# *performance: a transfer feature theory of processing in reading**

PATRICK J. FINN
*State University of New York at Buffalo*

HYPOTHESIZES THAT CLOZE EASINESS (the percent of subjects filling in the correct word in a cloze task) is a measure of *information* as defined by Shannon and Weaver (1949). A theory is developed that *information* is a function of 2 variables: a) the number of lexical markers associated with the word, and b) the number of lexical markers supplied for the word through what Weinrich (1966) calls transfer features. The theory states that frequent words are low information words because they have few lexical markers. Rare words are high information words because they have many lexical markers; however, the literature of automated indexing (*e.g.*, Carroll and Roeloffs, 1966) shows that rare words that are repeated in a text are closely associated with the topic of the text. It follows that rare words that are repeated in a text are low information words by virtue of extraordinary transfer feature support. The theory is shown to be consistent with a reanalysis of Bormuth (1966) data resulting from administering cloze tests to 675 elementary school subjects over 20 passages at 4 levels of difficulty. The theory is examined in light of 3 current models of reading. It is argued that the theory was consistent with some aspects of the models, that the theory may be more parsimonious than some aspects of the models, and that the theory suggests researchable questions which may refine, modify, or refute some aspects of the models.

# *La Fréquence des mots, la théorie de l'information et la performance sur les tests de Cloze: un indice lexical de la théorie du transfert des mots dans le procédé de la lecture*

ON ASSUME QUE LA FACILITÉ à compléter un test de Cloze est une information ainsi que l'ont définie Shannon et Weaver (1969) (indice que tire le lecteur d'un mot pour découvrir son sens). On a

*The author wishes to thank John R. Bormuth for the use of the data from Bormuth's 1966 study and S. David Farr for his advice in carrying out the statistical analysis for this paper.

développé une théorie qui postule que l'information est fonction de deux variantes: a) le nombre d'indices lexicaux (ce que les lettres et d'autres attributs indiquent au sujet des mots qu'ils forment) associés avec le mot, et b) le nombre d'indices lexicaux fournis pour le mot d'après ce que Weinrich (1966) nomme particularité de transfert. Cette théorie affirme que les mots qui ont une haute fréquence dans un texte ne donnent pas d'indications au lecteur car ces mots possèdent peu d'indices lexicaux. D'autre part les mots dont la fréquence est plus rare sont des mots à haute information parce qu'ils contiennent de nombreux indices lexicaux. Cependant d'après la recherche qui s'occupe du répertoire automatisé (études qui emploient un ordinateur afin de découvrir la fréquence moyenne d'un mot dans une langue) (e.g., Carroll et Roeloffs, 1966), les mots qui paraissent rarement dans un texte sont associés à son contenu. Il s'ensuit que plus la fréquence d'un mot est grande, moins le lecteur en assimile son sens. Cette théorie est compatible avec les données qui proviennent de l'administration de tests de Cloze à 675 élèves du niveau élémentaire sur 20 passages de lecture et à 4 niveaux de difficulté. Cette théorie enfin est examinée d'après 3 modèles de lecture. On remarque qu'elle est en accord avec certains aspects des modèles en question et qu'elle suggère la possibilité d'approfondir certaines questions qui à leur tour pourraient raffiner, modifier ou réfuter des aspects de ces modèles.

## *Repetición de palabras, teoría de información y efectividad del* cloze*: una teoría sobre indicadores lexicográficos y características expresivas de palabras en el proceso de lectura*

PLANTEA LA HIPOTESIS DE que el grado de facilidad de tareas tipo *cloze* (el porcentaje de sujetos que indican la palabra correcta en una tarea tipo *cloze*) es una medida de la *información*, según la definen Shannon y Weaver (1949). Desarrolla la teoría de que la *información* es una función de dos variables: a) el número de indicadores lixicográficos asociados a una palabra, y b) el número de indicadores lexicográficos provistos para una palabra a través de lo que Weinrich (1966) llama características expresivas de una palabra. La teoría expresa que las palabras de uso frecuente proveen poca información porque poseen escazos indicadores lexicográficos. Las palabras de uso poco frecuente proveen mucha informaciún pues tienen muchos indicadores lexicográficos; no obstante, la literatura que versa sobre la aplicaciún de computadores para contar las veces que usa una palabra (por ejemplo Carroll y Roeloffs, 1966) indica que las palabras de uso poco frecuente que aparecen repetidamente en un texto mantienen una estrecha relación con el tópico del texto. Consecuentemente, se deduce que las palabras infrecuentes repetidas

en un texto pasan a ser de escaza información en virtud del inusitado impulso de las características expresivas de la palabra. Se demuestra que la teoría es consistente con una modificación de los resultados de Boymuth (1966) al someter a 675 sujetos de la escuela primaria a pruebas tipo *cloze* en 20 trozos de grado 4 de dificultad. Se estudia la teoría en base a 3 modelos de lectura actuales. Concluye que la teoría es consistente con algunos aspectos de los modelos, y que además esta teoría sugiere temas de investigación que permitirían refinar, modificar o refutar algunos aspectos de los modelos.

An inductive study that reanalyzed data gathered by Bormuth (1966) reveals the relationship between the probability of subjects' success in supplying words verbatim in a cloze task and word frequency, and suggests a theory of processing in reading. The theory draws on information theory and a semantic theory.[1]

## *Theoretical background*

### Information theory and models of reading

In many of the models of reading proposed in the past decade, one finds a component dealing with the observation that the reader does not expend equal amounts of energy or attention or conscious effort in perception, recognition, or assigning meaning to every word in a written passage.

Goodman (1967) calls reading "a psycholinguistic guessing game." He proposes that the reader "picks up graphic cues guided by constraints set up by prior choices,"...[and] his language knowledge...."

Venezsky and Calfee (1970) and LaBerge and Samuels (1974) incorporate into their models the common observation that frequent words are processed by the reader with less effort than rarely encountered words. Venezsky and Calfee (1970) argue that:

> Extremely common words are retrieved from a highly organized store which can be searched rapidly on the basis of such features as intitial letters and length. Less common words are stored differently and are more accessible by sound than by any other form.

LaBerge and Samuels propose a model which focuses on the degree of automaticity of word processing. They propose that very common words are automatically coded into a visual word code which excites a meaning code. Unfamiliar words are automatically coded into spelling patterns and phonological codes, but it then requires the *attention* of the reader to refer to the episodic code (roughly speaking, the context of the word) and to excite the meaning code.

1. Writing the paper presented a problem: should one present the fairly complicated relationships in the data before explaining the theory which is developed from these relationships, or should one explain the theory first and then show how the theory is supported by the data? On the advice of several readers it was decided that one needs to understand the theory to make the description of the relationships in the data comprehensible. Therefore, the theory derived from the data is presented first and the data is presented second.

In this paper a theory will be developed to explain that common words are processed more rapidly and automatically than rarer words, and that even comparatively rare words require unequal amounts of time for recognition and processing in reading. The theory proposed here brings together Information Theory (Shannon and Weaver, 1949) and the semantic theory proposed by Weinrich (1966). The theory will be derived from observed results of subjects' success on cloze tasks.

First, it will be argued that the Cloze Easiness of a word (per cent of subjects supplying the exact word in a cloze task) is a measure of *information* supplied by the word in the passage. Second, several empirical observations will be stated: Cloze Easiness (or *information*) of a word is affected by the frequency of the word in English, by the level of difficulty of the passage, and by the number of times the word appears in the passage; furthermore, there is an interaction among these variables as they affect Cloze Easiness. Third, a theory of lexical markers and transfer features will be introduced and developed to explain the relationship among these 3 variables and their effect on Cloze Easiness or *information*.

## Information theory

Weaver (1949) points out that information theory is at once disappointing and bizarre—disappointing because it has nothing to do with meaning and bizarre because information and uncertainty turn out to be partners.

A typical problem in information theory is one where a person is receiving a message letter by letter (in Morse Code, for example) and because of interference on the channel, he is not always sure which letter code he hears. One's first thought might be that he has about 1 chance in 27 of guessing correctly, since there are only 26 letters (plus a space) to choose from. On second thought, one realizes that he would be wiser to choose *a* than *x*, for example. Since *a* appears more frequently in English spelling than *x*, the probability that the unknown letter is *a* is greater than the probability that the letter is *x*.

