ED 280 851 TM 870 151

AUTHOR	Stabler, Edward P., Jr.
TITLE	Possible Contributing Factors in Test Item Difficulty. Research Memorandum.
INSTITUTION	Educational Testing Service, Princeton, N.J.
REPORT NO	ETS-RM-86-7
PUB DATE	Oct 86
NOTE	26p.
PUB TYPE	Reports - Evaluative/Feasibility (142)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	*Difficulty Level; Higher Education; Language Processing; *Linguistic Difficulty (Inherent); *Psychometrics; Sentence Diagraming; *Sentence Structure; Standardized Tests; Syntax; *Test Construction; Test Format; *Test Items
IDENTIFIERS	Sentence Completion Test

ABSTRACT

A type of item used frequently in standardized testing involves the recognition of a sentence. Examples of such items are the sentence completion items, used in both the Scholastic Aptitude Test (SAT) and the Graduate Record Examination (GRE), and the sentence correction items used in the Test of Standard Written English. Tests that contain such items are constructed by a laborious process which, remarkably, does not involve at any point a detailed analysis of the semantic or syntactic properties of the sentence on which the item is based. This paper provides an initial exploration of the possibility that a mental model of the item solution process may provide indications of how difficult it is to solve a certain item correctly. Some of the cognitive theories of language comprehension are reviewed to identify factors that may affect the level of effort required to solve a sentence-based item. It is a first step towards a test development process that does not rely exclusively on empirical test data analysis and instead views the characteristics of items as a source of psychometric information. (Author/JAZ)

RESEARCH MEMORANDUM

POSSIBLE CONTRIBUTING FACTORS IN TEST ITEM DIFFICULTY

Edward P. Stabler, Jr.

Educational Testing Service
Princeton, New Jersey
October 1986

Possible Contributing Factors in Test Item Difficulty

Edward P. Stabler, Jr.

University of Western Ontario

London, Ontario, Canada

POSSIBLE CONTRIBUTING FACTORS IN TEST ITEM DIFFICULTY

Edward P. Stabler, Jr.

Abstract

A type of item used frequently in standardized testing involves the recognition of a sentence. Examples of such items are the sentence completion items, used in both the SAT and GRE, and the sentence correction items used in the Test of Standard Written English. Tests that contain such items are constructed by a laborious process which, remarkably, does not involve at any point a detailed analysis of the semantic or syntactic properties of the sentence on which the item is based. This paper provides an initial exploration of the possiblity that a mental model of the item solution process may provide indications of how difficult it is to solve a certain item correctly. Some of the cognitive theories of language comprehension are reviewed to identify factors that may affect the level of effort required to solve a sentence-based item. It is a first step towards a test development process that does not rely exclusively on empirical test data analysis and instead views the characteristics of items as a source of psychometric information.

POSSIBLE CONTRIBUTING FACTORS IN TEST ITEM DIFFICULTY[1]

Edward P. Stabler, Jr.

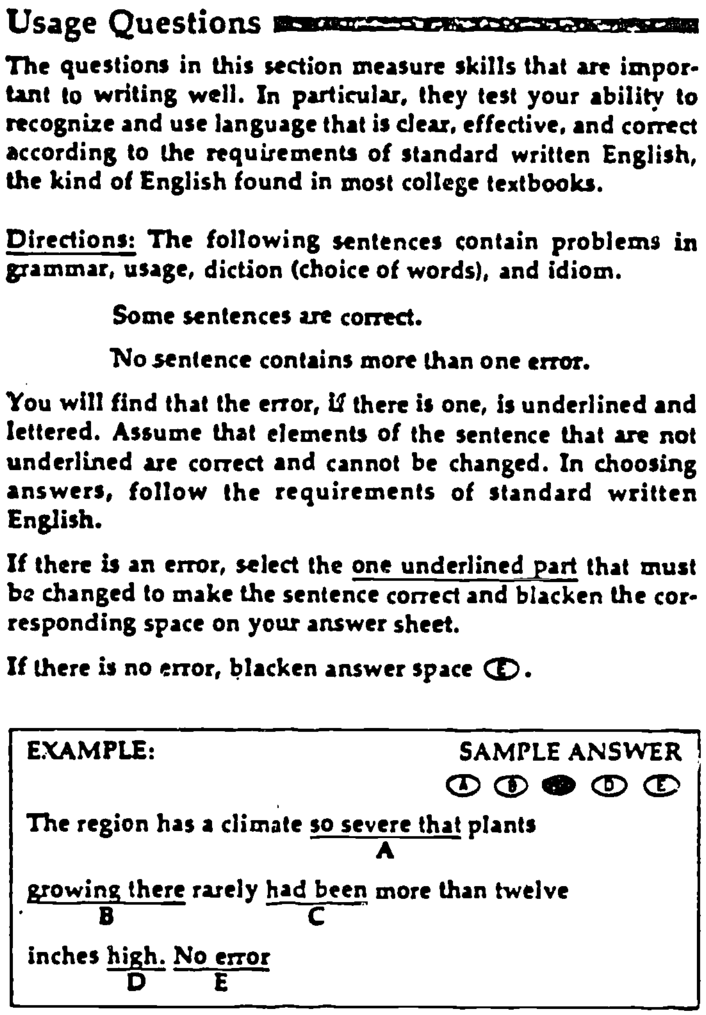
A type of item used frequently in standardized testing involves the recognition of a sentence. Examples of such items are the sentence completion items, used in both the SAT and GRE, and the usage (sentence correction) items used in the Test of Standard Written English. (Figures 1 and 2 show some typical items.) Tests that contain such items are constructed by a laborious process which, remarkably, does not involve at any point a detailed analysis of the semantic or syntactic properties of the sentence on which the item is based. Instead of a linguistic analysis, psychometric analyses of responses to the items are used to determine the relative difficulty and discrimination of the items. In addition, the items are submitted to multiple reviews to insure their adequacy and fairness. For example, the items are reviewed to insure that their content is not offensive to specific minority groups.

Were it not for the empirical analysis of responses to the items, the test developers would have very little indication of how hard or easy a given item is, yet that information is essential to the process of constructing multiple test forms that are comparable. Comparable forms are obtained by administering new items to examinees before the item is placed in a so-called final form. Although operationally this pretesting process is very smooth, it does complicate the logistics of test administration by requiring the insertion of sections into the test which contain pretest items only. If, as some coaching schools claim, students can be trained to identify the pretest sections then there is clearly a danger that some students would not cooperate and would decide to relax rather than respond to these pretest sections. If that were to occur frequently, the basis for the current method of test development could be in jeopardy.

