i$ are the same height, (b) the peak of $C_i$ is lower than the peaks of $C_{i-1}$ and $C_{i+1}$, and (c) $C_{i-1}$ and $C_{i+1}$ are comparable in height and width.

The illustrations and definitions in 8–11 are intended as idealizations. Actual VMPs almost never include identical successive constituents such as those depicted in 8a. Consequently, it is more accurate to treat coördination and subordination in VMPs as gradient pairs rather than as categorical opposites. Successive constituents of VMPs can be 'nearly' coördinate, and some subordinate constituents are clearly 'more subordinate' than others.

The definitions in 8–11 also fail to correct for the downward drift of VMPs resulting from the finite size of speakers' vocabularies. Because of this downward drift, uptrends in the VMP are more significant than downtrends as signals of discourse boundaries, subordination, and the like. Nevertheless, as long as 8–11 are understood to define ideal prototypes rather than clearcut categories, they will serve as useful guides for discussing the correlation between VMPs and constituent structure in discourse.

**4.2.** THE VMP AND NARRATIVE CONSTITUENTS IN 'THE DEAD'. I return now to the curve plotted in Fig. 7 to examine how closely this version of the VMP correlates with the narrative structure of the first part of 'The Dead'. The story begins with Lily, the caretaker's daughter, greeting guests at the door of the Misses Morkans' house. The Morkans' nephew Gabriel Conroy arrives and converses briefly with Lily, then leaves her in the entry hall and goes upstairs to his aunts' party. Thereafter, the story concentrates upon Gabriel. This shift of focus is a significant turning point in 'The Dead': a conversation ends (between Gabriel and Lily), the setting changes (from entry hall to drawing room), new participants are introduced (the guests at the party), and the central character of the narrative becomes Gabriel rather than Lily. This combination of changes defines a clearcut boundary in the narrative; thus, the first part of the story can be identified as a narrative constituent, call it 'the entry-hall episode'. Hereafter, I will concentrate upon this part of the story (words 1–1189) and its correlate VMP, Fig. 7.

In the tradition of immediate-constituent analysis and tree diagramming, we might ask next where this entry-hall episode is subdivided into its component narrative constituents. That is, what is the highest-level constituent boundary within the episode? According to Longacre 1983, the two major subtopics of discourse analysis are monologue and dialogue, which require different analytic techniques. Polanyi 1985, too, distinguishes discourses in which dialogue is the 'encoding norm' from those in which reported speech is an evaluative device for signaling especially salient information. The first three paragraphs of 'The Dead' (words 1–554) are entirely monologic; they contain no reported speech. At the arrival of Gabriel and Gretta Conroy, however, there begins an extended

section of dialogic discourse interspersed with monologic discourse (words 555–1189). This section includes fourteen separate speeches divided evenly between Gabriel and Lily, who answers the door. Hence, the monologic-dialogic distinction implies that the highest-level discourse boundary in this episode occurs with Lily's first speech.

Prior to this major subdivision in the entry-hall episode, Joyce includes three monologic paragraphs. According to Polanyi (1985:19), Main Story Line Events are '... only those main clauses which fulfill all event criteria (active, affirmative, punctual, noniterative, completive) and which are in the simple past (or historical present) tense ...', excluding '... event clauses which do not form part of the main time line (flashed clauses for example)'. Despite the length of Joyce's opening paragraphs, they contain no clauses that qualify as Main Story Line Events under Polanyi's criteria. Until Lily's first line of dialogue (*—O, Mr Conroy, said Lily*), Joyce reserves the simple past tense for iterative and durative-descriptive clauses; otherwise, he uses past perfect or modal auxiliaries. Following Polanyi's criteria, we can classify the first 554 words of 'The Dead' as background or introductory material—one more justification for dividing this section from the conversation after it.

Hence, both Longacre's and Polanyi's criteria imply that the major subdivision within the entry-hall episode is between the three introductory monologic paragraphs and the dialogue that follows. This dividing point corresponds exactly with the lowest valley in the first 1500 tokens of the VMP in Fig. 7. This deepest valley (five types at 538 tokens) signals an upturn at the end of the thirty-five-word interval, with the 555th word *O* in 12:

(12) ... that was what brought them every two minutes to the banisters to ask Lily had Gabriel or Freddy come.
—O, Mr Conroy, said Lily to Gabriel when she opened the door for him, Miss Kate and Miss Julia thought you were never coming ...

In this case, then, the VMP gives exactly the right signal. The lowest valley in the curve pinpoints the exact boundary between the two major constituents of the episode. Signals from the VMP are rarely this accurate, however, and even in this case it is probably more correct to count the interjection *O* as a function word rather than as a content word. If so, then the upturn in the VMP begins with Gabriel's last name *Conroy*, which is still gratifyingly close to the boundary between the two major discourse constituents.