Consider now a person who has received 3 letters of a message over a noiseless channel but does not hear the fourth letter because of noise. The 3 letters received are *t-h-r*. Upon the slightest reflection, one knows that the doubtful letter must be a vowel. In this context, he is not choosing from 27 symbols, but from 5; and even among these 5, he would do well to guess *e*, since it is the most probable of the vowels in this context. In the terminology of information theory, written English is said

to be a *stochastic* process because it consists of a finite set of symbols whose probabilities of occurrence are not equal (*a* is more probable than *x*) and a *Markoff* process, because the probability of occurrence of a symbol is affected by the actual occurrence of symbols preceeding it. The probability of *e* following *t-h-r-* is much greater than the probability of *e* in the language in general.

Because of the *stochastic* and *Markoff* characteristics of written English, the probability of guessing the next letter in a message fluctuates between instances where the receiver is 99 per cent certain of what the next letter will be to instances where no letter is impossible and many letters are equally probable. In the former case, where the receiver is certain of the next letter, the actual physical occurrence of the letter code yields virtually no information—if there is no doubt, there can be no information. In the latter case, however, where the receiver is in great doubt as to which of many letters might be correct, the physical occurrence of the letter yields a great deal of information. That is what Weaver meant when he said that information and uncertainty are partners. The information in a symbol is a function of the amount of doubt the receiver has about its identity before the symbol appears.

When one considers words rather than letters in written English, the same *stochastic* and *Markoff* characteristics appear. At certain points in continuous discourse, the receiver is quite certain of what the next word will be before he actually sees it. At other points, he may be at a loss to guess what the next word will be. When the receiver knows what the next word is before the word comes over the channel, the word yields no information. When the receiver thinks the next word might be any one of a dozen, the actual appearance of the word yields a great deal of information.

## Information and meaning

It is important that the reader not confuse the ideas of *information* and *meaning*. To demonstrate the independence of these 2 ideas, the following paragraph was chosen from a passage used in this study:

> Cotton plants grow seeds covered with many fine threads. The threads are called fibers. The cotton fibers are pulled off the seeds by a machine called the cotton gin.

The words in this paragraph are listed in Table 1. Following each word is the per cent of subjects (Cloze Easiness) who supplied the word in a cloze

*Table 1* Cloze Easiness for each word in a sample paragraph

| Word # | Cloze Easiness | Word # | Cloze Easiness | Word # | Cloze Easiness |
|---|---|---|---|---|---|
| 1 Cotton | 10 | 11 threads | 45 | 21 the | 43 |
| 2 plants | 13 | 12 are | 59 | 22 seeds | 17 |
| 3 grow | 10 | 13 called | 26 | 23 by | 29 |
| 4 seeds | 10 | 14 fibers. | 42 | 24 a | 46 |
| 5 covered | 00 | 15 The | 50 | 25 machine | 42 |
| 6 with | 47 | 16 cotton | 22 | 26 called | 30 |
| 7 many | 7 | 17 fibers | 27 | 27 the | 45 |
| 8 fine | 13 | 18 are | 60 | 28 cotton | 60 |
| 9 threads. | 23 | 19 pulled | 10 | 29 gin. | 36 |
| 10 The | 46 | 20 off | 13 | | |

task. Notice that *cotton* is the first, sixteenth, and twenty-eighth word. The Cloze Easiness for *cotton* is 10, 22, and 60 per cent in its 3 appearances. There is no reason to believe that any one of these appearances of *cotton* has more meaning than any other. There is empirical evidence that *cotton* as word 1 has more *information* than *cotton* has as word 28: Few persons guessed what *cotton* was when word 1 was deleted in the cloze task; a majority of the subjects guessed what it was when word 28 was deleted. The same discussion applies to the multiple appearances of the words *seeds*, *threads*, and *fibers* in the paragraph.

In a *Markoff* system, one can specify the amount of *information* in a symbol (*e.g.*, a word) in terms of the number of symbols there are to choose from at a given point. Imagine a situation where a gambling casino has a radio connection to the stadium where the World Series is being played between the Sox and the Cubs. The following message is received over a very noisy channel—the words represented by *x*'s are garbled beyond recognition: "The xxxxx won the final game of the World Series by a score of 8 to xxxxx." The context tells one that the first garbled word can be only one of 2 words (*Sox* or *Cubs*), while the second garbled word could be one of 8 words (*0, 1, 2, 3, 4, 5, 6, 7*). Therefore, technically, the first garbled word represents a loss of less *information* than the second.

One way of looking at this is to imagine that the receiver must determine the garbled words by playing Twenty Questions with the sender. To get the first word, he need ask only 1 question: "Did the Sox win?" The answer to that question, whether *yes* or *no*, will identify the garbled word. For the second garbled word, the most efficient procedure

would be to reduce doubt by 1-half with each question. If the correct answer were 0, the questioning might go like this: "Is it greater than 3?" (*no*) "Is it greater than 1?" (*no*) "Is it 1?" (*no*)

In the technical vocabulary of information science, the first word has *1 bit* of information. The receiver has only 2 words to choose from. One yes-no question resolves the doubt. The second word has *3 bits* of information. The receiver has 8 symbols to choose from. One must ask a minimum of 3 yes-no questions to resolve the doubt.

The amount of *information* in the 2 garbled words can be specified precisely without referring to the *meaning* of either word. Yet one feels uncomfortable with the idea that information has nothing to do with meaning. It will be proposed later that meaning and information are related through the idea of lexical markers and transfer features, concepts which will be developed in the course of this discussion. For the present, the important idea is that *information* and *meaning* are not synonymous and that *information* can be determined without reference to *meaning*.

It will be argued in this paper that in context, extremely common words are not processed with relative rapidity and effortlessness because they are stored differently from less common words (*cf.*, Venezsky and Calfee, 1970). It will be argued that extremely common words are processed quickly and easily because they contain little information. The reader suspects with considerable certainty what such words are before he "comes to them." When words are very low in information, the reader does not have to find the lexical entry to see what is there, he brings the lexical entry to the printed word—thus, the rapidity in processing.

The theory developed in this paper proposes that *information* determines the degree of automaticity or ease of processing (*cf.*, LaBerge and Samuels, 1974). In Table 1, the word *cotton*, when it is the first word in the passage, would be processed with more attention or with less automaticity than the word *cotton* when it is the twenty-eighth word in the passage. This would follow from the fact that as the first word, Cloze Easiness is 10 (hard to guess—high in *information*) and as the twenty-eighth word, Cloze Easiness is 60 (comparatively easy to guess—low in *information*).

What is needed is a theory which will account for the fact that frequent words are usually low information words—but sometimes rare words are even lower in information than frequent words.

## Variables affecting Cloze Easiness or *information*

*Standard Frequency Index.* The measure of word frequency used in this study is the Standard Frequency Index reported in the *Word Frequency Book* (Carroll, Davies, and Richman, 1971). Five and one-half million words taken from American school books were tabulated, and a statistically estimated frequency for each word type in the sample is reported. This estimate reflects the number of times an average American student is likely to encounter each of the word types. The Standard Frequency Index ranges from 88.6 for *the* (meaning that the average student is likely to encounter *the* once in every 10 words of his school book reading) to 02.5 for *incarnation* (meaning that the average student is likely to encounter *incarnation* less often than once in every billion words of his school book reading).

The research reported under "An empirical test of the relationships" in this report demonstrates that there is a basic relationship between Standard Frequency Index and *information* because rare words were supplied on a cloze test by few subjects and frequent words were supplied by a large number of subjects. However, this relationship is affected by 2 other variables—Passage Difficulty and Text Frequency.

*Passage Difficulty.* In this study Passage Difficulty is determined by the Dale Chall Readability Formula (1948). Cloze Easiness for words having any Standard Frequency Index is likely to be higher in easy passages than in more difficult passages. For example, when very frequent words (such as *the* and *of*) are cloze items in a fourth grade passage, they are on the average, supplied correctly by more subjects than the same words are supplied correctly in a seventh grade passage.