The present paper does not aim to provide a solution to the problem described above. It is a first step, however, towards a test development process that does not rely exclusively on empirical test data analysis and instead views the characteristics of items as a source of psychometric information. That is, a mental model of the item solution process may provide indications of how difficult it is to solve a certain item correctly. If these indications of difficulty correspond with psychometric measures of item difficulty we can then use this information to create additional items with better control of their psychometric properties.

This paper reviews some of the factors that may affect the level of effort required to solve a sentence-based item. Specifically, the literature on sentence comprehension will be reviewed with an eye to identifying potential contributors to item difficulty of sentence-type items.

[1] I am indebted to Isaac I. Bejar for making this research possible, and for providing substantial suggestions for improving this paper. The comments of Erich Woisetschlaeger were also valuable. Of course, I am responsible for any remaining errors.



Usage Questions

The questions in this section measure skills that are important to writing well. In particular, they test your ability to recognize and use language that is clear, effective, and correct according to the requirements of standard written English, the kind of English found in most college textbooks.

Directions: The following sentences contain problems in grammar, usage, diction (choice of words), and idiom.

Some sentences are correct.

No sentence contains more than one error.

You will find that the error, if there is one, is underlined and lettered. Assume that elements of the sentence that are not underlined are correct and cannot be changed. In choosing answers, follow the requirements of standard written English.

If there is an error, select the one underlined part that must be changed to make the sentence correct and blacken the corresponding space on your answer sheet.

If there is no error, blacken answer space Ⓔ.

EXAMPLE: SAMPLE ANSWER

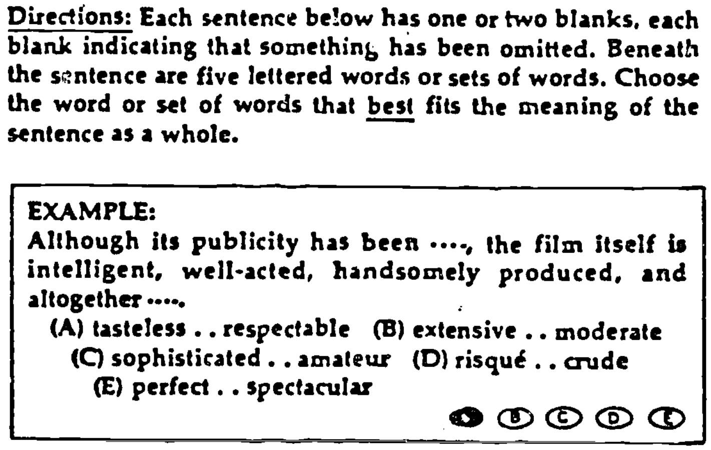
Ⓐ Ⓑ ● Ⓓ Ⓔ

The region has a climate so severe that plants
A

growing there rarely had been more than twelve
B C

inches high. No error
D E

Figure 1.



Directions: Each sentence below has one or two blanks, each blank indicating that something has been omitted. Beneath the sentence are five lettered words or sets of words. Choose the word or set of words that best fits the meaning of the sentence as a whole.

EXAMPLE:
Although its publicity has been ...., the film itself is intelligent, well-acted, handsomely produced, and altogether ....

(A) tasteless . . respectable (B) extensive . . moderate
(C) sophisticated . . amateur (D) risqué . . crude
(E) perfect . . spectacular

● Ⓑ Ⓒ Ⓓ Ⓔ

Figure 2.

Pyschological Models of Sentence Understanding. Almost all sentence processing models presuppose that in any task that involves sentence recognition and understanding, a number of representations of the linguistic input are formulated, whether the input is visual, auditory, or tactile. In reading, it is assumed that representations are formulated at each of the levels indicated in the following diagram:

stimulus
↓
1. graphemic/phonological
processing
↓
2. morphological
processing
↓
3. syntactic
processing
↓
4. semantic processing/
evaluation/
response formulation
↓
response

The complexity of formulating representations at each of these levels presumably contributes to overall task complexity in any task that involves reading. However, since test items are usually presented in a large set, with no measure or time limit on any particular item, much of the variability in the psychological complexity of sentence recognition will have no discernible effect. For example, evidence that recognizing a sentence with one structure takes some few milliseconds longer than recognizing a sentence with another structure is not really very good reason to think that sentences with the former, more complex, structures will increase the "difficulty" of an item, where this latter notion is related to the chances of choosing a solution that good students choose, or anything like that. It is possible that the structures which are demonstrably harder to process at the superficial levels are more liable to be misunderstood; indeed, in some cases, there is certainly a connection of this sort, as we will see. But in any case, this preliminary review may set the stage for a more careful discrimination of those factors that are worthy of further study from those that apparently have no bearing.

The theory of natural language processing is, unsurprisingly, most securely established at the most superficial levels, and in tasks where those superficial levels of processing most clearly play a critical role.[2] So, for example, there are fairly robust results and relatively secure theories

[2] The "natural languages" are those that humans can learn as a first language. Given the obvious diversity of languages and cultures, it is easy to forget the striking commonality in human linguistic experience -- the most striking common feature being the fact that children learn their native tongue, whatever it is, at about the same age and at about the same rate. Recent theoretical linguistics has provided evidence that all natural languages share many deep structural features which are manifested in superficially different systems.

about performance in relatively simple tasks like sentence-nonsense discrimination tasks, especially when the word strings presented to the subject contain familiar but not predictable words, when the strings themselves are unfamiliar, and when the input is not too hard to see (or hear). It is typically harder to explain what is going on in cases where the input is unclear (e.g., presented for a *very* short time, or masked by scribbled lines), or where the input is very familiar (i.e., cases in which the subject already has firm beliefs about what is being presented), or where the input is nonsense or obviously false, and so on. In these cases, people seem to resort to strategies that are not specifically linguistic -- strategies that draw extensively on background knowledge and problem-solving skills that are relatively poorly understood. That is why the above diagram of levels of linguistic representation relegates so much to the last stage. Semantic processing, evaluation and response formulation obviously are a most crucial and time-consuming part of many linguistic activities, but relatively little is known about what goes on here.

Many test items call upon the subject to analyze ill-formed linguistic input. This is yet another good example of a process that is very poorly understood, since it requires the subject to find some analysis of an input which has no proper, grammatical analysis. Even stating the requirements of this sort of task in formal terms is something that is beyond current theories of language (and so it is no surprise that computer simulation of human linguistic abilities falls particularly short in error recovery, in making sense of what is not literally correct). Responding correctly to many test items obviously depends on having some acquaintance with rather unusual vocabulary items and with stylistic rules that often conflict with the spoken and written material that is most prevalent in the subjects' environments. Again, these are the sorts of situations in which performance is hardest to study and explain, and about which relatively little is known. We begin with our review of some of the best established results about each level of linguistic representation, beginning with the relatively superficial levels.

