[From: Allen Kent (ed.) *Information retrieval and machine translation*
(New York: Interscience, 1959)]

CHAPTER 31

The Isolation of Elements for a Grammatical Description of Language*

LILA R. GLEITMAN

University of Pennsylvania, Philadelphia, Pennsylvania

A necessary step in the construction of a recognition grammar is the establishment of the elements which are said to be the components of sentences. The description of the language will differ as the elements chosen for the analysis are different. We try to choose that assignment which results in the simplest grammatical description, maintains information, and reflects the notions of the speaker about the language. It is difficult to achieve such results without setting up a rigorous procedure for deciding on the disposition of each element. Criteria of meaningfulness result in unreliable judgments because of the number of different semantic, morphological and grammatical concepts which are subsumed under that name.

In this paper, we will show how elements can be isolated for use within a particular model of language structure. In so doing, we will describe the general method for assigning elements to classes, since these two operations are interdependent. The grammatical model for which this assignment of elements will be done is known as a transformational grammar. A transformation is a relation whose domain is a class of sentences characterized by a succession of word-classes and whose counterdomain is a class of sentences characterized by a succession of word-classes each of whose members also occurred (perhaps differently ordered) in the domain, and possibly some constant elements. For example, for every sentence of the form:

$$N_1 V_i N_2, \dagger$$

*This paper is based on the work done by a group in the Department of Linguistics at the University of Pennsylvania under the direction of Zellig S. Harris. Besides the author, the other members of this group are Aravind Joshi, Henry Hiż, Naomi Sager and Bruria Kaufman. The project is sponsored by the National Science Foundation.

† N = noun, V_i = A class of verbs; roughly, the class of verbs completable by a noun phrase.

there is a sentence:

$$N_2 \text{ is } V_i\text{-en by } N_1$$

En and by are constant elements in the counterdomain. That is, if "John eats the apple," is an occurring sentence in English, we know mechanically that "The apple is eaten by John." is also a legitimate sentence in English.

The assignment of elements to classes depends on the set of transformations assumed for the language and the participation of the elements in these transformations. When two elements behave identically under transformations, they are members of the same class. For this kind of model, words, defined in English orthography by word-space, provide a good first approximation to elements. In large part, sentences can be described as some sequences of units which are divided by word-space. However, it is sometimes the case that word-length and element-length are not conveniently equated. A variety of sequences of letters within and over word-space provide more useful units for analysis. Any workable mechanical description must take into account the fact that the conventions of print are partially arbitrary, or else recognize that they will show up as cumbrous and artificial adjustments in the functional description of the language. This is also a problem in the spoken language, although the problems there are somewhat different.

There are two cases in which word-space division is insufficient for a transformational model. First and most frequent is the case of more than one grammatical or informational function within a single word. A familiar example is the affixing process of English which, besides adding certain informational properties to a root, changes its grammatical properties and thus alters its positional occurrence in sentences. Presumably no descriptive system for English will succeed without recognizing the partial similarity of affixed and unaffixed forms and characterizing the grammatical operators which make them partially different. In addition, there are words or roots which appear both independently and together with other roots inside word-spaces. An example of a word containing two roots is ferricyanide. These combinations never affect the class-membership of words, but the components must be recognizable as occurrences of more than one item of information, just as they are in hyphenated forms.

The second problem which word-space fails to account for is the rare case in which two or more concatenated words, each of which is an element assignable to a class, on occasion do not behave grammatically like other concatenated members of their word-classes. Such sequences are known as "idioms." An idiom exists only relative to a particular grammatical system with a particular assignment of elements to classes. It represents an occurring sequence of classes which cannot be defined or which must be defined as not allowable within the grammar. This is not the case where multiple classifica-

tion of an element provides a simple solution; that is, where the position and word-classes of the other elements in the sentence suffice to determine the appropriate assignment. It is the case in which we need to know precisely which members of which word-classes are co-occurring in order to recognize the structure of the sentence. In English, sequences like according to and because of have particular functions in sentences different from all other sequences of the same classes. They are more easily isolated in the spoken language, where characteristic pitch and stress serve to identify them. The written language gives no such clue, evidently relying on the speakers' recognition of the high probability of an idiomatic construction when its parts appear contiguously.

Another class of these so-called word-complexes arises in print because of certain arbitrary orthographic conventions. Some prefixes appear occasionally with word-space, sometimes with a hyphen, and sometimes without either when they are newly added to a word; this is usually a matter of editorial judgment, but we must recognize the same entity in either case. Words like can't and shouldn't, spelled with an intervening apostrophe which otherwise occurs at word-space introduce a related problem.