*Text Frequency.* The number of times a particular word appears in a text is referred to in this paper as the Text Frequency of the word. If *commerce* or *democracy* appears only once in a passage of 250 words, it is very unlikely that subjects will supply it in a cloze task, regardless of passage difficulty. If it appears 5 times in a 250-word passage, the probability that subjects will supply it in a cloze task is enhanced considerably.

# *A lexical marker theory of the relationship between* information *and frequency*

## Lexical markers

Figure 1 shows 3 categories of lexical classification proposed

by Katz and Fodor (1964).[2] In this scheme, *grammatical markers* indicate where the word can appear in a sentence and what kind of grammatical information can appear with it. The fact that *bachelor* is a common noun indicates that it can be preceded by an article (*the bachelor*). The fact that *bachelor* is a count noun indicates that it can take a plural form (*bachelors*).

*Semantic markers* and *distinguishers* are related to the meaning of the word. Semantic markers express dichotomous categories which are systematic in language. Most nouns can be categorized in terms of whether or not they are *Concrete*. Most concrete nouns can be categorized in terms of whether or not they are *Animate*. Most animate nouns can be categorized in terms of whether or not they are *Human*. Most human nouns can be categorized in terms of whether or not they are *Male*. Thus, in Figure 1, *Human* and *Male* are classified as semantic markers. Distinguishers express categories which are relevant to only a small set of words. Whether or not one has ever married is not relevant to the meaning of most human nouns. Therefore, in Figure 1, *has never married* is classified as a *distinguisher*.

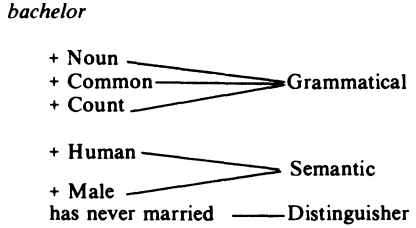


*Figure 1*
Lexical Entry for One Meaning of *Bachelor* Showing 3 Categories of Lexical Classification

For the present discussion, another kind of marker will be considered for words in sentences. This will be called a *case* marker after Fillmore (1968). In *The bachelor hit the ball, bachelor* has the feature +Agent. In *The bachelor knows Latin, bachelor* has the feature

2. The Katz/Fodor semantic scheme described here has come under considerable criticism. For example, Bolinger (1965) and Weinreich (1966), whose theory of transfer features is drawn later in this paper, point out that air-tight categorization of semantic features as syntactic and semantic markers and distinguishers is not really possible. However, the scheme is easy to understand and therefore useful for describing the concept of lexical markers. Alternative semantic theories, though theoretically more defensible, are unnecessarily abstract for the purposes of this paper.

+Dative. Case markers express relationships between verbs and nouns in a sentence.

## Transfer features

Words in discourse convey not only their own lexical markers; through what have been called "side effects" (Friedman and Bredt, 1968) or "transfer features" (Weinrich, 1966), the inclusion of a particular word in a discourse can dictate some lexical features for other words in the discourse.

Suppose Sentence 1 were presented to 8 persons but that each person received a version of the sentence with a different word deleted. How would the task of each person filling in the blank correctly be explained in terms of lexical markers and transfer features?

1] The man drove from Chicago in a car.

*Function words.* In the context of the example Sentence 1, *man* and *car* have the grammatical features +Noun, +Common, and +Singular. These features dictate that somewhere before the nouns there must appear a word with the features +Determiner and +Singular. Knowledge of the transfer feature +Determiner for the missing words in Sentences 2 and 3 reduces the possible choices to several words. Knowledge of the additional feature +Singular reduces the possible choices to only a few.

2] _____ man drove from Chicago in a car.

3] The man drove from Chicago in _____ car.

Features of the verb *drove* and the nouns *Chicago* and *car* indicate that the missing words in Sentences 4 and 5 have the feature +Preposition and that both prepositions are likely to have features signifying spatial relationships. Knowledge of each feature reduces the number of choices possible for the missing words.

4] The man drove _____ Chicago in car.

5] The man drove from Chicago _____ a car.

It is a fact that on cloze tasks, subjects nearly always fill the blank with the same part of speech as the deleted word even if the word is not correct. Fillenbaum *et al.* (1963), for example, found that when subjects were asked to fill in fifth word deletions in a transcript of spoken English, words of the same form class (Noun, Verb, Pronoun, etc.) were supplied 78 per cent of the time for the deleted words. It is also a fact that

function words have few lexical markers beyond grammatical markers. That is why they are called *function* words as distinguished from *content* words. Function words are very low information words. Fillenbaum *et al.* report that in every fifth word deletion cloze tasks, function words were supplied verbatim 65 per cent of the time, while content words were supplied only 32 per cent of the time.

*Content words.* The word *drove* has certain transfer features regarding the kinds of nouns that relate to it. The subject will most probably be +Agentive and must be +Animate and probably will be +Human and +Adult. The Locative (perhaps a better word here would be the Source, following Frederiksen, 1975) must be +Concrete. The Instrumental is probably a vehicle. So if *man* were the deleted word, as in Sentence 6, the subject knows it is a noun, it is animate, and most probably human and adult. There is really only 1 marker missing (+/- Male). If *car* were the deleted word, as in Sentence 7, the subject knows that it is a noun and probably a vehicle. *Car* has very little meaning beyond that for the subject to guess. *Chicago*, on the other hand, is simply a high information word. The only transfer features for the missing word in Sentence 8 are +Noun and +Source. Since there are many different words having these features, there are many different words to choose from.

6] The _____ drove from Chicago in a car.

7] The man drove from Chicago in a _____ .

8] The man drove from _____ in a car.

But suppose the sentence used in this example were Sentence 9 rather than 1.

9] The *baritone* drove from Chicago in a car.

We need not speculate on the particular markers on *baritone*. What is important is that it has all the markers as *man* plus some others. Subjects would not have nearly the probability of guessing *baritone* in Sentence 6 because its features in addition to the features of *man* are not supplied by transfer features.

The form of the argument would be the same if the sentence were Sentence 10. *Ambulance* has all the features of car, plus some others. Subjects would not be nearly as apt to guess *ambulance* as *car* in Sentence 7 because those additional features of *ambulance* are not supplied by transfer features.

10] The man drove from Chicago in an *ambulance*.

## Word frequency and lexical markers

It is proposed here that rarer words have more lexical markers than more frequent words. Our intuition is consistent with the hypothesis. Think of a common word and then think of a word that has all its features plus some others, like *ask* and *demand*; *walk* and *saunter*; *dancer* and *ballerina*; *hot* and *sultry*; *say* and *argue*. In every case the word with additional features is rarer than its partner.

## *Information*, lexical markers, and meaning

The theory proposed is that *information* in a word is a function of the number of features *not* supplied by transfer features in discourse. It is reasonable to suppose that the more features a word has the more likely that some features will not be supplied through transfer features. Therefore, in general, more frequent words will be low information words, while rare words will be high information words. The hypothesis is also consistent with our intuition that there is a connection between information and meaning. If "amount of meaning" is defined as the number of features associated with a word and if the number of features associated with a word is predictive of information, then words with more meaning would tend to be higher in information.

But there are 2 variables that can affect the number of features on a word not supplied by transfer features in discourse. One is the number of features on the word in question; the second is the number of transfer features supplied by the discourse for the word in question.

## Transfer features and Text Frequency

In a 250-word passage, the words *with, by,* and *is* were supplied by around 40 per cent of the subjects—making them low information words. But in the same passage, the word *intestine* appeared 5 times; and 4 out of the 5 times, it was supplied by more than 40 per cent of the subjects. Obviously, *intestine* has many more lexical markers than *with, by,* and *is*; but in this passage the word *intestine* was lower in *information* than these high frequency function words.

When comparatively rare words appear 5 times in a passage of 250 words, it is probably because they are intimately related to the topic of the discourse. The validity of this assertion has been tested by information scientists who discovered that a computer programed to identify rare words that appear repeatedly in a book will produce essentially the same list of words as a person who is asked to list the words in the book that should appear in the book index (Carroll and Roeloffs, 1969).