1. Graphemics/phonology. The task begins with the recognition of the letters and graphemes (roughly, "letter groups") in the linguistic presentation. It is well known that this process is more or less complex depending on the identity of the letters and graphemes, the context of each particular letter and grapheme, and so on (McClelland and Rumelhart, 1981; Underwood and Bargh, 1982). Another factor that is related to recognition complexity is the regularity of the phoneme-grapheme correspondence: words whose pronunciation is in accord with regular rules of pronunciation tend to be easier to recognize, especially when they are relatively unfamiliar (McCusker et al., 1981; Underwood and Bargh, 1982). A word that is phonologically similar to another may also be harder to recognize, as evidenced, for example, by the finding of Rubenstein et al.(1971) that pronounceable pseudowords were easier to identify when they were not homophonic with any actual, familiar word.

2. Morphological processing. The recognition of morphemes (that is, roughly, the recognition of the meaningful words and word compounds) given the graphemic structure is similarly more or less complex depending on familiarity, Cloze value, etc.[3] (See, e.g., Miller and Isard, 1963, for a review.) Another factor influencing the psychological complexity of sentence recognition is lexical ambiguity, which may or may not be resolved by context. For example, in the sentence,

I bank at First Trust.

the noun-verb ambiguity is resolved by syntactic context, and the meaning ambiguity (among the readings for the verb) is (at least partially) resolved by the meaning of other words in the sentence and background knowledge. Even in cases like this, where the probable reading is easy to see, there is some evidence that the presence of the alternative readings increases recognition complexity (Swinney, 1979; Tannenhaus, Leiman and Seidenberg, 1979). Again, this is a feature whose relation to actual item complexity could easily be studied.

In sum, a number of features of lexical items themselves are known to affect sentence complexity. The influences here are so diverse and so clear that it would really be a surprise if they did not affect test item difficulty in many cases. A study to determine whether word frequency results and scales of Cloze value are in fact related to actual item difficulty looks like a promising line of research.

3. Syntactic processing. The recognition of the syntactic structure of linguistic input, given its morphological structure, is influenced by the number of syntactic structures that can be assigned, and by the complexity of those structures. A sentence is said to be syntactically ambiguous if it can be assigned more than one syntactic structure. We will consider some of the results on the psychological complexity of various sorts of sentence structures before considering the influence of ambiguity at this level of processing.

A. Structural complexity. It is well known that sentences with certain sorts of syntactic features are harder to understand than others. We will review only some of the best known results.

i. Length. The number of morphemes in the sentence is the most obvious determinant of sentence complexity. Other things being equal, short sentences tend to be understood more quickly and with less demand on memory. One would expect this to be a factor in test item difficulty.

ii. Center embeddings. There are a number of different sorts of center embedding constructions in English (and other natural languages). A phrase of any category (e.g., a sentence(S), verb

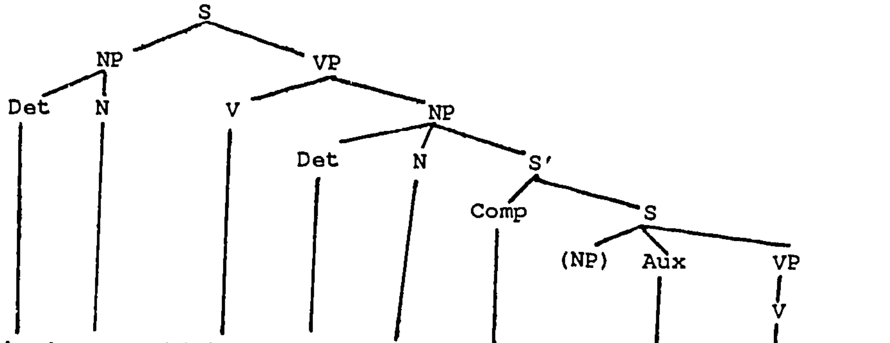
[3] A word is said to have "high Cloze value" in a particular context if, roughly, its occurrence in that context is expected, as the word "pepper" in "The soup is spiced with salt and pepper." As Fodor(1983) has pointed out, though, the common demonstrations of the importance of Cloze value should be tempered by considering results like those of Fishler and Bloom(1980), showing that in recognizing spoken words, the effect of high Cloze almost vanishes when the words are very clear, and they vanished entirely when the words were presented quickly, in which case, presumably, there was not enough time for any "guessing" strategy.

phrase(VP), noun phrase(NP), or whatever) is said to be **self-embedding** if it includes a phrase of the same category, or if it includes a phrase which is self-embedding. This feature is easy to see in a tree diagram. The sentences of Figure 3 are self-embedded, where X and Y are arbitrary categories.

```
          S                     S
         / \                   / \
        X   Y                 X   Y
       / \ / \               / \ / \
    ... S   ...           ...    Y ...
       / \                      / \
       ...                      ...
```

Figure 3. Self-embedding trees.

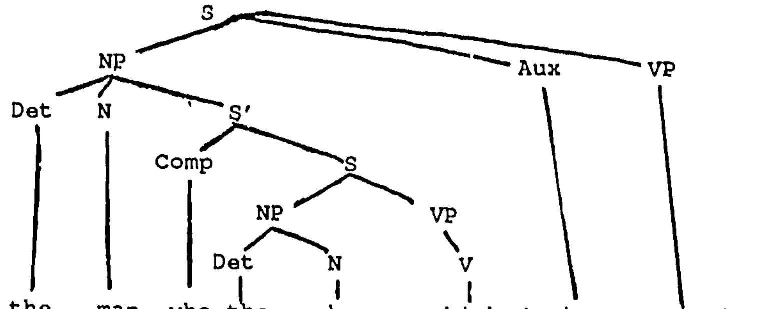
A phrase X of any category is **center embedded** if it contains a phrase of the same category **and** that phrase is neither the leftmost nor the rightmost phrase of X. If the included phrase is the rightmost we say the phrase is right embedding; if it is leftmost, we say it is left embedding. Relative clauses in English provide an example of right embedding and center embedding constructions. Examples are shown in Figures 4 and 5, respectively.



(i) the horse kicked the man who is recovering.

(ii) the jockey likes the h cse who kicked the man who is re:overing.

Figure 4. Sentences with right embedding through S.



