CONGRAMMATICALITY, BATTERIES OF TRANSFORMATIONS AND GRAMMATICAL CATEGORIES

BY

HENRY HIŻ

I am dealing here[1] with that branch of grammar which takes a body of sentences given empirically as its starting point, which compares sentences, which arranges them in grammatically connected sets and, by pointing out their similarities and differences arrives at structures of sentences and at smaller units into which sentences are analyzable. This procedure is manifestly the reverse of a constructive or generative path in grammar where one attempts to build up sentences from more elementary constituents.

Between some sentences of a language a specific grammatical relation holds which may be called congrammaticality. This can be illustrated by a pair of sentences of which one is a negation of the other or a pair where one sentence is a question to which the other is an answer. Congrammaticality is, roughly speaking, the relation of being sentences for similar reasons. This intuitively obvious relation is cumbersome to describe precisely, but it can be defined within the usual conceptual apparatus. Congrammaticality is a relation between sentences and not between abstract structures, like trees, even though abstract structures may be needed to define it. Just as the relation of successor occurs between natural numbers, but to define it one uses tools from set theory.

To define congrammaticality I make six assumptions.

(1) It is understood what a segment of a text is. In many examples that follow, segments will be continuous fragments of written texts, fragments that may be composed of half words but not of half letters. For other purposes one can take phonetic fragments of utterances as segments.

(2) Concatenation[2] of segments leads to new segments. Thus concatenation of two segments that actually occur in a language may lead to a segment that does not occur in the language. In English an initial $l \frown r$ does not occur though its two components do occur in other environments. The notion of a segment is therefore revealing about the language under study only in this: a segment is divisible into segments each of which occurs in the language; some segments are never parts of sentences.

(3) Conformity[3] is a relation between segments by virtue of which two segments are recognized as looking the same.

(4) The unspecific large quantifier 'there are many' is used in the way

[1] This paper was written while I was engaged in the Transformations and Discourse Analysis Project at the University of Pennsylvania. The project is sponsored by the National Science Foundation. All co-workers on the project contributed substantially to the ideas presented in this paper.

[2] About concatenation see [8, pp. 172–174] and [5, p. 217].

[3] The notion of conformity was first introduced into the study of (formal) languages in [6].

suitable for linguistics. To say, e.g., that in a large segmented text there are many segments, or discontinuous combinations of segments, satisfying a given condition means that a suitably large proportion of segments of the text, or of such combinations, satisfy the condition. To say 'there are many' is not the same as to say 'there are some' nor 'there are infinitely many.' The claim that there are many adjectives in Latin does not assert that they are infinitely numerous, but that the set of Latin adjective constitutes, say, 4 per cent of the total Latin vocabulary. The assertion that there are many Latin prepositions is false, because it is easy to list all the Latin prepositions.

(5) A list of specific segmental constants

$$C_1, C_2, \cdots, C_n$$

is presupposed. The choice of constants for the list is not unique for a given language, and grammars of the language may slightly differ due to different choices of constants. In practice every such list will contain most frequent "grammatical morphemes" and will resemble the list of chapter titles in a grammar primer. Prepositions, flexion suffixes and prefixes, articles, classifiers, conjunctions are typical constants useful as a basis for a grammar. It seems that grammars are often built by distinguishing some classes of segments with only a few members, each of the other classes containing very many members.

(6) Lastly, it is assumed that sentences are segments which are sharply distinguished from segments that are not sentences. Though in practice there may be hesitations about sentencehood of many segments, the grammatical considerations that follow are relative to a decision (possibly an arbitrary one) in each case.

As a technical device I shall use the concept of a sequence of symbols each of which is either a constant C_i from the assumed list or a numeral. A numeral is used here just as a place-holder. Two sequences may be such that the very comparison of them forces us to distinguish between some elements but does not require us to distinguish between some others; so that, two symbols in succession may be taken as if they were a single symbol. Thus in the pair of sequences

$$\langle 1, 2, 3\rangle \text{ and } \langle 3, 1, 2\rangle$$

1 and 2 are not distinguished but 3 is.