It follows that words intimately related to the topic (words with a low Standard Frequency Index but a high Text Frequency) will receive a good deal of transfer feature support from other words in the discourse. The passage referred to is about the process of digestion. When *intestine* is the deleted word, transfer features may include such highly specified distinguisher features as that the word refers to an organ that holds food during digestion and absorbs the digested foods into the blood stream. Because so many distinguisher features are supplied, the subject has very few possible words to choose from, and the word is a low information word.

In contrast, imagine a passage of the same length which is about a shooting, not about digestion, and the only occurrence of *intestine* is deleted from the sentence: *The bullet pierced the victim's* _____ . Here the same word, because it is not intimately related to the topic of the discourse, receives little transfer feature support from the discourse and is a high information word, as one would expect from the general principle that rare words have many lexical features and are therefore high information words.

## Transfer features and Passage Difficulty

Another factor that may influence the number of transfer features a given word receives from other words in a text is Passage Difficulty. An easy passage may be described as one where the reader is familiar with the vocabulary and the topic, and where the reader is fully capable of dealing with the syntax. Presumably, a difficult passage is one where the reader finds the vocabulary or topic unfamiliar or finds the syntax too complex for his capabilities. Each of these factors would influence the transfer feature phenomenon for a reader. If a word is totally unfamiliar to a reader, it cannot supply transfer features for words around it. For example, if *drive* in Sentence 1 is not a familiar word, the reader would not know that the Agentive is probably human and adult. If a topic is unfamiliar to a reader, he is deprived of transfer features that would be known to him if he knew the topic. For example, if a reader knew nothing about the art of fencing, he would not know features of the Instrumental typically associated with the verb *thrust* in a passage about fencing.

Finally, the whole phenomenon of transfer features is dependent on the reader's correct perception of syntax. For example, a reader cannot employ what he knows to be probable features of nouns associated with a verb if there is a breakdown in his perception of the grammatical relationships between the verb and nouns in a sentence.

The lexical marker—transfer feature theory is consistent with the fact that words having the same Standard Frequency Index have higher Cloze Easiness in fourth grade passages than in seventh grade passages. Words having the same Standard Frequency Index would, according to the theory, have the same number of lexical markers, but they would receive more transfer feature support in easier passages than in more difficult passages.

*Summary.* The theory proposed here is that *information* in a word is a function of the number of lexical markers associated with the words that are not supplied through transfer features. The number of lexical markers associated with a word that are not supplied through transfer features is, in turn, a function of 2 variables: 1) the number of lexical features associated with the word and 2) the number of lexical features supplied through transfer features.

It is proposed here that Standard Frequency Index correlates highly with the number of lexical features associated with a word and is, therefore, a good predictor of Cloze Easiness or *information* in a word in context. It is further proposed that when comparatively rare words are repeated in a text, they are probably intimately related to the topic of discourse and would therefore have many lexical markers supplied through transfer features. Therefore, Text Frequency is a good predictor of Cloze Easiness or *information* for rare words in context. Finally, it is proposed that the transfer feature phenomenon operates optimally when the reader is familiar with the vocabulary and the topic and when the syntax is easily perceived. As 1, 2, or 3 of these characteristics of the passage increase in difficulty, the transfer feature phenomenon will begin to operate at less than optimal efficiency. Therefore, knowledge of Passage Difficulty will enhance one's accuracy in predicting Cloze Easiness after the Standard Frequency Index and Text Frequency of a word are known.

## *An empirical test of the relationships*

The data available from an extensive 1966 study was reanalyzed to demonstrate whether in a cloze task:

1] Words having a high Standard Frequency Index are more likely to be supplied by subjects than words having a low Standard Frequency Index.

2] Any word, regardless of Standard Frequency Index, is more likely to be supplied by subjects when it appears in a passage

having a low Passage Difficulty than when it appears in a passage having a high Passage Difficulty.

3] Words having low Standard Frequency Index are more likely to be supplied by subjects if they have a high Text Frequency than if they have a low Text Frequency.

4] Standard Frequency Index, Passage Difficulty, and Text Frequency interact in facilitating subject's ability to supply words.

## The original data

Bormuth (1966) identified 20 passages in 5 subject areas (literature, history, geography, biology, and physics) at 4 levels of difficulty (grade levels 4, 5, 6, and 7) using the Dale-Chall Readability Formula. Each passage was about 250 words long. He administered 5 alternate forms of fifth word deletion cloze tests (so that each word in the passage was deleted on one of the forms) to 675 subjects. Each form of each passage was administered to 135 subjects, grades 4 through 8 in Wasco Union Elementary School Districts in California. Groups assigned to each form were matched for reading ability on the *California Reading Test.*

## The reanalysis variables

For the present study, the data on the 5,185 words from the Bormuth study were reanalyzed for each of the following variables:

1] Cloze Easiness (CE)—the per cent of correct cloze responses for the word from Bormuth's 1966 study.

2] Standard Frequency Index (SFI)—the estimated frequency of the word-type in an infinitely large sample of American School Books calculated from actual frequencies in a 4,500,000 word count (Carroll, Davies, and Richman, 1971).

3] Test Frequency (TF)—the number of times the word-type appears in the given passage.

4] Passage Difficulty (PD)—the difficulty of the passage (as measured by the Dale-Chall Readability Formula) in which the word appeared.

5] Token Position (TP)—the position of the token in the passage. Token Positions 1 and 250 refer to the 1st and the 250th words in a passage.

6] Type Number (TN)—the specification for which appearance a

particular token is for its type (for words that appear more than once in a passage).

The following sentence was taken from a fifth grade geography passage in the study: *More than a million other people live in the suburbs of Washington.* Table 2 shows the data collected for each word in this sentence. For example, the word *Washington* was supplied by 30 per cent of the subjects in the study (CE=30); according to the Carroll, Davies, and Richman list, it appears about once in 10,000 (SFI=59); *Washington* appears 9 times in the text from which the sentence was taken (TF=9); this token of *Washington* is the 193rd word (TP=193) and is the seventh time *Washington* appears (TN=7); the passage in which *Washington* appears is of fifth grade readability according to the Dale-Chall Formula (PD=5).

*Table 2* Example sentence showing the values for the 6 variables collected for each word in the study

| Word Token | Cloze Easiness | Standard Frequency Index | Text Frequency | Token Position | Type Number | Passage Difficulty |
|---|---|---|---|---|---|---|
| More | 45 | 72 | 1 | 182 | 1 | 5 |
| than | 66 | 71 | 1 | 183 | 1 | 5 |
| a | 32 | 83 | 5 | 184 | 4 | 5 |
| million | 26 | 60 | 1 | 185 | 1 | 5 |
| other | 13 | 73 | 4 | 186 | 2 | 5 |
| people | 55 | 71 | 2 | 187 | 2 | 5 |
| live | 38 | 66 | 1 | 188 | 1 | 5 |
| in | 68 | 82 | 14 | 189 | 10 | 5 |
| the | 57 | 88 | 25 | 190 | 18 | 5 |
| suburbs | 00 | 44 | 1 | 191 | 1 | 5 |
| of | 46 | 84 | 15 | 192 | 9 | 5 |
| Washington | 30 | 59 | 9 | 193 | 7 | 5 |

Finn (1975) demonstrated a relationship between Standard Frequency, Text Frequency, Passage Difficulty, and Cloze Easiness. Because of the *Markoff* characteristics of language, 2 other variables were added to the present investigation—Token Position and Type Number. These variables were included to examine the possibility that words encountered later in a passage (reflected by Token Position) are easier to guess, and that words repeated in the passage are easier to guess on their second or third, etc. appearance (reflected by Type Number).

## *The Reanalysis*

### Description of the relationships among the variables

Means and standard deviations for all variables are reported in Table 3. Correlations among the variables are reported in Table 4. All

of the variables correlate with Cloze Easiness (p. < .0001). However, 2 things must be taken into consideration: With 5,185 cases, very low and uninteresting correlations may reach a high level of statistical significance. Secondly, many of these variables would be expected to correlate highly with each other. For example, only words with a high Text Frequency can have high Type Numbers. Text Frequency and Type Number correlate .84 with each other. The question arises whether knowing Type Number for a word adds to one's ability of predicting Cloze Easiness if the Text Frequency is known.