(iii) the man who the horse kicked is recovering.

(iv) the man who the horse that the jockey likes kicked is recovering.

(v) the man who the horse that the jockey the agent told you about likes kicked is recovering.

(vi) the man who the horse that the jockey the agent who fixed the race that the audience which was terribly uninformed liked told you about likes kicked is recovering.

Figure 5. Sentences with center embedding through S.

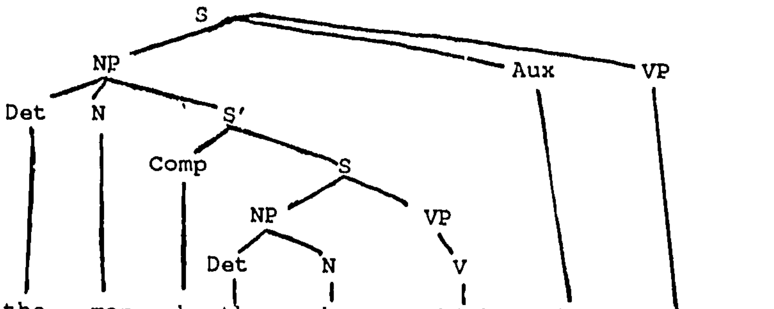
English tends to embed "to the right", as the "right heavy" trees of Figures 4 and 5 illustrate. Left embedding also occurs in English, but not so commonly, and not in relative clauses. One example of right embedding is provided by possessive constructions like the following:

John's father
John's father's uncle
John's father's uncle's brother
John's father's uncle's brother's mother

That these phrases are, in fact, left embedding, is intuitively clear when one notices that "John's father" is a subphrase of every one of these phrases, but the word string "father's uncle" is not a subphrase in any of them -- the thing that modifies "uncle" in all of these is not "father's" but "John's father's".

All self-embedded phrases may increase processing complexity slightly, but it is well known that center-embedded phrases are particularly difficult to understand.[4]Phrases with more than 3 senter embedded phrases are practically impossible to understand without paper and pencil, as example (vi) in Figure 5, above, illustrates. Phrases with more than 3 right embeddings, on the other hand, are relatively easy to understand (as in the children's story, "The House that Jack Built").

[4] Miller and Chomsky(1963) noticed the interesting coincidence that this special difficulty in human language understanding corresponds to the result that a certain kind of abstract automaton, a "finite state machine," is capable of dealing with right or left embedding but cannot recognize a language with arbitrarily deep center embedding.



(iii) the man who the horse kicked is recovering.

(iv) the man who the horse that the jockey likes kicked is recovering.

(v) the man who the horse that the jockey the agent told you about likes kicked is recovering.

(vi) the man who the horse that the jockey the agent who fixed the race that the audience which was terribly uninformed liked told you about likes kicked is recovering.

Figure 5. Sentences with center embedding through S.

English tends to embed "to the right", as the "right heavy" trees of Figures 4 and 5 illustrate. Left embedding also occurs in English, but not so commonly, and not in relative clauses. One example of right embedding is provided by possessive constructions like the following:

John's father
John's father's uncle
John's father's uncle's brother
John's father's uncle's brother's mother

That these phrases are, in fact, left embedding, is intuitively clear when one notices that "John's father" is a subphrase of every one of these phrases, but the word string "father's uncle" is not a subphrase in any of them -- the thing that modifies "uncle" in all of these is not "father's" but "John's father's".

All self-embedded phrases may increase processing complexity slightly, but it is well known that center-embedded phrases are particularly difficult to understand.[4]Phrases with more than 3 senter embedded phrases are practically impossible to understand without paper and pencil, as example (vi) in Figure 5, above, illustrates. Phrases with more than 3 right embeddings, on the other hand, are relatively easy to understand (as in the children's story, "The House that Jack Built").

[4] Miller and Chomsky(1963) noticed the interesting coincidence that this special difficulty in human language understanding corresponds to the result that a certain kind of abstract automaton, a "finite state machine," is capable of dealing with right or left embedding but cannot recognize a language with arbitrarily deep center embedding.

Why are center embedded constructions so hard to process? One reason is certainly that they create "discontinuous dependencies;" that is, recognizing the sentence as grammatical and understanding it requires relating sentence constituents that are not contiguous. In the last example, the last two words of the sentence ("is recovering") can be properly parsed and understood only if they are associated with the first two words of the sentence ("the man"). A phrase or other grammatical constituent is, by hypothesis, recognized as a unit and treated as such, and so any substantial intervening subphrase will delay that recognition. A language which allows self-embeddings will, given some common assumptions about the grammar, allow arbitrarily deep self-embeddings, and this allows subphrases to be arbitrarily complex. These phrases can have arbitrarily deep "nested dependencies." Other sorts of discontinuous dependencies, and even nested dependencies, will be considered below, but the relative clause constructions just considered provide the most deeply nested structures found in common English. Although deeply center-embedded constructions are unlikely to occur in test items (or in most other texts!), other sorts of "discontinuous dependencies" will be found, and these may quite generally have some influence on recognition and on item difficulty.

It should be noted that the influence of any of these factors on complexity may be hidden by other influences. One of the most powerful influences in almost all recognition tasks is semantic plausibility. For example, Anderson(1976) has noted that center embedded constructions are more readily understood when there are helpful semantic cues. So for example, the first of the following two sentences is more readily understood for this reason:

The cat the dog chased meows pitifully.
The dog the cat chased meows pitifully.

Semantic plausibility has an important effect on almost every complexity measure - even on the very short response latency measures of the simplest sentence decision tasks.

ii. Crossing dependencies. This is one sort of structure that is particularly difficult to parse with most sorts of computing systems.[5] There is some controversy about whether this sort of structure actually occurs in English, and about how it should be described if it does. The best known cases, first noticed by Bar-Hillel and Shamir(1960), involve the word "respectively." In the following sentences, for example, "beautiful" is understood to modify "the man," and "intelligent" is understood to modify "the woman:"

The man and the woman are beautiful and intelligent, respectively.

[5] In their classic paper, Miller and Chomsky(1963) drew attention to the observation of Bar-Hillel and Shamir(1960), that this sort of construction seems to occur in English, though it is beyond the capability of a certain kind of abstract automaton, a "pushdown automaton," which is capable of dealing with right, left, and center embedding.