We will first discuss in detail the isolation of elements smaller than the word. As we have stated, it is by seeing how elements behave under transformations that we establish their identity and relationships. Should a sequence of letters smaller than a word frequently participate in a fixed way in a given transformation, we would call that sequence an element. For example, the sequence of letters -tion is often affixed to verbs with the effect of nominalizing them. Therefore we establish a class of nominalizing suffixes. Should this same sequence of letters occur after a sequence which cannot be construed to be a verb, the element is said not to have occurred. Given the word nation we know immediately that the sequence of letters does not represent the suffix, since there is no verb nate. However, the case is rarely that simple. Very frequently we run into this kind of problem: given the word ration and the rules by which spelling is affected by the addition of the suffix, we could claim that ration = rate + tion. From this it can be seen that the suffix cannot be isolated mechanically in any simple way. Another problem forces us to the same conclusion: suffixes need to be assumed for a simple grammatical description when the sequence of letters gives little or no cue to the fact of a relationship. Examples are go + ed, giving went; sheep +s, giving sheep, think + a nominalization, giving thought.

Taking the example of rate and tion, we now apply a transformational test for a relationship. Tion participates in the following transformation with a set of verbs, V_j;* Given the sentence:

Sums are calculated by machine.

---

*Roughly, the set of verbs which can be completed by a single noun.

We obtain the nominalized sentence:

The calculation of sums by machine.

Given:

Words are created by John.

We obtain:

The creation of words by John

But given:

Banks are rated by Dun and Bradstreet.

We do not obtain:

The ration of banks by Dun and Bradstreet.

Ration is then excluded from the class of Verb + tion and is considered by virtue of its positional occurrence in sentences as a single noun or verb, unrelated to the verb rate. Unless such tests can be met, two words are considered unrelated transformationally to each other, even when a morphological relation can be shown; that is, when it is clear that the word was originally formed with the affix, but any transformational relationship with the unaffixed form has been lost.

Transformational tests are made by speakers of the language with altogether reliable results. A particular kind of semantic satisfaction is simultaneously achieved. Without such a method, speakers confuse morphological, semantic, and syntactic relationships. The transformational criterion succeeds in separating out a given relationship.

Perhaps the most interesting example of words consisting of more than one element is the set who, whom, which, what. By division of these words we achieve a simplification of the grammatical description as a whole. Before explicating this example, it is necessary to point out this much about the general grammatical description which we use. We assume that the English sentence can be described as a noun phrase, followed by a verb phrase with the object of that verb, plus a set of transformations. The verb-object can be of many kinds, and the isolation and description of verb-objects is central in the grammar. A sentence is considered well formed only if the subject, verb, and object can all be shown to have occurred. The sentence may contain additional such units provided a conjunction appears between them. In the general case, it is claimed that conjunctions join two sentences. That is, conjunctions are constant elements in transformations taking $S_i + S_i$ .....+ S_n into S.* Parts of conjoined sentences may be "zeroed;" that is, if a complete sentence does not appear to the right or left of the conjunction, we are able to specify exactly those elements which can be inserted to form two complete sentences.

*S = sentence.

With these facts in mind, we will consider the forms who, whom, which, what. To show that these forms have a pronominal function is simple. Verbs which do not take a pronoun as their object cannot directly precede these forms, and verbs which may be completed by a pronoun always can take these forms. For example, one can say:

I take what I like.

But never:

I live what …

In fact, these forms are no different in their characteristic noun-replacing functions from other pronouns, so that we say that who is equivalent to he or she, and whom to him or her, etc. We can easily specify the positions in which we expect the wh- form or the normal pronominal form. One can show further that this set of pronouns may serve simultaneously as the object of one verb and the subject or object of the next, a so-called word-sharing structure common in English. For example, in the sentence:

I know whom I like.

We have two verbs which each require at least an object-noun for completion. Whom is the only choice for both verbs in this sentence. Let us return to the example:

I take what I like.

Since we have shown that two well-formed sentences:

I take it. (or them)
I like it. (or them)

have occurred, we are required to accept one of two conclusions: either that the verb take accepts a complete sentence as its object, or that a conjunction has occurred between two sentences. Whenever a sentence-object has occurred for a verb, we know that the form that either appears or can be inserted between the verb and its object. It is not insertable here—one cannot say "I take that what I like."—and furthermore this particular verb, take, in no other environment accepts a sentence as its object. We are thus forced to the conclusion that a conjunction has occurred. The only candidate for the status of conjunction is again the word whom. For this reason, we subdivide the word into a conjunction, wh-. with special positional properties, and a pronoun, it or them.