The next step is to compare the narrative structure of the introductory section (1–554 words) with the VMP in Fig. 7. Joyce divides this section into three paragraphs, the first of which is quoted in 13. (I have interpolated signals from the VMP within the text.)

(13) Lily, the caretaker's daughter, was literally run off her feet. Hardly had she brought one gentleman into the little pantry behind her office on the ground floor and helped him off with his overcoat than the wheezy hall-door bell clanged again and she had to scamper along the bare hallway to let in another guest. It was well for her she had not to attend to the ladies also. But Miss Kate and Miss [UPTURN after a valley of ten types at 58 tokens] Julia had thought of that and had converted the bathroom upstairs into a ladies dressing-room. Miss Kate and Miss Julia were there, gossiping and laughing and fussing, walking [DOWNTURN after a minor peak of thirteen types at 85 tokens] after

each other to the head of the stairs, peering down over the banisters and calling down to Lily to ask her who had come.

It was always a [UPTURN after a valley of eleven types at 115 tokens] great affair, the Misses Morkans' annual dance.

Hereafter, I will use an abbreviatory convention to represent turning points in the VMP: V10,58 means 'a valley of ten types at 58 tokens', and P13,85 is 'a peak of thirteen types at 85 tokens'.

An inspection of the VMP in Fig. 7 reveals that V11,115 near the end of 13 is followed by a major peak P17,149. Hence, this valley successfully signals the onset of a new paragraph. The first sentence of the second paragraph begins with four function words (*It was always a*), so the upturn in the VMP is delayed until the first content word in the predicate (*great*). This pattern is typical, since repeated vocabulary (like given information) is more common at the beginning of sentences, and new vocabulary (like new information) is more common at the end.[4] Consequently, upturns signaling the onset of new paragraphs typically begin in the predicates of sentences at or near the paragraph boundary. In 13 the major turning point after V11,115 is in the first sentence of the new paragraph, so its location is nearly ideal.

The VMP in Fig. 7 also implies that the first paragraph is subdivided into two parts, somewhere near the minor upturn following V10,58, which occurs with the word *Julia* in *But Miss Kate and Miss Julia had thought of that*. In this case, too, the correlation between discourse structure and the VMP is almost ideal, because this clause marks a change in primary participants, from Lily in the first half of the paragraph to Miss Kate and Miss Julia in the second half.

The early signals given by the VMP in Fig. 7 are misleading in only one respect: they seem to imply that the second half of the first paragraph is presubordinate to the second paragraph rather than postsubordinate to the first half-paragraph. That is, the first valley (V10,58) is slightly lower than the second one (V11,115), showing the intervening peak as subordinate to the following rather than the preceding VMP constituent. However, the two valleys are almost the same height, and the peak is a minor one (P13,85). Hence, it is reasonable to classify this constituent as transitional rather than subordinate. From the point of view of information management, this classification is plausible, because the constituent looks backward (to Lily and her duties) as well as forward (to the next paragraph, which begins with an account of Kate and Julia's party).

The second paragraph is much longer (words 128–445). Another clearcut valley (V7,449) signals the transition between the second and third paragraphs, although in this case the turning point is delayed until the second sentence of the new paragraph:

(14) But Lily seldom made a mistake in the orders so that she got on well with her three mistresses. They were fussy, that was all. But the only thing they would not stand was back answers.

[4] This observation holds true for discourses in English and probably most other languages. However, Grimes (1986:2) identifies two South American languages in which '... communicative dynamism may decrease rather than increase during a sentence'.

> Of course they had good reason to be fussy on such a night. And then it was long after ten [UPTURN after V7,449] o'clock and yet there was no sign of Gabriel and his wife.

This valley (V7,449) is followed by a clearcut peak (P12,478). Once again, the VMP successfully signals the onset of a new paragraph.

One of the prominent narrative devices that Longacre (1983:9) identifies is '... tail-head linkage (in which the last sentence of one paragraph cross-references to the first sentence of the following paragraph)'. Joyce uses this device in 14 with the sequence *they were fussy ... they had good reason to be fussy*. The same device seems to be even more common in expository writing. Whenever such tail-head repetitions occur, they inevitably lower the VMP. This explains why the VMP's turning point in 14 is delayed until the second sentence of the paragraph.