A tree diagram of this sentence that tried to mark these relations by putting "beautiful" in the same phrase with "the man" and "intelligent" in the same phrase with "the woman" would have crossing branches. An alternative way to mark the crossing dependencies in a tree is with indices:

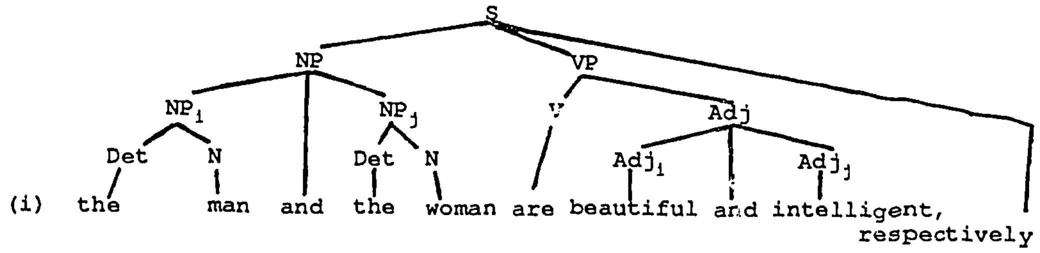

S
NP
VP
NP$_i$
NP$_j$
V
Adj
Det
N
Det
N
Adj$_i$
Adj$_j$
(i) the man and the woman are beautiful and intelligent, respectively

Figure 6. Sentence with crossing dependency marked by indices.

Of course, the important thing is not what notation we use to represent these crossing associations, but the fact that they must be recognized by a subject that understands the sentence.

The fact that these crossing dependencies really do correspond to something specifically linguistic is illustrated by the translations of sentences like these into languages like French in which many adjectives require number and gender agreement with the noun they modify. In English, we mark number agreement in the verb, and so we can construct cases like the following:

The man and the women smokes and drink, respectively.

Here it is the man who smokes and the women who drink, but the sentence sounds distinctly awkward. Sentences with more than two crossing pairs are harder to understand and sound even more awkward:

The man, the women and the child are beautiful, intelligent, and cheerful, respectively.
The man, the women, and the child smokes, drink, and plays, respectively.

It is no surprise that these more exotic cases do not occur very often in speech or writing, and so they would not be expected to occur in the items of most verbal aptitude tests. They have a clear theoretical interest, however, because if (abstracting away from memory limitations) they are counted as part of the language that humans can recognize, this would have important implications about the capabilities of the language recognition system.

iv. Filler-gap dependencies. One of the main levels of linguistic representation in Chomskian theoretical linguistics is the level called "S-structure" (Chomsky, 1981, 1982). Theories that do not propose S-structures typically propose a level which is quite similar. At this level, sentences are represented in tree-structures similar to those used in the figures above. One important feature of this level of representation is its ability to represent constituents that do not correspond to any word in the sentence. These so-called "empty categories" are really very well motivated, and are posited by all of the dominant traditions in western theoretical linguistics. There are usually assumed to be various types of empty categories. In the Chomskian tradition, these are the NP-trace, the wh-trace, PRO, VP-gap and so on. In some cases, these categories can be viewed as holding the former place of a lexically realized phrase which occurs elsewhere in the sentence. In these cases, we can regard the empty category, or "trace" as an unrealized element that refers to whatever the moved phrase refers to. This sort of coreference relation occurs with almost all empty categories: they must be "coreferential" with some non-empty constituent in the sentence.

It is easy to illustrate in a preliminary way the grammatical motivation for empty categories. One popular example (cf., Radford, 1981) involves the verb "put," which in simple declarative sentences requires a direct object and a locative phrase. Consider the following strings:

John put the car in the garage.
*John put in the garage.
*John put the car.

The second string is not a grammatical sentence because there is no direct object. The third lacks a locative. (The asterisk is the standard indicator of ungrammaticality in the literature of theoretical linguistics.) The verb "put" occurs in some grammatical strings, though, without being followed by a direct object:

Which car did John put in the garage?

In sentences like these, the direct object is regarded as having been "moved" to the front of the sentence. The sentence is, in effect, asking about the direct object of the verb. In recent transformational grammar, this relation between the wh-phrase and the object position is marked by putting an empty category, a "trace," in the object position and "coindexing" it with the moved phrase to indicate that the two must be coreferential:

[np_i Which car]did John put [np_i t] in the garage?

Traces themselves are not pronounced, of course, but are supposed to influence pronunciation. The association between the trace and the moved NP or wh-phrase must be noted in understanding these sentences.

When a sentence like the last example is processed from left to right in reading (or from the first word to last in listening), the processor must keep track of the fact that a wh-phrase has been heard, so that it will not look for (or listen for) a direct object for 'put'. Keeping track of this places an extra load on the system, which is evidenced in various complexity measures (e.g., Frazier et al., 1983).

Dependencies involving fronted wh-phrases can occur in various sorts of constructions and can extend over arbitrarily many words:

Who$_i$ did you see [np$_i$t]? (wh question)
I wonder who$_i$ you saw [np$_i$t]. (embedded question)
I saw a man who$_i$ you know [np$_i$t]. (restrictive relative clause)
Who$_i$ did you say that Bill said that...Mike said I saw [np$_i$t]?
I wonder who$_i$ you said that Bill said that...Mike said I saw [np$_i$t].

One might speculate that the complexity of such sentences is proportional to the complexity of the material that separates the dependent items. Many other sorts of constructions involve similar dependencies:

John [kissed Mary], and I think that Frank said that Mary thought that Harry would have [vp e] too.

[np$_i$ The city] was destroyed [np$_i$t] by the enemy.
[np$_i$ John] seems [np$_i$t] to like Mary.
[np$_i$ John] seems [np$_i$t] to appear [t]$_i$ to like Mary.
[np$_i$ The city] seems [np$_i$t] to have been destroyed [np$_i$t] by the enemy.
[np$_i$ Which violins] are [np$_i$ these sonatas] easy to play [np$_i$t] on [np$_i$t]?
[np$_i$ Who] did [np$_j$ you] ask [np$_i$t] whether [np$_j$t] to blame yourself$_j$?

These are not all treated in the same way in current linguistic theory, but are similar in having some sort of filler-gap dependency. Notice that the dependencies in the second to the last example are nested, and in the last example they are crossing.

It is plausible that the mere number of filler-gap dependencies can dramatically influence complexity. On the Berwick and Weinberg (1984) model of human sentence recognition, a sentence of n words can be parsed in an amount of time proportional to n if it contains no more than 1 NP-trace or wh-trace; otherwise it may take an amount of time proportional to n squared. (This upper bound has not been confirmed by psychological study, but neither has it been disconfirmed.) Frazier ct al. (1983) found that sentences with nested filler-gaps were actually harder to process than sentences with no overlapping filler-gap dependencies. Another interesting study is Wanner's (1968, reported in Fodor, Bever and Garrett, 1974). He found that in a prompted recall task, for a sentence like:

The governor asked the detective to cease drinking.

the phrase 'the detective' is a better prompt than it is for a sentence like:

The governor asked the detective to prevent drinking.