In what way does this simplify the description of English grammar? First, if whom is not considered a pronoun, we must assume that, given such forms, verbs can appear without their objects. Second, it is not considered a conjunction, we must assume that all verbs which can take any object at all can take a sentence as their object. Either assumption forces us to consider legitimate and well formed a large group of non-occurring sentences in English. The

proposed division, on the other hand, requires no new grammatical statements beyond the definition of the conjunction. The inverted order: object, subject, verb, which occurs for the second sentence must be accounted for for reasons external to such forms, since we have sentences like:

This I believe.
In God we trust.

The classification into two elements accounts for the positional occurrence of the class. When we handle them as single units, we must write descriptions of at least two new and otherwise nonoccurring types of well-formed sentence.

We can now consider the unit larger than a word. We say that such a unit exists when we discover a sequence of words in which the individual members do not retain their normal grammatical functions; that is, sets of words occurring contiguously whose individual grammatical values taken in sequence differ from the grammatical function of the sequence as a whole. For example, take the sequence because of. A machine dictionary must classify because as a conjunction, and of as a preposition, but the two together generally behave like a single preposition, and not the preposition of. The sentence:

Men retire because of illness.

Would be written by the machine as:

Noun Verb Conjunction Preposition Noun.

If because of is a case of Conjunction + Preposition and it is held that conjunctions join words or phrases of like structure, it would have to be concluded that some Preposition + Noun, or that some Preposition precedes the conjunction, or else that Preposition + Noun is a well-formed sentence. The first conclusion is patently false, and the second destroys grammatical restrictions which generally hold for the written language. For this reason, we reclassify the sequence because of as a new single preposition, and ignore the word-space.

The presence of anomalous structures of this kind in a sentence might perhaps be discoverable by operations on the otherwise analyzed sentence, if a nonwell-formed sentence resulted from regular analysis, but the limits and function of the structure could not be precisely identified. In other words, such phrases are additions to the grammar and cannot be subsumed under any prior economical description of the language. In a writing system which mirrored a grammar, such sequences would have unique spellings. In fact, they are generally marked off by punctuation, and they take a different stress in the spoken language. For example, compare:

Men retire because, of the partners, they are the eldest.

Unfortunately, commas are not used reliably in English. Since these so-called word-complexes are few, it is possible to list them and decide individually on their disposition.

Because their members occur separately, it is impossible to deal with word-complexes in the dictionary. All one can do in the original classification is to identify words which could participate in idiomatic structures. After the machine has completed the dictionary look-up operation, we require that it make a pre-run through the text material to isolate these sequences. A library of such forms gives the relative location for the word or words which complete the word-complex and direct that the composition be assigned the status of a single element of specified grammatical classification.

Of less theoretical interest, but of prime importance in mechanical recognition are orthographic anomalies. For example: in the word "shouldn't" we do not wish to admit two words between which the punctuation mark, apostrophe, occurs, since apostrophe then becomes non-unique in interpretation, complicating the grammar unnecessarily. Another case is the word percent which sometimes receives word-space and sometimes does not. Either way we wish to say that the same single element has occurred, that it is not the concatenation of the independent words per and cent; the existence of these separate words in the language requires that the situation be resolved outside the framework of the dictionary.

We resort in this case to precisely the same method for mapping two words onto a single element. As the previous examples show, the contiguous occurrence of the parts of these structures does not automatically determine that the word-complex has occurred, although the probability that it has is extremely high. When the word-complex is the result of a purely orthographic convention, and in a few other cases, the sequence can indeed be guaranteed to be a special structure without reference to the sentence in which it is embedded. For this latter type, we replace the individual class names of the members by zero, giving the value of the total sequence to the final word in the sequence. Otherwise, we make a tentative decision in favor of the high probability word-complex value, and use this value for making local decisions about surrounding words and phrases. We reserve the final decision for comparison of the analyzed sentence with a description of well-formed sentence types. Very occasionally, both readings are legitimate, and the machine must state that an ambiguity exists. For example, in the sequence according to, we can construct the ambiguous sentence:

Caesar's slave is according to Caesar that which is Caesar's.

In effect, the "absolute word-complex" after initial discovery is rewritten from

$$X + Y + Z \text{ to } O + O + Q,$$

where Q is the word-complex value, and the "tentative word-complex" from

$$X + Y + Z \text{ to } O(X) + O(Y) + Q(Z).$$

In summary, we use word-space only as a starting point for the identification of the combinatory elements of language, but every deviation from the word-space description is justified on grammatical rather than semantic grounds. That the grammatically justifiable solution satisfactorily mirrors the semantic intuitions of speakers is itself the strength of this kind of description of English.