The third paragraph is 109 words long—very close to the typical length for intermediate-sized VMP constituents in 'The Dead'—and it contains just one peak. By contrast, the second paragraph is 328 words long, and its corresponding VMP implies considerable internal structure, with two major peaks (P17,149 and P17,223), a rather ambiguous series of plateaus (P13,309), and a final major peak (P15,405). The first VMP constituent (centered at the first peak) corresponds to Joyce's comments about 'the Misses Morkans' annual dance'. Joyce then shifts to a description of their house, which corresponds to the second VMP peak. The clearcut valley (V9,268) at the midpoint of the paragraph signals a shift to detailed comments about the Misses Morkans' niece, Mary Jane. The final peak corresponds with the re-emergence of Lily as the primary subject of the discourse. Hence, the VMP correlates closely with the narrative structure of the paragraph: the four peaks in the VMP correspond with four major topics.

The second part of this opening section of 'The Dead' (words 555–1189) is more complex, because it includes dialogue interspersed with narration and description. Although English orthography traditionally indents each change of speaker, Longacre (1983:47ff., following Klammer 1971) points out that successive speeches can be combined naturally to form larger dialogue 'paragraphs', which are composed of an Initiating Utterance (IU), optional Continuing Utterances (CU), a Resolving Utterance (RU), and an optional Evaluating Utterance. The conversation between Gabriel and Lily in this passage from 'The Dead' conforms to Longacre's taxonomy exactly. The characters' speeches fall into simple IU-RU pairs until the end of the passage, an IU-CU-CU-RU sequence in which Lily briefly resists accepting a Christmas tip.

Longacre points out (1983:47) that 'Along with the nucleus of a dialogue paragraph there may occur ... an outer periphery which consists of non-dialogue material ...'. Hence, a dialogue paragraph may include narration and description as well as direct quotation. Longacre's discussion appears in a chapter about repartee, so it is natural for him to treat dialogue as central and narration/description as peripheral. But from the viewpoint of information management these roles are often reversed. In this passage from 'The Dead', for example, the dialogue consists of short speeches (twenty words at the longest), whereas

the 'interpolated' description and narration are much longer. Consequently, the nondialogue passages tend to convey more information than Gabriel's and Lily's short speeches do. Furthermore, the speeches in this passage are mostly formulaic small talk; their communicative function is largely phatic rather than informative. As a result, the VMP constituents corresponding with words 555–1189 of 'The Dead' tend to coincide with descriptive and narrative topics rather than with Longacre's 'dialogue paragraphs', as the following analysis will demonstrate.

The dialogic section of the story (words 555–1189) reaches its first peak (P14,595) after the opening IU-RU exchange between Lily and Gabriel followed by a brief description of Gabriel:

(15) He stood on the mat, scraping the snow from his galoshes, [DOWNTURN after P14,595] while Lily led his wife to the foot of the stairs and called out ...

The description of Gabriel in 15 is interrupted by a brief greeting (including dialogue and embraces) between the Conroys and Gabriel's aunts. This greeting corresponds with a minor peak on the VMP (P9,658) bounded by two valleys (V6,636 and V7,674). After this interruption, Joyce resumes his description, which includes the next upturn in the VMP:

(16) He continued scraping his feet vigorously while the three [UPTURN after V7,674] women went upstairs, laughing, to the ladies' dressing-room. A light fringe of snow lay like a cape on the shoulders of his overcoat ...

The sequence *He stood on the mat, scraping* and *He continued scraping* is an example of head-head linkage, in which 'the first sentence of one paragraph cross-references to the first sentence of the following paragraph' (Longacre 1983:9). Given that 16 is an explicit continuation of 15, we can just as easily treat the intervening dialogue as an interpolation in the narrative, rather than treating the narration as an interpolation in the dialogue.

The description of Gabriel's snow-covered coat leads to Lily's IU in 17. I have included a complete 'dialogue paragraph' (in Longacre's sense) to illustrate that the VMP is out of phase with this constituent.

(17) —Is it snowing again, Mr Conroy? asked Lily.

She had preceded [DOWNTURN following P15,743] him into the pantry to help him off with his overcoat. Gabriel smiled at the three syllables she had given his surname and glanced at her. She was a [UPTURN following V8,773] slim, growing girl, pale in complexion and with hay-coloured hair. The gas in the pantry made her look still paler. Gabriel had known her when she was a child and used to sit on the lowest step nursing a rag doll. [DOWNTURN following P14,813]

—Yes, Lily, he answered, and I think we're in for a night of it.

Lily is helping Gabriel remove his snow-covered coat, so her question at the beginning of this passage is not a genuine request for information. She knows very well that it is snowing, and she is just making small talk about the weather. Consequently, the onset of new information in this 'dialogue paragraph' does not come at the beginning, but in the middle, with the shift to the detailed description of Lily, *She was a slim, growing girl* ....