A recursive definition of a distinguished symbol in a set of sequences says that:

(a) every constant is distinguished;

(b) if a sequence in the set is composed of one single symbol only, then the symbol is distinguished;

(c) every symbol that appears in two sequences of the set with different immediate neighbors is distinguished (more precisely this is composed of three conditions: (c′) if there are sequences $X = \langle x_1, x_2, \cdots, x_k\rangle$, $Y = \langle y_1, y_2, \cdots, y_m\rangle$, $Z = \langle z_1, z_2, \cdots, z_p\rangle$ (Y and Z not necessarily different, but each different from X) and there are i, j, h (not necessarily different) such that x_i, y_j and z_h all conform to each other and x_{i+1} does not conform to y_{j+1} and x_{i-1} does not conform to z_{h-1} where $1 < i < k$, $1 < j < m$, $1 < h < p$, then x_i is distinguished; (c″) if there are X and Y as before and there is j $(1 < j < m)$ such that x_1 conforms to y_j and x_2 does not conform to y_{j+1}, then x_1 is distinguished; (c‴) if there are X, Y, and j as before such that x_k conforms to y_j and x_{k-1} does not conform to y_{j-1}, then x_k is distinguished);

(d) if the immediate neighbors are distinguished, then the symbol is distinguished (this is composed again of three conditions on $X = \langle x_1, x_2, \cdots, x_k \rangle$ in the set: (d′) if x_{i-1} and x_{i+1} $(1 < i < k)$ are distinguished, then x_i is distinguished; (d″) if x_2 is distinguished, then x_1 is distinguished; (d‴) if x_{k-1} is distinguished, then x_k is distinguished);

(e) if a symbol conforms to a distinguished symbol, it is distinguished.

If in a (finite) set of sequences every symbol is distinguished and the set of numerals occurring in one is the same as the set of numerals occurring in any other sequence, then the sequences of the set are called related. Thus if sequences are related, they differ at most by constants, permutations, or repetitions.

Here are a few examples of sets of related sequences.

(1) $\langle 1 \rangle$, $\langle 1 \rangle$;
(2) $\langle 1, C \rangle$, $\langle 1, C \rangle$;
(3) $\langle 1, 2, 3, 4, 5 \rangle$, $\langle 5, 1, 2, 3, 4, 3, 1 \rangle$;
(4) $\langle 1, 2, 3, 1 \rangle$, $\langle 3, 2, 2, 3, 1 \rangle$;
(5) $\langle 3, 1, 2, 4 \rangle$, $\langle 3, 1, 2, 3, 4 \rangle$, $\langle 2, 3, 4, 1 \rangle$;
(6) $\langle 1, 2, 3, 4 \rangle$, $\langle 1, 2, 4, 3 \rangle$, $\langle 3, 2, 4, 1 \rangle$.

About sets of related sequences one can prove the following theorem.

If, in a set of related sequences, **a, d** occurs and **b, e** occurs, then the set of sequences resulting from replacing **a** by **b, c** and **e** by **c, d** is also a set of related sequences.

The theory of related sequences, besides its applications to linguistic analysis, may be of interest in itself. Sequences considered here resemble the kind of graphs Goodman and Pownall[4] consider. An obvious, but not easy, generalization may be to sets of n-dimensional matrices; another one to infinite sequences.

If in a set of related sequences one replaces numerals by linguist segments uniformly (conforming segments for conforming numerals) and one obtains a sentence in place of each sequence of the set, then these sentences are locally congrammatical. "Locally" here means "irrespectively of the rest of the language."

For instance

(1) The pill hits the window

and

(2) The wind hits the pillow

are locally congrammatical. They are obtained from the pair of related sequences $\langle 1, 2, 3, 4, 5 \rangle$ and $\langle 1, 4, 3, 2, 5 \rangle$ by replacements: *the* for 1, *pill* for 2, *hits the* for 3, *wind* for 4, *ow* for 5.