Using the standard transformational theory of the 1960's as his framework, Wanner explained this sort of result by maintaining that understanding these sentences involved formulating their deep structures, and 'the detective' occurs three times in the deep structure of the former sentence and only twice in the latter. In current transformational grammar, many deep structure relations (perhaps all of them -- this is part of current controversy) are marked at a single level, S-structure, and semantic interpretation is presumed to be definable on S-structure, so the explanation of Wanner's results should be different. The analogous account would maintain that understanding a sentence involves formulating its S-structure, and the S-structure of the former sentence,

the governor [asked [the detective]$_i$ [[comp t$_i$]][PRO$_i$ to cease [t$_i$ drinking]]]

contains more empty categories co-indexed with 'the detective; than the S-structure of the latter does,

the governor [asked [the detective]$_i$ [[PRO$_i$ to prevent [[PRO drinking]]]]

(Notice that no distinction between PRO and NP-trace is needed for this explanation. The difference comes from the fact that in the latter sentence, the PRO in the object clause of prevent is not "subject controlled"; i.e., it is not necessarily coreferential with the subject of 'prevent'.)

v. Other discontinuous dependencies. Many of the constructions noted so far involve non-adjacent items which are in some sense dependent on one another. This can happen in many other sorts of constructions as well. For example, subject-verb agreement may need to be enforced across arbitrarily much intervening material:

The house - and I do not mean to brag - is beautiful.

* The house - and of course I do not mean to brag to you of all people - are beautiful.

There seem to be problems here.

* There seems likely to seem likely to be problems here.

It is plausible that the complexity of the material intervening between the subject and the verb increases the difficulty of enforcing the proper agreement. Chomsky (1963) noted that we get similar discontinuous dependencies in constructions like 'either...or', 'if...then' 'both...and'. A similar dependency holds between a reflexive pronoun and its antecedent, in certain verb particle constructions, in tag questions, and many other cases.

vi. Reversible passives. It was originally thought that passive constructions like "The city was destroyed [t]" were all harder to process than simple active sentences (and this was taken to be evidence for the "derivational theory of complexity"(DTC)), but it was later discovered that only "reversible" passives are relatively complex (Slobin, 1966; Gough, 1965). Reversible passives are just passives which would make sense even if the subject and object were switched; thus

"John was hit by Mary" is reversible, but "The cookies were smelled by John" is not. Forster and Olbrei (1973) argue, however, that when overall semantic plausibility is controlled for, reversibility effects disappear, and passives remain generally hard to process. This idea is consistent with the treatment of (most) passives as the result of NP-movement, producing a NP-trace relation that must be recognized.

vii. Inversions. It has been suggested that departures from the standard subject-verb-object ordering of phrases in English may increase complexity, but this conjecture is not supported by the evidence. (This has been discussed at length in various places, again because of its relation to what is called the derivational theory of complexity (DTC). See, e.g., Fodor, Bever and Garrett, 1974; Berwick and Weinberg, 1984.) Thus, some psychological models of sentence parsing (e.g., Marcus, 1980) count among their virtues the fact that questions with subject-auxiliary inversion are parsed by their models in (almost) the same time as the corresponding declarative forms:

You are coming to the show.
Are you coming to the show?
You will be coming to the show.
Will you be coming to the show?

viii. Garden paths. Some sentences are hard to understand because they have ambiguous beginnings. There are a wide range of such sentences, and it is controversial whether they are all hard for exactly the same reasons. Some examples are the following:

The horse raced past the barn fell.
I told the boy the dog bit Sue would help him.
The grocer always orders a hundred pound bags of sugar.
The prime number few.
The man who hunts ducks out on weekends.
Cotton clothing is made of grows Mississippi.

It is possible that some such sentences occur in the item pool and contribute to task complexity. There is some controversy about how to define this class of sentences formally, and the proposed definitions are rather complicated. Sentences of this class could be recognized with a parser of sufficiently broad coverage, or with hand calculation by someone who has been trained to use a particular definition.

B. Ambiguity. The second main contributor to the complexity of syntactic processing is syntactic ambiguity. We say that a string of morphemes is ambiguous if it has more than one grammatical sentence structure. It is no surprise that ambiguous sentences would be more complex, since the processor must either formulate all of the acceptable structures or else decide which structure to formulate. Either option would be expected to demand some resources. Usually, the different structures correspond to different meanings, as in the famous example:

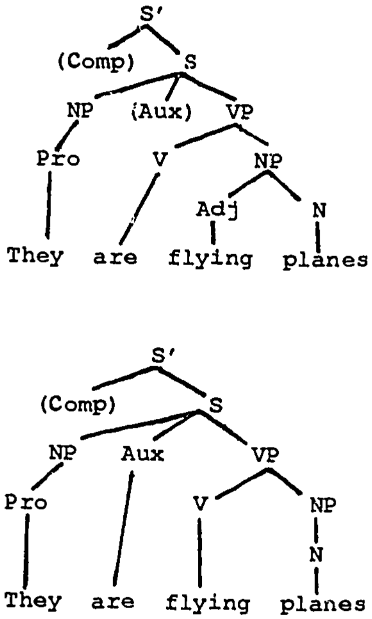


Figure 6. Syntactic ambiguity

The first of these structures corresponds to the interpretation of the sentence according to which it tells us what kind of planes they are; the second might be interpreted as telling us something about what the Navy is doing these days. Sometimes, there will be some pragmatic, semantic or even structural bias in favor of only one of the structures, and in these cases the ambiguity is not noticed by the speaker. Even when the ambiguity is not noticed, though, it can be shown to increase processing complexity (e.g., MacKay 1966 on the effect of ambiguity in a sentence completion task). However, there is also evidence that in a biasing context, the subject very quickly settles on the preferred reading and seems unable to reinterpret the sentence without reprocessing it (e.g., Carey et al. 1970). So it is not clear that the effects of ambiguity will generally show up in a task of the sort of interest here. These effects would be expected only in cases where the preferred structure turns out to be the incorrect one.

In any case, there are a number of very common sorts of syntactic ambiguity that can be noted. They all pose serious problems for mechanical parsers because there is no adequate account of the structural preferences and biasing contexts.

i. Prepositional phrase attachment. In many cases, there will be more than one position for a prepositional phrase (PP) in a grammatical structure. Consider the following sentence:

Bring the book from the library.