The next major VMP constituent in Fig. 7 is a minor peak (P12,922) flanked

by clearcut valleys (V7,864 and V7,971). This portion of the curve fits the definition in 11 for 'transitional' VMP constituents. However, the most interesting thing about the passage is that it is the only one in this section of 'The Dead' to consist primarily of dialogue. It includes four conversational turns (two IU-RU exchanges), which are interrupted only by speech tags and by one fourteen-word narrative clause. This rapid-fire IU-RU exchange accounts for the distinctively prickly, porcupine look of the VMP. Elsewhere in the narrative, the VMP tends to rise rapidly to a clearcut peak, then fall quickly to a clearcut valley. In this case, however, the VMP is crowned by three prominent spikes with almost the same height (P11,896; P12,932; P11,939). These spikes (which are roughly coördinate) correspond to conversational turns. Hence, the VMP for this continuous stretch of Joyce's dialogue looks distinctively different from the VMPs for his descriptive-narrative passages. It is all the more surprising, then, that the length of this constituent (108 words) is nearly identical with that of the other intermediate-length VMP constituents in 'The Dead'.

After the valley V7,976 the VMP rises rapidly to the highest peak (P19,1052) in Fig. 7. This prominent VMP constituent coincides with a paragraph-long description of Gabriel, which begins in 18:

(18) Gabriel coloured as if he felt he had made a mistake and, without looking at her, kicked off his galoshes and [UPTURN after V7,976] flicked actively with his muffler at his patent-leather shoes.

He was a stout tallish young man. The high colour of his cheeks pushed upwards ...

Not only does the VMP come within one predicate of marking the onset of the new paragraph in 18, but the curve also depicts this paragraph as the most salient peak in the first 2000 words of the story. This is an ideal result, because the paragraph includes the first extended description of Gabriel Conroy, who emerges as the main character in the story. Indeed, the entire preceding conversation between Lily and Gabriel can be thought of as a transitional constituent, a passing of the baton as it were from Lily to Gabriel as the viewpoint character in 'The Dead'.

The last, minor VMP constituent in this part of the story corresponds to the final exchange between Gabriel and Lily, when Gabriel gives Lily a Christmas tip and goes upstairs to his aunts' party. Overall, then, the correlation between VMP constituents in Fig. 7 and narrative constituents in 'The Dead' is remarkably close. The VMP seems to be even more accurate as a measure of information flow in the discourse; roughly speaking, VMP peaks and valleys can be equated with information peaks and valleys as well.

**4.3.** VMPs FOR JOYCE'S 'EVELINE' AND *FINNEGANS WAKE*. Based upon my examination of VMPs for a wide variety of stories, essays, and letters, I have come to believe that many of the characteristics of the VMP described in §4.2 are typical of discourse in general. This is too broad a hypothesis to test in a single introductory article, but examples of VMPs from other narratives will add some plausibility to this claim. The VMP for Joyce's story 'Eveline' is shown in Figure 9. 'Eveline' is 1826 words long (considerably shorter than 'The Dead'), and Fig. 9 is the VMP for the entire story. The curve shows the by-

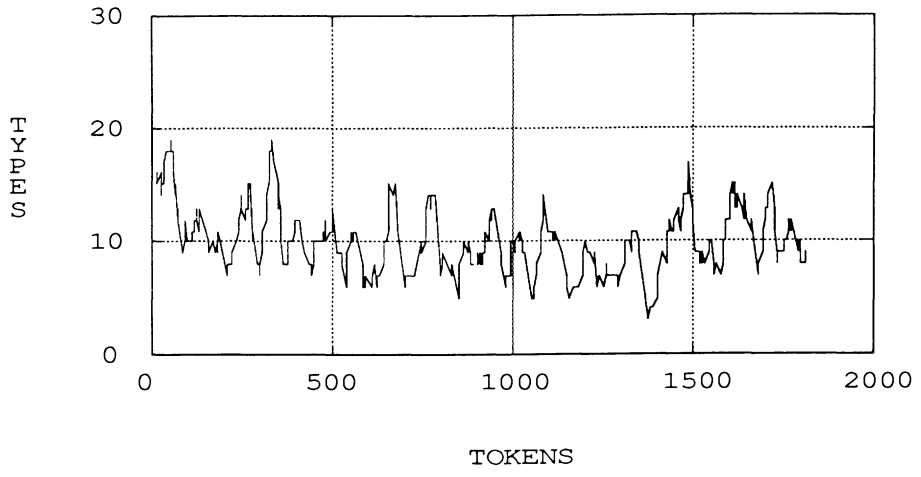


FIGURE 9. The VMP for James Joyce's 'Eveline' ($\Delta x = 35$).

now familiar pattern of regular alternations between new and repeated vocabulary; in fact, Fig. 9 looks even more regular than Fig. 7. This greater regularity results partly from the fact that 'Eveline' is almost pure narration and description. It includes no IU-RU conversations at all, just nine isolated quotations totaling only thirty-seven words.