*Table 3* Description of the data for 6 variables for all 5,185 words

| | MEAN | STANDARD DEVIATION |
|---|---|---|
| 1. Cloze Easiness | 29.5 | 21.6 |
| 2. Standard Frequency Index | 68.3 | 14.1 |
| 3. Text Frequency | 5.5 | 7.3 |
| 4. Passage Difficulty* | 2.5 | 1.1 |
| 5. Token Position | 146.6 | 75.4 |
| 6. Type Number | 3.4 | 4.6 |

*Passage Difficulty was converted from Grades 4, 5, 6 and 7 to levels of difficulty 1, 2, 3 and 4.

*Table 4* Correlations among 6 variables for all words (N=5,185)

| | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| 1. Cloze Easiness | .99 | | | | | |
| 2. Standard Frequency Index | .55 | .99 | | | | |
| 3. Text Frequency | .48 | .54 | .99 | | | |
| 4. Passsage Difficulty | -.38 | -.07 | .05 | .99 | | |
| 5. Token Position | -.10 | -.03 | .00 | .08 | .99 | |
| 6. Type Number | .39 | .45 | .84 | .05 | .25 | .99 |

Multiple regression analysis was run with Cloze Easiness as the dependent variable and with the independent variables selected in order of the amount of independent variance they accounted for in Cloze Easiness. The summary table is reported in Table 5. Standard Frequency Index is the best predictor of Cloze Easiness ($R=.551$; $R^2=.304$). Passage Difficulty and Text Frequency account for an additional .116 and .064 of the variance in Cloze Easiness. The multiple correlation between the 3 independent variables—Standard Frequency Index, Passage Difficulty, and Text Frequency—and the dependent variable, Cloze Easiness, is .696. The variance accounted for in Cloze Easiness is .484.

Although Token Position accounts for statistically significant variance, it accounts for such a small change in $R^2$ as to be of no further interest in this discussion. Type Number did not account for enough independent variance to be taken into the multiple regression formula. The comment may be in order, however, that the nature of the cloze tests

*Table 5* Multiple regression summary table for all words—Cloze Easiness is the dependent variable

| Step Number | Variable | R | $R^2$ | Increment in $R^2$ |
|---|---|---|---|---|
| 1 | Standard Frequency Index | .551 | .304 | .304* |
| 2 | Passage Difficulty | .648 | .420 | .116* |
| 3 | Text Frequency | .696 | .484 | .064* |
| 4 | Token Position | .698 | .487 | .003* |

*$p < .0001$

where subjects can go back and change an answer or fill in an answer after reading beyond the blank may explain why Token Position and Type Number have such a small effect on Cloze Easiness. Neville and Pugh (1976) found that better middle grade readers were more successful on regular cloze tests than on modified cloze tests, where the subject was not permitted to read beyond a blank before he responded and was not permitted to change a response after he had read the text beyond a blank to which he had responded.

## Non-linearity of the relationships

*Cloze Easiness and Standard Frequency Index.* The relationship between Cloze Easiness and Standard Frequency Index was fitted to first through fifth degree polynomial curves to determine whether these relationships were linear. The correlation between Cloze Easiness and Standard Frequency Index increases from .55 to .59 for the second degree polynomial. For third degree and beyond, the increase in $R^2$ is 0. The gain from .55 to .59 is modest. The graph of the second degree polynomial (Figure 2) indicates that the relationship between Cloze Easiness and Standard Frequency Index is nearly linear for words having a Standard Frequency Index of 50.0 or greater. For words having a Standard Frequency of less than 50, the curve flattens out.

*Cloze Easiness and Text Frequency.* The relationship between Cloze Easiness and Text Frequency was also fitted to polynomials. The correlation between Cloze Easiness and Text Frequency increases from .48 to .58 for the third degree polynomial. For the fourth and fifth degree, the increase in $R^2$ is near 0. The graph of the third degree polynomial (Figure 3) indicates that the relationship between Text Frequency and Cloze Easiness is nearly linear for words appearing 10 times or less in a passage. The curve flattens out beyond Text Frequency = 10, indicating that as Text Frequency increases beyond 10, Cloze Easiness remains constant.

*Text Frequency and Standard Frequency Index.* Text Frequency and Standard Frequency Index are highly correlated. This is

an extraordinarily reliable finding dating back to Zipf (1935). The simple fact is that the same words that have a high Standard Frequency Index will have a high Text Frequency (*e.g.*, *the* and *of*), and if words with low Standard Frequency Index appear in a text, they will have a comparatively low Text Frequency. For the passages used in this study, the linear correlation between Standard Frequency Index and Text Frequency is .54 (Table 4). However, fitting the data to polynomial equations reveals that this relationship is not linear and that the correlation for the fifth degree polynomial is .85. The graph of the fifth degree polynomial (Figure 4) shows that Text Frequency remains fairly constant (varies between 2 and 3) as Standard Frequency Index varies between 40 and 75. As Standard Frequency Index increases above 75, there is a very dramatic increase in Text Frequency. (These are words like

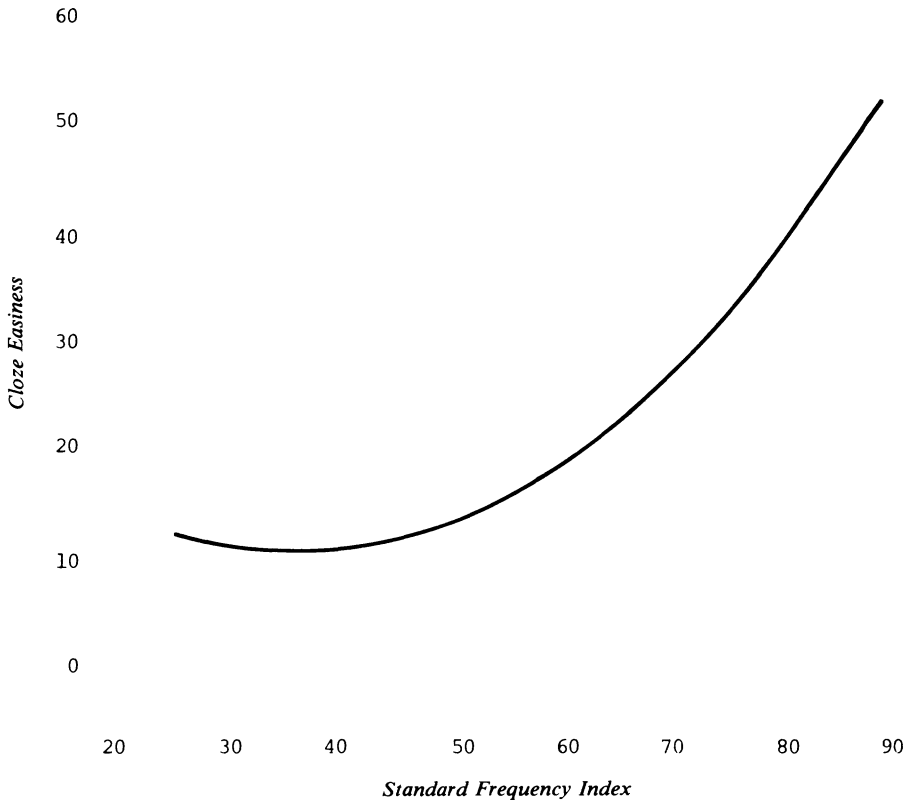


*Figure 2*
Second Degree Polynomial Relating Standard Frequency Index to Cloze Easiness for All Words (N = 5185, $r$ = .59)