In this sentence the PP can be part of the NP or it can be part of the VP separate from the NP. This structural distinction corresponds to a difference in meaning: the sentence can either be a request to bring the book that is in the library, or a request to bring the library book from whatever location. Things get more complicated when there is more than one PP. Consider:

Put the block in the box on the table.

ii. Clause attachment. A similar sort of attachment ambiguity can arise in sentences with embedded or coordinated clauses. Consider:

A student is expected to study the material until the term ends.

This could be assigned either of the following structures:

A student is expected to [vp study [the material] [until the term ends]].
A student is [vp expected [s' to study the material] [until the term ends]].

iii. Noun-noun modification. This sort of modification can be either left or right branching -- the appropriate structure is usually determined on semantic grounds. In the following phrases, for example, the structures indicated are obviously the ones preferred on semantic grounds -- others are syntactically correct:

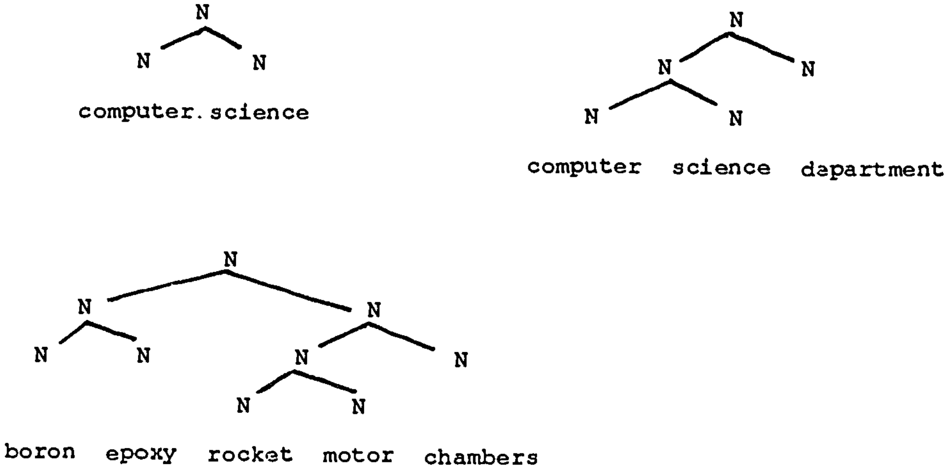


Figure 7. Intended structures of some noun-noun compounds

The treatment of these expressions as N-compounds rather than as phrases follow Selkirk (1982).

iv. Coordinate structure reductions. Coordinate structures (e.g., conjunctions with 'and' disjunctions with 'or', etc.) may generally add to the difficulty of comprehension, even in simple sentences like,

I like Mary and Sally.

Coordinate structures can be more complicated, though, when an ambiguity is introduced by the possibility that the structure is "reduced." For example. all of the following sentences can be used to mean the same thing:

Mary goes to the best school in the bus with John, and Sally goes to the best school in the bus with John, too.

Mary goes to the best school in the bus with John, and Sally goes to the best school in the bus, too.

Mary goes to the best school in the bus with John, and Sally goes to the best school too.

Mary goes to the best school in the bus with John, and Sally goes too.

The latter sentences may mean that same thing as the first, but they may not. This is sometimes treated as a syntactic matter. Another similar sort of case is the following:

Give me the names of students in English 101 and English 102.

The last 3 words of this sentence could be a simple NP conjunction, in which case the sentence requests the names of each student who is taking both classes. On the other hand, the sentence could be reduced, in which case the sentence requests the names of students in English 101 and of students in English 102.

v. *Quantifier scoping*. Another sort of ambiguity that increases complexity is introduced by interactions between quantifiers (like 'every', 'some', 'each', 'all', 'any', 'each', 'several', 'five') and determiners (like 'a', 'the', 'those') in a sentence. Consider:

Every woman loves some man.

This could be interpreted as meaning either that there is a particular man (Arnold Schwartznegger?) that every woman loves, or that for any woman you consider, there is some man that she loves. A similar ambiguity is present in the sentence:

You can fool all of the people some of the time.

Does this mean that there is some particular time (from 3 to 4 a.m.) at which all of the people can be fooled? No, commonsense resolves the ambiguity in favor of the interpretation according to which it means that for any person you take, there will be some time or other at which that person can be fooled.

Things get even more complicated when there are more quantifiers and determiners. It has been suggested (Hobbs, 1983) that the following sentence has 120 (i.e., 5 factorial) different interpretations:

In most democratic countries most politicians can fool most of the people on almost every issue most of the time.

It is implausible that there are 120 psychologically distinct readings of this sentence, but there is no question that multiple quantifiers may increase complexity. Of course, some pairs of quantifiers will not interact. For example, in the following sentences we do not get the "every...some" interactions noted in the earlier examples:

Every man breathes, and some woman knows it.

Every man knows that I believe some woman is superior.

4. Character of errors. There has not been very much study of the processing of ungrammatical strings, but some of the available results suggest that the type of error in a sentence can have a definite influence on the difficulty of identifying and correcting that error.

A. Constraint vs. rule violations. Many of the linguistic theories that have been proposed distinguish between "constraints" which hold for all human languages and "rules" which define the particular properties of each individual language. The constraints are often assumed to be "innately given", rather than learned. This simplifies the learning task by making it unnecessary for the learner to eliminate lots of possibilities about the language spoken. So, for example, the learner never needs to consider the hypothesis that the set of grammatical strings of English includes all word sequences whose lengths are prime numbers.

There is some evidence (Freedman, 1982) that constraint violations are actually harder to identify than rule violations. So, for example, fewer subjects identify the first of the following sentences as ungrammatical:

* What did you buy Picasso's painting of?

* Mary were writing a letter to her husband.

Crain and Fodor (1984) did not find the same difference between

* Which dog did the man think that had bitten him?

* John and there was a fly in the soup.

In any case, different sorts of ungrammatical strings are certainly identified as such with different degrees of success by fluent speakers of the language. Crain and Fodor (1984) suggest that "correctability" may be a factor -- it is harder to correct 'What did you buy Picasso's painting of?' than it is to correct 'Mary were writing a letter to her husband'.

B. Stylistic errors and pragmatic plausibility. Many -- probably most -- of the errors that the schools are concerned with are not "grammatical" errors in the linguists' or psychologists' sense; rather, they are stylistic errors. That is, they are "errors" only in the sense that the sentences are not part of the dialect that is "in style." (Sometimes linguists use the adjective 'stylistic' in a different, unrelated sense, to refer to rules that are not like move-NP or move-Wh.) Linguists and psychologists assume that the grammar is what enables us to speak our language. We do not use this system consciously, but very quickly and automatically in an unconscious manner. The rules of the internalized grammar need to be distinguished from stylistic rules and from mere pragmatic peculiarity.