Another striking feature of Fig. 9 is that once again the VMP peaks are separated by intervals of about a hundred words (intervals only slightly shorter than those for 'The Dead'). In addition, the peaks in the first part of the story tend to be narrower (and closer together) than those in the latter part of the story. This reflects a gradual change in the pace of narration. The first part of 'Eveline' includes more summarized narrative (iterative events, telescoped time, a variety of settings); it covers ground more quickly. The later part of the story shifts toward dramatized narrative (noniterative events; unity of time, place, and action); the pace is slower, and the detail is greater. Hence, the VMP in Fig. 9 can be interpreted as a visual approximation not only of the constituent structure, but also of the narrative pace, in 'Eveline'.

The deepest valley in Fig. 9 is at V3,1377. This valley is followed by a major peak (P17,1484) and another valley (V7,1562); this VMP constituent should therefore coincide with a major narrative constituent, and it does. The upturn preceding P17,1484 begins with the word *illness* in 19. Immediately afterward, Eveline has a particularly vivid memory that is recounted in dramatized narrative:

(19) She remembered the last night of her mother's [UPTURN after V3,1377] illness; she was again in the close dark room at the other side of the hall ... [There follows a dramatized description of the deathbed scene.]

She stood up in a sudden impulse of terror. Escape! [DOWNTURN after P17,1484] She must escape!

Not only does a major VMP constituent in Fig. 9 correspond to a major narrative constituent in 19, but the most salient VMP peak toward the end of the

story also coincides with the climactic exclamatory word *Escape*. In this case, then, the correlation between 'Eveline' and its VMP is ideal.

'The Dead' and 'Eveline' are conventional fictional narratives, relying upon traditional narrative techniques. By contrast, *Finnegans Wake* is Joyce's most unconventional narrative. Its organizing principle is not temporal sequence or logical structure, but psychological and verbal association. Given the notoriously idiosyncratic nature of psychological associations in general and of *FW* in particular, we might expect the VMPs for *FW* to be less regular than those for 'The Dead' and 'Eveline'. In fact, the VMP in Figure 10 for the first 2000 words of *FW* includes several well-defined peaks and valleys, some of which occur at fairly regular intervals; but Fig. 10 also includes arrhythmic stretches, most notably the section between 1135 and 1500 tokens. Reflecting the extraordinarily large vocabulary of *FW*, the overall curve in Fig. 10 is much higher

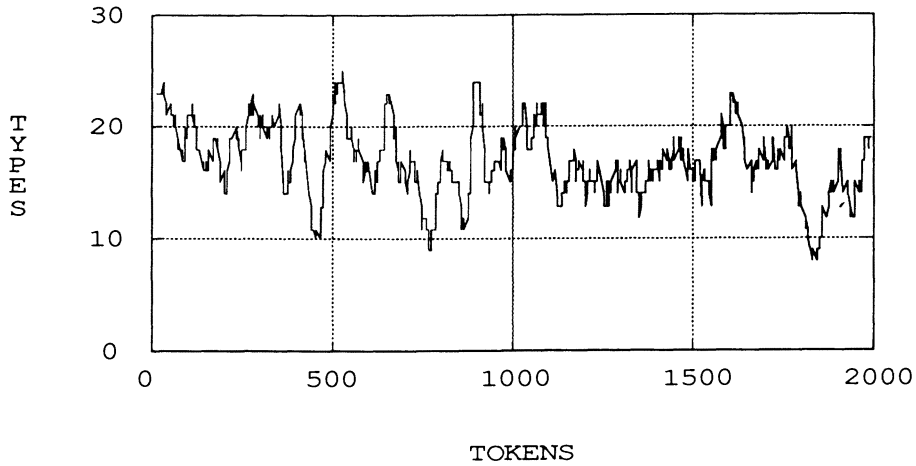


FIGURE 10. The VMP for the beginning of *Finnegans Wake* ($\Delta x = 35$)

than that for any other VMP I have examined. Although the VMP for *FW* is less orderly than the VMPs for 'The Dead' and 'Eveline', it does include clear-cut VMP constituents, which often correspond to distinct episodes in the narrative. One example is the deepest valley on the curve (V9,1839), which is followed by a well-defined peak (P18,1906). The upturn after this valley begins with the word *saloos* in 20.

(20) ... For her passkey supply to the janitrix, the mistress Kathe. Tip.
This is the way to the museyroom. Mind your hats goan in! Now yiz are in the Willingdone Museyroom. This is a Prooshious gunn. This is a ffrinch. Tip. This is the flag of the Prooshious, the Cap and Soracer. This is the bullet that byng the flag of the Prooshious. This is the ffrinch that fire on the Bull that bang the flag of the Prooshious. [UPTURN after V9,1839] Saloos the Crossgunn! Up with your pike and fork! Tip ...