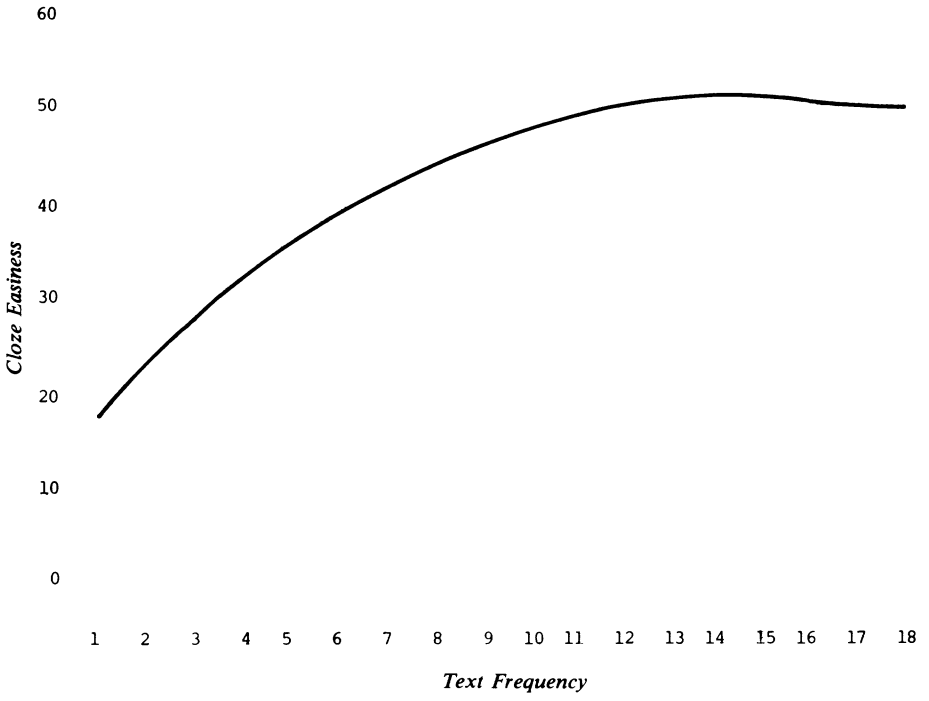


*Figure 3*
Third Degree Polynomial Relating Cloze Easiness to Text Frequency for All Words (N = 5185, $r$ = .57)

*the* and *of*.) As Standard Frequency Index decreases below 40, there is a decided increase in Text Frequency. (These are words like *commerce* and *democracy*.)

*Lexical marker—transfer feature interpretation of the relationship between Cloze Easiness and Standard Frequency.* The relationship between Cloze Easiness and Standard Frequency Index shown in Figure 2 is consistent with the following hypothesis: Words having very high Standard Frequency Index have few lexical markers, and those they have are grammatical markers. There is evidence that grammatical markers are reliably supplied through transfer features. Therefore, such words have very high Cloze Easiness. Words having a lower Standard Frequency Index have increasingly more lexical markers of the semantic and distinguisher variety which are less likely to be supplied through transfer features. Therefore, Cloze Easiness decreases as Standard Frequency Index decreases.

As Standard Frequency Index decreases to 50 and continues to diminish, the number of lexical markers on words reaches a critical

number such that it becomes unlikely that all markers will be supplied through transfer features. As Standard Frequency Index decreases below 50, Cloze Easiness bottoms out around 12. As Standard Frequency Index continues to decrease below 40, there is a slight increase in Cloze Easiness. This increase is consistent with the overall theory, since the theory predicts that words with low Standard Frequency Index but high Text Frequency will receive exceptional transfer feature support and increase in Cloze Easiness. Figure 4 shows that as Standard Frequency Index decreases to and below 40 there is a sharp increase in Text Frequency.

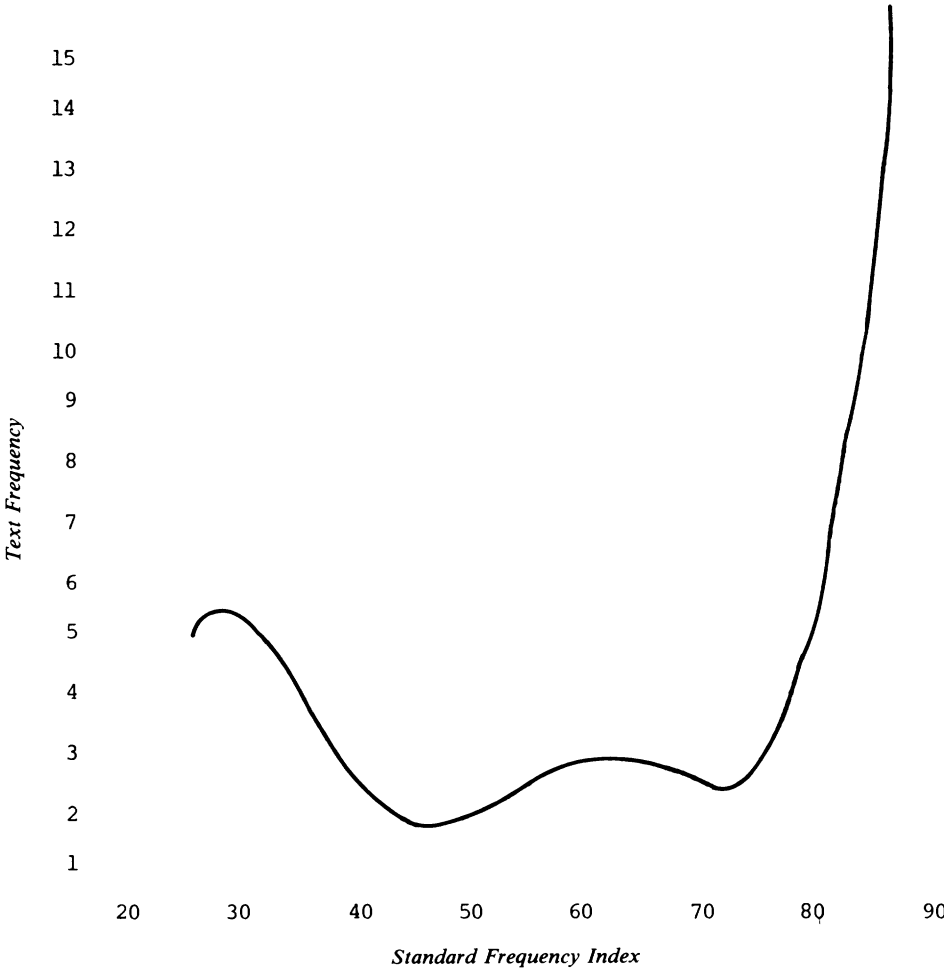


*Figure 4*
Fifth Degree Polynomial Relating Standard Frequency Index to Text Frequency for All Words (N = 5185, $r$ = .85)

Another way of showing the consistency of the data with the theory is to divide the data into 2 parts. From the theory, one would expect that for words at the high end of the Standard Frequency Index spectrum, Standard Frequency Index would account for a great deal of variance in Cloze Easiness and that Text Frequency would account for little additional variance. On the other hand, one would expect that for words at the low end of the Standard Frequency Index spectrum, Text Frequency would account for a great deal of variance in Cloze Easiness and that Standard Frequency Index would account for little additional variance.

## Rare words and frequent words defined

In order to explore the hypothesis based on rare words and frequent words, it had to be determined what Standard Frequency would be the dividing line between the 2 word groups. The words were divided into 2 groups using 80, 70, 60, and 50 as the dividing point. Multiple regression analysis was run on both sets of the divided data with Cloze Easiness as dependent variable and Standard Frequency Index, Text Frequency, and Passage Difficulty as the independent variables—chosen in order of the amount of independent variance they accounted for in Cloze Easiness. Standard Frequency 50 was chosen as the dividing line because this division resulted in the maximum amount of variance accounted for in Cloze Easiness for both Rare Words and Frequent Words. For the remainder of this report, "rare words" are defined as words having a Standard Frequency Index of less than 50 and "frequent words" are defined as words having a Standard Frequency of 50 or higher.

## Description of relationships among the variables for rare words and frequent words

Correlations among the 4 variables still under consideration were computed, and regression analysis was run on the 2 sets of words in the same manner as was reported on the whole corpus. Means and standard deviations for the variables for the 2 sets of words appear in Tables 6 and 9. The correlation matrices appear in Tables 7 and 10. The summary tables for the multiple regression analysis appear in Tables 8 and 11.