In school we are taught how to write and speak "properly." We are taught to make our speech and prose correspond to various conventional "stylistic rules." This teaching is quite unlike what goes on in first language learning (cf., Brown and Hanlon, 1970; or Pinker, 1979 on the lack of training and on the futility of training in first language acquisition). And -- as is introspectively obvious -- what goes on in the application of a stylistic rule is quite different from what goes on in unselfconscious use of the language. It is plausible, though, that the stylistic rules, with practice, may be internalized and become rules of "grammar" (in the linguists' sense of that term). This amounts to learning a new dialect.

Both grammatical and stylistic rules need to be distinguished from mere pragmatic peculiarities. As Radford (1975, ch.1) points out, a noun phrase like 'the tree who we saw' may seem ungrammatical at first, but it is perfectly appropriate in a story in which trees have human characteristics.

5. Conclusions. A number of the well-studied contributors to sentence understanding complexity that have been reviewed above appear to be plausible contributors to test item difficulty. Some of these factors are investigated in Bejar and Stabler(1986). The influences on actual test performance are so various that detecting the influence of, for example, structural complexity may be rather difficult, but it is hard to believe that such psychologically important aspects of sentence understanding would not be relevant. The line of study begun here and in Bejar and Stabler(1986) may provide valuable insights for test item development and assessment.

REFERENCES

Anderson, J.R. (1976) Language, Memory and Thought (Hillsdale, NJ: Lawrence Erlbaum).

Bejar, I.I. and Stabler, E.P. (1986) Syntactic complexity and psychometric difficulty: a preliminary investigation. ETS Technical Report.

Berwick, R.C. and Weinberg, A.S. (1984) The Grammatical Basis of Linguistic Performance (Cambridge, MA: MIT Press/Bradford).

Brown, R. and Hanlon, C. (1970) Derivational complexity and order of acquisition in child speech. In J.R. Hayes (ed.) Cognition and the development of Language (NY: Wiley).

Carey, P.W., Mehler, J. and Bever, T. G. (1970) Judging the veracity of ambiguous sentences. Journal of Verbal Learning and Verbal Behavior, 9:243-254.

Chomsky, N. (1963) Formal properties of grammars. In R.D. Luce, R. Bush and E. Galanter, eds., Handbook of Mathematical Psychology, Volume 2 (NY: Wiley).

Chomsky, N. (1981) Lectures on Government and Binding (Dordrecht: Foris Publications).

Chomsky, N. (1982) Some Concepts and Consequences of the Theory of Government and Binding (Cambridge, MA: MIT Press).

Church, K.W. (1980) On memory limitations in natural language processing. Unpublished MIT technical report, MIT/LCS/TR-245.

Crain, S. and Fodor, J.D. (1984) How can grammars help parsers? in D.R. Dowty, L. Kartunnen, and A. Zwicky, eds., Natural Language Processing (NY: Cambridge University Press).

Fodor, J.A. (1983) The Modularity of Mind (Cambridge, MA: MIT Press/Bradford).

Fodor, J.A., Bever, T.G., and Garrett, M. F. (1974) The Psychology of Language (NY: McGraw-Hill)

Forster, K. (1979) Levels of Processing and the structure of the language processor. In W. Cooper and E. Walker, eds., Sentence Processing (Hillsdale, NJ: Lawrence Erlbaum, pp. 27-86.

Forster, K. and Olbrei, I. (1973) Semantic heuristics and syntactic analysis. Cognition, 2:319-347.

Frazier, L., Clifton, C. and Randall, J. (1983) Filling gaps: decision principles and structure in sentences comprehension. Cognition, 13:187-222.

Freedman, S.A. (1982) Behavioral Reflexes of Constraints on Transformations. Unpublished PhD Dissertation. Monash University, Australia.

Gough, D. (1965) The verification of sentences. The Journal of Verbal Learning and Verbal Behavior, 5:107-111.

Hobbs, J.R. (1983) An improper treatment of quantification in ordinary English. Proceedings of the 21st annual meeting of the Association of Computational Linguistics, pp. 57-63.

Langendoen, T. (1976) On the adequacy of type 3 and type 2 grammars for human languages. CUNY Forum, 1:1-12.

McClelland, J.L. and Rumelhart, D.E. (1981) An interactive activation model of context effects in letter perception: Part 1. An account of basic findings. Psychological Review, 5:375-407.

McCusker, L., Hillinger, M.L. and Bias, R.G. (1981) Phonological recoding and reading. Psychology Bulletin, 89:217-245.

MacKay, D.G. (1966) To end ambiguous sentences. Perception and Psychophysics, 1:426-436.

Marcus, M. (1980) A Theory of Syntactic Recognition for Natural Language (Cambridge, MA: MIT Press).

Miller, G.A. and Chomsky, N. (1963) Finitary models of language users. In R.D. Luce, R. Bush and E. Galanter, eds., Handbook of Mathematical Psychology, Volume 2 (NY: Wiley).

Pinker, S. (1979) Formal models of language learning. Cognition, 7:217-283.

Pullum, G.K. and Gazdar, G. (1983) Natural languages and context free languages. Linguistics and Philosophy, 4:471-504.

Radford, A. (1981) Transformational Syntax: A Student's Guide to Chomsky's Extended Standard Theory (NY: Cambridge University Press).

Rubenstein, H., Lewis, S.S. and Rubenstein, M.A. (1971) Evidence for phonemic recoding in visual word recognition. Journal of Verbal Learning and Verbal Behavior, 10:645-657.

Selkirk, E.O. (1982) The Syntax of Words (Cambridge, MA: MIT Press).

Slobin, D. (1966) Grammatical transformations and sentence comprehension in childhood and adulthood. Journal of Verbal Learning and Verbal Behavior, 5:219-227.

Swinney, D. (1979) Lexical access during sentence comprehension: (Re) consideration of context effects. Journal of Verbal Learning and Verbal Behavior, 18:645-660.

Underwood, G. and Bargh, K. (1982) Word shape, orthographic regularity, and contextual interactions in a reading task. Cognition 12:197-209.

Tannenhaus, M., Leiman, J., and Seidenberg, M. (1979) Evidence for multiple stages in the processing of ambiguous words in syntactic contexts. Journal of Verbal Learning and Verbal Behavior, 18:427-441.