The VMP correctly signals a major transition in 20—to the Willingdone Museyroom episode of *FW*—although in this case the upturn in the VMP does not occur until 60 words into the paragraph. This delay is caused by Joyce's

deliberately repetitive, 'this is the house that Jack built' style at the beginning of the paragraph. Such repetitions tend to depress the VMP, but the transition to a new episode eventually results in an increase in the rate of introduction of new vocabulary.

By comparison with clearcut constituents such as the one beginning at V9,1839, the VMP in Fig. 10 becomes decidedly arrhythmic between 1135 and 1500 words. This portion of the curve looks a bit like a longer version of the spiky, porcupine section of the VMP corresponding to the brief passage of sustained dialogue in 'The Dead'. The excerpt in 21 gives the flavor of this portion of *FW*.

(21) ... He's stiff but he's steady is Priam Olim! 'Twas he was the dacent gaylabouring youth. Sharpen his pillowscone, tap us his bier! E'erawhere in this whorl would ye hear sich a din again? With their deepbrow fundigs and the dusty fidelios. They laid him brawdawn alanglast bed. With a bockalips of finisky fore his feet. And a barrowload of guenesis hoer his head. Tee the tootal of the fluid hang the twoddle of the fuddled, O!

Hurrah, there is but young gleve for the owl globe wheels in view which is tautaulogically the same thing. Well, Him a being so on the flounder of his bulk like an overgrown babeling, let wee peep, see, at Hom, well, see peegee ought he ought, platterplate ...

Small wonder that the VMP becomes arrhythmic at this point in the narrative. Here and elsewhere in *FW*, transitions between thoughts and episodes are often hard to follow; partly for this reason, critics have found *FW* an extraordinarily difficult book to read. The VMP in Fig. 10 gives graphical evidence of the qualitative difference between the 'stream-of-consciousness' style of passages such as 21 and conventional narration in stories such as 'The Dead' and 'Eveline'.

**4.4.** VMPs IN ESSAYS. So far I have focused exclusively on narratives by Joyce, but VMP constituents correspond with structural constituents in essays as well as in stories. George Orwell's novel *1984* is a useful test case because it includes an essay, 'The Principles of Newspeak', as an appendix, allowing a comparison between narrative and expository passages within a single work. The VMP for the first 2000 words of the novel is plotted in Figure 11. The VMP in Fig. 11 looks surprisingly similar to the VMPs for 'Eveline' and 'The Dead'. Once again, the alternation between new and repeated vocabulary falls into a nearly regular pattern, with roughly hundred-word-long intervals between successive peaks and valleys.

The correspondence between VMP constituents in Fig. 11 and narrative constituents in *1984* is also similar to that found for Joyce. One example will have to suffice. A major peak at P20,1149 is flanked by major valleys at V4,1106 and V9,1216. This VMP constituent corresponds with a description of the Ministry of Love that begins in 22.

(22) ... Their names, in Newspeak; Minitrue, Minipax, Miniluv, and Miniplenty.

The Ministry of Love was the really frightening one. There were no windows in it at all. Winston had never been inside the Ministry of Love, nor within half a kilometer of it. It was a place impossible to enter except on [UPTURN after V4,1106] official business, and then only by penetrating through a maze of barbed-wire entanglements, steel doors, and hidden machine-gun nests ...

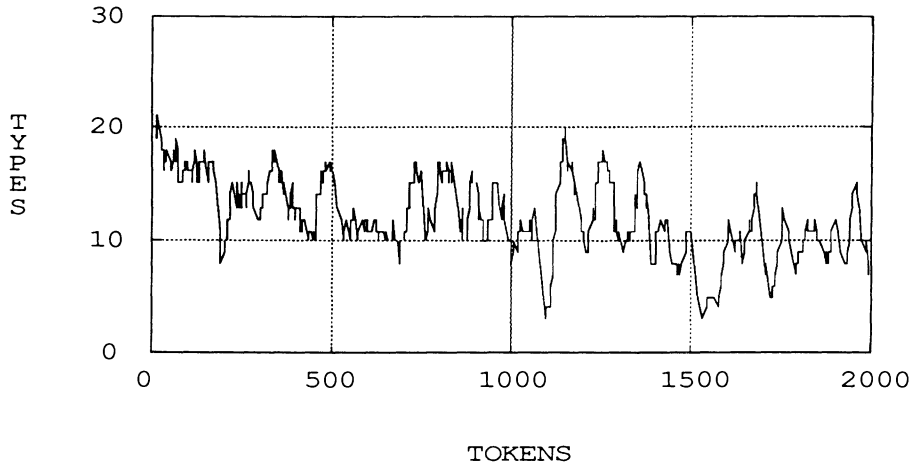


FIGURE 11. The VMP for the beginning of *1984* ($\Delta x = 35$).