*Frequent words.* Both the correlation matrix (Table 7) and the multiple regression summary table (Table 8) for frequent words are very similar to the related tables for all words (Tables 4 and 5). The greatest

difference in the correlation matrix is that the correlation between Standard Frequency and Text Frequency has increased from .54 for all words to .62 for frequent words. One can predict with more confidence from the Standard Frequency Index of a word how often it will appear in a text. The greatest change in the multiple regression summary table is that the amount of independent variance accounted for in Cloze Easiness by Text Frequency has decreased from .064 for all words to .032 for frequent words. These data support the hypothesis stated above: For frequent words, the most important predictor of Cloze Easiness is Standard Frequency Index. Text Frequency is highly correlated with Standard Frequency Index and accounts for very little independent variance.

*Table 6* Description of the data for 6 variables for frequent words (N=4,677)

| | MEAN | STANDARD DEVIATION |
|---|---|---|
| 1. Cloze Easiness | 31.5 | 21.0 |
| 2. Standard Frequency Index | 71.3 | 10.9 |
| 3. Text Frequency | 5.9 | 7.6 |
| 4. Passage Difficulty | 2.5 | 1.1 |

*Table 7* Correlations among 6 variables for frequent words (N=4,677)

| | | 1 | 2 | 3 | 4 |
|---|---|---|---|---|---|
| 1. | Cloze Easiness | .99 | | | |
| 2. | Standard Frequency Index | .56 | .99 | | |
| 3. | Text Frequency | .46 | .62 | .99 | |
| 4. | Passage Difficulty | -.34 | .02 | .09 | .99 |

*Table 8* Multiple regression summary table for frequent words (SFI $>$ 50)—Cloze Easiness is the dependent variable

| Step Number | Variable | R | $R^2$ | Increment in $R^2$ |
|---|---|---|---|---|
| 1 | Standard Frequency Index | .561 | .315 | .315* |
| 2 | Passage Difficulty | .663 | .440 | .125* |
| 3 | Text Frequency | .687 | .472 | .032* |

*$p < .0001$

*Rare words.* Tables 10 and 11 show that for rare words, Text Frequency accounts for half the variance ($R^2$=497) in Cloze Easiness, and that Standard Frequency Index accounts for no independent variance in Cloze Easiness. This is consistent with the hypothesis that for rare words, Cloze Easiness is a function of transfer feature support rather than a

function of the relative number of lexical features associated with a word. Standard Frequency Index (the hypothetical correlate of the number of lexical markers associated with a word) does not correlate with Cloze Easiness, but Text Frequency (the hypothetical correlate of transfer feature support for rare words) accounts for half the variance in Cloze Easiness.

There are several other relationships which emerge in Tables 10 and 11 which support the lexical marker—transfer feature theory. For rare words, there is no correlation between Standard Frequency Index and Text Frequency (.00). This is consistent with the hypothesis that Text Frequency is a function of the degree of intimacy between rare words and the topic and not a function of the relative frequency of rare words in the language in general (reflected by Standard Frequency Index).

The correlation between Passage Difficulty and Cloze Easiness is higher for rare words (-.54) than for all words (-.38) and frequent words (-.34). This is consistent with the theory which states that Passage Difficulty will affect a reader's ability to utilize transfer feature support for all words in a passage, but that transfer feature support is the single important factor in determining the information value (and therefore the Cloze Easiness) of rare words.

Of further interest, the correlation between Passage Difficulty and Text Frequency is very high for rare words (-.42) compared to all words (.05) and frequent words (.09). This indicates that rare words tend to have higher Text Frequency in easier passages and lower Text Frequency in more difficult passages. If the theory is correct, easy passages would tend to have a dearth of high information words because potentially high information words (rare words—hypothesized to have many lexical markers) tend to have high Text Frequency and therefore much transfer feature support. This conclusion is in keeping with our intuition about Passage Difficulty and lends support to the theory.

Table 11 shows that Passage Difficulty accounts for less independent variance in Cloze Easiness for rare words (.074) than for all words (.116) or for frequent words (.125). This is explained by the fact that Passage Difficulty is highly correlated with the best predictor (Text Frequency) for rare words, but it is not at all correlated with the best predictor of Cloze Easiness (Standard Frequency Index) for all words and for frequent words.

## The lexical marker—transfer feature theory compared to current models of reading

*The Goodman model.* The theory is wholly consistent with the Goodman model of reading, and it suggests ways the Goodman model

*Table 9* Description of the data for 6 variables for rare words (N=508)

| | MEAN | STANDARD DEVIATION |
|---|---|---|
| 1. Cloze Easiness | 12.1 | 19.0 |
| 2. Standard Frequency Index | 40.5 | 8.9 |
| 3. Text Frequency | 2.5 | 2.7 |
| 4. Passage Difficulty | 2.9 | 1.0 |

*Table 10* Correlations among 6 variables for rare words (N=508)

| | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| 1. Cloze Easiness | .99 | | | |
| 2. Standard Frequency Index | -.05 | .99 | | |
| 3. Text Frequency | .70 | .00 | .99 | |
| 4. Passage Difficulty | -.54 | -.01 | -.42 | .99 |

*Table 11* Multiple regression summary table for rare words (SFI < 50)—Cloze Easiness is the dependent variable

| Step Number | Variable | R | $R^2$ | Increment in $R^2$ |
|---|---|---|---|---|
| 1 | Text Frequency | .705 | .497 | .497* |
| 2 | Passage Difficulty | .756 | .571 | .074* |

*p< .001

might be further researched and refined. The theory predicts that words having low Standard Frequency Index but high Text Frequency and words having high Standard Frequencies (regardless of Text Frequency), would be words requiring little attention to graphic cues, requiring little phonetic analysis, and not likely to be the source of miscues resulting in loss or distortion of meaning. Words with low Standard Frequency and low Text Frequency would be words requiring great attention to graphic cues, sometimes requiring substantial phonetic analysis, and would be likely to be the source of miscues resulting in the loss or distortion of meaning.

*The Calfee-Venezsky model.* Calfee and Venezsky posit 2 lexical lookups. Frequent words are "found" on the basis of such features as initial letter and length; less common words are stored differently and are more accessible by sound than by any other form. These 2 lexicons account for the fact that frequent words are processed rapidly and with little effort while less common words are processed with greater attention to phonic analysis.

A more parsimonious explanation would be that lexical markers are stored with words in a *single* lexicon, but that lexical markers are discovered through transfer features as well as by referring to

one's lexicon for each individual word. Since all, or nearly all, markers of frequent words tend to be supplied by transfer features, the reader does not have to look up frequent words to discover their features. This would account for the rapidity and apparent ease of processing frequent words. It would also predict that rare words having high Text Frequency would be rapidly and effortlessly processed as well—a phenomenon that runs counter to the 2-lexicon theory.

*The LaBerge and Samuels model.* LaBerge and Samuels have shown that readers give evidence of recognizing very common words while attention is deliberately diverted to another cognitive task but that they fail to recognize less common words under similar circumstances. They conclude that automaticity of decoding is a function of word frequency. They propose that the reader has a 2-fold task.

> For present purposes we find it convenient to separate comprehension from word meaning. By word meaning we refer to the semantic referent of a spoken or written word, morpheme, or group of words that denote a meaningful unit. By comprehension, on the other hand, we mean the organization of these word meanings. To do this, the meaning units presumably are scanned one by one by attention and organized into coherent wholes.... If a subject maintains attention solely on single word codes, this would constitute a rather low form of comprehension, much like viewing characters in a play one by one and ignoring their interactions. On the other hand, for high level comprehension of passages attention must be directed to organizing these meaning codes and presumably, this is where effort enters into reading just as it does in understanding difficult sentences.
>
> So long as word meanings are automatically processed the focus of attention remains at the semantic level and does not need to be switched to the visual system for decoding nor the phonological level for retrieving the semantic meanings. (1974, pp. 319, 320)

The lexical marker—transfer feature theory is consistent with the LaBerge-Samuels model insofar as both theories predict that frequent words will require little attention on the part of the reader. There is, however, an essential difference between the 2 theories. LaBerge and Samuels propose a 1-way process in which the physical occurrence of the word excites a meaning code. Their theory focuses on whether this process occurs automatically or with attention. They claim that automaticity is a function of frequency (in a Standard Frequency Index sense). They also take into account the fact that repetition of words (in a Text Frequency sense) facilitates automaticity. "When the child reads text in which the same vocabulary is used over and over again, the

repetitions will certainly make more automatic the perceptions of each word unit...." (p. 315). However, in both cases, they attribute automaticity to frequency (repetition of stimuli facilitates automatic responses). In contrast, the lexical marker—transfer feature theory suggests that automaticity of processing is a function of the amount of *information* in words and that it is only incidentally true that Standard Frequency Index and Text Frequency are highly correlated with *information*.