This passage illustrates the tail-head structure identified by Longacre. The first paragraph ends by mentioning the Ministries of Truth, Peace, Love, and Plenty (and their Newspeak names), and the second paragraph begins by singling out the Ministry of Love for further comment. The repetition inherent in such tail-head constructions tends to lower the VMP. In addition, most of the content words in the second and third sentences are also repetitions (*windows, Winston, Ministry of Love, kilometer*); the only new content word is *half*. Consequently, the upturn in the VMP does not begin until the predicate of the fourth sentence, when new descriptive details about the ministry start to accumulate.

Tail-head and head-head constructions are even more common in Orwell's essay on Newspeak. Upturns in the VMP for this essay therefore tend to lag a sentence or two behind paragraph boundaries. The VMP for the first 2000 tokens of the essay is plotted in Figure 12.

The excerpt in 23 includes an especially clear example of a tail-head construction and its effect on the corresponding VMP. The beginning of the paragraph is signaled by an upturn after V10,540, which rises to a major peak at P17,589. Tail-head repetition delays this upturn until the predicate of the second sentence in the new paragraph.

(23) ... The grammatical peculiarities of the language can be dealt with in the section devoted to the A vocabulary, since the same rules held good for all three categories.

The A vocabulary. The A vocabulary [UPTURN after V10,540] consisted of the words needed for the business of everyday life—for such things as eating, drinking, working, putting on one's clothes, going up and down stairs, riding in vehicles, gardening, cooking, and the like. It was composed almost entirely of words that we already possess—words like hit, run, [DOWNTURN after P17,589] dog, tree, sugar, house, field—but in comparison with the present-day English vocabulary, their number was extremely small, while their meanings were far more rigidly defined. All ambiguities and shades of meaning had been purged out of them. So far as it could be achieved, a Newspeak word of this class was [UPTURN after V7,642] simply a staccato sound ex-

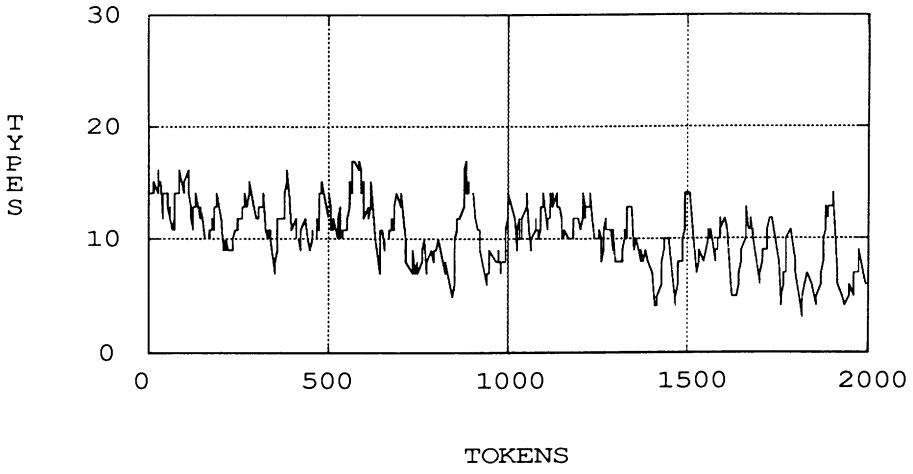


FIGURE 12. The VMP for the beginning of Orwell's essay 'Newspeak' ($\Delta x = 35$).

> pressing one clearly understood concept. It would have been quite impossible to use the A vocabulary for literary purposes or for political or philosophical discussion. It was intended only to express simple, purposive thoughts, usually involving concrete objects or physical actions.

Another characteristic trait of VMPs is revealed in the relation between Fig. 12 and the tell-tale formulas: *for such things as ...* and *words like ...* in 23. Expressions such as these (along with *for example* and *for instance*) frequently precede lists of specific examples; one can almost hear the prescription of composition teachers: 'Be specific'. Such lists typically include a high density of new content words, and in this example the first of the two lists in 23 causes a rapid rise in the VMP to a peak that occurs in the midst of the second list. Almost inevitably, the VMP begins to decline sharply after the completion of this second list. The density of function words and repeated content words increases: *English vocabulary, meanings, Newspeak, word*, and *class* have all been mentioned before. Hence, this portion of the paragraph tends to elaborate old topics rather than to introduce new ones. The decline in the VMP reverses itself with the predicate *simply a staccato sound ...*, which is followed by examples of possible and impossible uses for the A vocabulary—a new topic in the essay.