Another way of contrasting the 2 theories would be to say that LaBerge and Samuels claim that if the search for the word meaning can go on automatically, the reader can devote his attention to the more difficult problem of semantic processing or comprehension. The lexical marker—transfer feature theory proposes that so long as the semantic process is progressing satisfactorily, the assigning of word meaning will proceed automatically. In other words, the lexical marker—transfer feature theory states that semantic processing facilitates automaticity of word recognition whereas the LaBerge-Samuels theory states that automaticity of word recognition facilitates semantic processing.

## *Suggestions for further research*

Further evidence is needed to support the hypothesis that Standard Frequency Index correlates with the number of lexical markers associated with words and to support the transfer feature hypothesis. Three methods of enquiry are suggested here.

One might devise an experiment where passages are presented tachistascopically and where the subject's attention is diverted to another task while certain words were exposed. The lexical marker-transfer feature theory would predict that low information words would be perceived by the subject under these conditions, but high information words would not. Such a finding would support the LaBerge-Samuels position as well. However, the LaBerge-Samuels theory would not predict what "mistakes" the subject would make in his perception of words or what words he might use if asked to guess a word he did not perceive because his attention was diverted. The lexical marker-transfer feature theory would predict that both mistakes and guesses would be words having a subset of the lexical markers of the correct words—those features satisfying transfer feature requirements—and that mistakes and guesses would be words having a higher Standard Frequency Index than the correct words.

One might collect data on incorrect responses in cloze tasks. The lexical marker-transfer feature theory would predict that incorrect

responses would have a subset of lexical markers of the correct response—those markers satisfying transfer feature requirements—and that incorrect responses would tend to be words having a higher Standard Frequency Index than the correct responses.

Further evidence might be pursued in miscue data. There is evidence that miscues tend not to distort meaning. The theory would predict that miscues would be words having a subset of the actual word, that the miscue would have features that satisfied transfer feature requirements, and that the miscue would tend to have a higher Standard Frequency Index than the actual word.

If it is true that subjects systematically supply incorrect words having a subset of lexical features of correct words in the 3 conditions described above, one could formulate a kind of "Principal of Least Effort" (*cf.*, Zipf, 1935) of the reading process—that is, subjects find words having necessary and sufficient features to satisfy transfer feature requirements, but subjects do not speculate on possible features beyond these necessary to satisfy transfer feature requirements.

## *Summary*

The theory developed in this paper grew out of several observations. The cloze task has much in common with classical information theory experiments. A word's frequency and a word's Cloze Easiness are highly correlated. Frequent words tend to be function words and function words have few lexical markers. Weinrich's theory of transfer features would lead one to believe that frequent words tend to have high Cloze Easiness because they have so few features and that those they have (grammatical markers) are typically supplied through transfer features. Rare words that are repeated in a text are good candidates for a text index because they are important words in relation to the topic of the text. Such words tend to have high Cloze Easiness and such words would tend to have strong transfer feature support.

From these observations a theory was developed that *information* in a word in a text is a function of the number of lexical markers associated with the word not supplied through transfer features. Rare words tend to have more lexical features than frequent words, and therefore, frequent words are *low information* words (high Cloze Easiness). Rare words are generally *high information* (low Cloze Easiness), but when they are repeated in a text, it is because they are intimately related with the topic. Thus one would predict that they would receive extraordinary transfer feature support, and become *low information* words (high Cloze Easiness).

The theory was shown to be consistent reanalyzing data resulting from a study that administered cloze tests to 675 subjects over 20 passages at 4 levels of difficulty.

The theory was examined in light of 3 current models of reading. It was argued that the theory was consistent with some aspects of the models, that the theory may be more parsimonious than some aspects of the models, and that the theory suggests researchable questions which may refine, modify or refute some aspects of the models.

## REFERENCES


BOLINGER, DWIGHT. The atomization of meaning. *Language*, Oct.-Dec. 1965, *41*, 555-573.

BORMUTH, JOHN R. Readability: a new approach. *Reading Research Quarterly*, Spring 1966, *1*, 79-132.

CARROLL, JOHN B.; DAVIES, PETER; & RICHMAN, BARRY. *Word frequency book*. Boston: Houghton Mifflin, 1971.

CARROLL, J.N., & ROELOFFS, R. Computer selection of key words using word-frequency analysis. *American Documentation*, July 1969, *20*, 227-233.

DALE, EDGAR, & CHALL, JEANNE. A formula for predicting readability. *Educational Research Bulletin*, Jan. 21, 1948, *27*, 11-20, 37-54.

FILLENBAUM, SAMUEL; JONES, LYLE V.; & RAPOPORT, AMNON. The predictability of words and their grammatical classes as a function of rate of deletion from a speech transcript. *Journal of Verbal Learning and Verbal Behavior*, August 1963, *2*, 186-194.

FILLMORE, CHARLES A. The case for case. In E. Bach & R. Harms (Eds.) *Universals in linguistic theory*. New York: Holt, Rinehart and Winston, 1968. Pp. 1-88.

FINN, PATRICK J. The use of the relative frequency to identify high information words and to measure readability. In G. H. McNinch & Wallace D. Miller (Eds.) *Reading: convention and inquiry*. Clemson, South Carolina: National Reading Conference, 1975. Pp. 286-290.

FREDERIKSEN, CARL H. Representing the logical and semantic structure of knowledge acquired from discourse. *Cognitive Psychology*, July 1975, *1*, 371-458.

FRIEDMAN, JOYCE, & BREDT, THOMAS H. *Lexical insertion in transformational grammar*. Stanford University Computer Science Department, Computational Linguistics Project, 1968.

GOODMAN, KENNETH S. Reading: a psycholinguistic guessing game. In Harry Singer & Robert Ruddell (Eds.) *Theoretical models and processes of reading*. Newark, Delaware: International Reading Association, 1970. Pp. 259-271.

KATZ, J.J., & FODOR, J.A. The structure of semantic theory. In J. J. Katz and J. A. Fodor (Eds.) *The structure of language*. Englewood Cliffs, New Jersey: Prentiss Hall, 1964. Pp. 479-518.

LABERGE, DAVID, & SAMUELS, S. JAY. Toward a theory of automatic information processing in reading. *Cognitive Psychology*, April 1974, *6*, 293-323.

NEVILLE, MARY H., & PUGH, A.K. Context in reading and listening: variations in approach to cloze tasks. *Reading Research Quarterly*, 1976-1977, *12(1)*, 13-31.

SHANNON, CLAUDE E., & WEAVER, WARREN. *The mathematical theory of communication*. Urbana, Illinois: University of Illinois Press, 1949.

VENEZSKY, RICHARD L., & CALFEE, ROBERT C. The reading competency model. In Harry Singer & Robert B. Ruddell (Eds.) *Theoretical models and processes of reading*. Newark, Delaware: International Reading Association, 1970. Pp. 273-291.

WEAVER, WARREN. Recent contributions to the mathematical theory of communication. In Claude E. Shannon & Warren Weaver (Eds.) *The mathematical theory of communication*. Urbana, Illinois: University of Illinois Press, 1949. Pp. 1-28.

WEINREICH, URIEL. Explorations in semantic theory. In Thomas A. Sebeok (Ed.) *Current trends in linguistics, Vol. III*. The Hague: Mouton, 1966. Pp. 395-477.

ZIPF, G.K. *The psycho-biology of language*. Boston: Houghton Mifflin, 1935.