The pattern of Orwell's paragraph in 23 is repeated again and again in the fifty-some student essays I have examined so far. Tail-head and head-head constructions are the norm in these essays; consequently, an upturn in the VMP typically occurs two or three sentences into a paragraph. After this upturn, the paragraph rises to an information peak. (Long paragraphs nearly always include more than one peak.) *For example* is the students' favorite formula for introducing lists, and such phrases usually precede major VMP peaks.

The most prominent VMP constituent in Fig. 12 is the valley at V5,845 followed by the major peak P17,881 and the valley V6,941; somewhat ironically,

this VMP constituent is introduced by not just one, but two *for example*-style formulas:

(24) There was, for example, no such word as [UPTURN after V5,845] cut ...

**5.** FURTHER APPLICATIONS OF VMPs. The VMP is a new analytic tool; hence, it is difficult to predict what uses will be found for it. If I am correct in claiming that VMPs correlate closely with constituent structure and information flow in discourse, then VMPs should prove very useful in discourse analysis, especially since they allow the examination of large quantities of text by computer.

In this article I have focused upon narratives in English written by professional authors. Questions naturally arise about the VMPs for conversation, nonprofessional writing, different genres, different languages, and so on. My prediction is that morphemic rhythmicity will prove to be a universal tendency in discourse; that is, I suspect that the VMPs for all fluent speech and writing will tend to show regular alternations between new and repeated words and morphemes, reflecting two motivating polarities of discourse—innovation and coherence. Regardless of the truth or falsehood of this hypothesis, the VMP appears to be a promising tool for further research in discourse analysis. More broadly, it illustrates that quantitative methods can be used more widely in linguistics—to supplement and to elicit (rather than to replace) the judgments of native speakers.

## REFERENCES

CARROLL, JOHN B. 1968. Word frequency studies and the lognormal distribution. Proceedings of the conference on language and language behavior, ed. by Eric M. Zale, 213–35. New York: Appleton-Century-Crofts.

——; PETER DAVIS; and BARRY RICHMAN. 1971. Word frequency book. New York: American Heritage.

CHAFE, WALLACE L. 1974. Language and consciousness. Lg. 50.111–33.

——. 1987. Cognitive constraints on information flow. Coherence and grounding in discourse, ed. by Russell S. Tomlin, 21–51. Philadelphia: John Benjamins.

CHOMSKY, NOAM. 1958. Review of Langage des machines et langage humain, by Par Vitold Belevitch (Bruxelles: Office de Publicité, 1956). Lg. 34.99–105.

FRANCIS, W. NELSON, and HENRY KUČERA. 1982. Frequency analysis of English usage: Lexicon and grammar. Boston: Houghton Mifflin.

GRIMES, JOSEPH. 1975. The thread of discourse. The Hague: Mouton.

——. 1986. Sentence initial devices. Dallas, TX: The Summer Institute of Linguistics.

HALLE, MORRIS. 1958. Review of Language as choice and chance, by Gustav Herdan (Groningen: P. Noordhoff, 1956). Kratylos 3.20–28.

HALLIDAY, M. A. K., and RUQAIYA HASAN. 1976. Cohesion in English. London: Longman.

HERDAN, GUSTAV. 1960. Type-token mathematics: A textbook of mathematical linguistics. 's-Gravenhage: Mouton.

KATZ, JERROLD J., and JERRY FODOR. 1963. The structure of a semantic theory. Lg. 39.170–210.

KLAMMER, THOMAS P. 1971. The structure of dialogue paragraphs in written dramatic and narrative discourse. Ann Arbor: University of Michigan dissertation.

KUČERA, HENRY, and W. NELSON FRANCIS. 1967. Computational analysis of present-day English. Providence, RI: Brown University Press.

LABOV, WILLIAM. 1972. Language in the inner city. Philadelphia: University of Pennsylvania Press.
LONGACRE, ROBERT E. 1983. The grammar of discourse. New York: Plenum Press.
OGDEN, C. K. 1934. The system of Basic English. New York: Harcourt, Brace.
POLANYI, LIVIA. 1985. Telling the American story: A structural and cultural analysis of conversational storytelling. Norwood, NJ: Ablex.
PRINCE, ELLEN F. 1981. Toward a taxonomy of given-new information. Radical pragmatics, ed. by Peter Cole, 223–55. New York: Academic Press.
TANNEN, DEBORAH. 1984. Conversational style: Analyzing talk among friends. Norwood, NJ: Ablex.
YOUMANS, GILBERT. 1990. Measuring lexical style and competence: The type-token vocabulary curve. Style 24.584–99.

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

[Received 12 September 1990;
revision received 22 May 1991;
accepted 12 June 1991.]