If we pay due respect to the rest of the language we define a set of sentences to be congrammatical when they are locally congrammatical and for each

[4] See [3] and [7].

segment used in replacing numerals of the related sequences there are many and varying segments which could have been used instead.

(1) and (2) are congrammatical only locally but not globally.

Instead of *the* in the sentences (1) and (2) one could use *a*, *my*, *a hard* and many others, but instead of *pill* one can use only either a very few words like *bill* or segments always ending on one of those few words—not satisfying the requirement of variety of segments which one could use.

Note that the requirement that a numeral of the sequence be replaceable by many and varying segments preserving sentencehood in each case is to be understood in such a way that one changes replacements for a single numeral holding all the other segments the same. Otherwise nearly all the sentences of a language may prove to be congrammatical.

The same pair of sequences, ⟨1, 2, 3, 4, 5⟩ and ⟨1, 4, 3, 2, 5⟩ has many replacements that are globally congrammatical, e.g.

(3) Almost all young, vigorous, healthy boys know how to swim

(4) Almost all healthy, vigorous, young boys know how to swim.

In case of the pair of sentences

(5) A young and healthy boy swims

(6) A healthy and young boy swims

the use of the same sequences ⟨1, 2, 3, 4, 5⟩ and ⟨1, 4, 3, 2, 5⟩ is not the most convenient. For, instead of *a* (or instead of *and*), not a great variety of segments can be used. But *a* and *and* are among most convenient constants for English. (5) and (6) are globally congrammatical via the pair of sequences: ⟨a, 1, and, 2, 3, s⟩ and ⟨a, 2, and, 1, 3, s⟩.

We defined congrammaticality of sentences using a method based on comparing sentences with the same material (except for constants) and discovering where and how they differ and whether the difference is a systematic feature of the language or only a local, isolated phenomenon. Note that congrammaticality is reflexive, symmetrical, and transitive. But congrammaticality with respect to a sequence (the notion more useful than the general congrammaticality) is not transitive.

Obvious changes in the preceding definitions will extend the notion of congrammaticality to cover a case when a sentence is congrammatical not with a single sentence, but with two, three, or any finite set of sentences. Thus

(7) John likes to swim but the weather is unstable

is congrammatical with the set of two sentences

(8) John likes to swim

and

(9) The weather is unstable.

It may happen that two sentences are congrammatical via a given set of two related sequences and two other sentences are congrammatical via the same set. This, as a rule, is not sufficient to conclude that there is a strong grammatical connection between one pair of sentences and the other pair. Thus

(10) John thinks she is pretty

and

(11) John thinks that she is pretty

can be obtained from the set of related sequences $\langle 1, 2\rangle$, $\langle 1, \text{that}, 2\rangle$. From this set one obtains similarly

(12) John likes painting

and

(13) John likes that painting.

Obviously, the sentences (10) and (11) are grammatically affiliated in many ways in which (12) and (13) are not.

The dissimilarity of these pairs may be explained in two ways.

Firstly, the set of admissible replacers for *she is pretty* in (10) and (11) is considerably different than the set of admissible replacers for *painting* in (12) and (13); many segments which can replace *she is pretty* in the first pair cannot replace *painting* in the second. One can define a co-occurrence[5] difference, on a particular position, between similarly congrammatical pairs of sentences as the set of those segments which can occur at this position in both sentences of the first pair and not in both sentences of the second pair. Slightly more precisely: If X is a sequence in a set A of related sequences, a sentence α is obtained from X by replacing segments for numerals, and every segment obtained by similar replacements (conforming segments for conforming numerals) from every sequence in A is a sentence, then we say that α enters A at X.

If α, α' and β all enter A at X and α' is like α except for containing a segment γ in every place in which α contains a segment that replaces the numeral x_i of X and the segment β' which is like β except for containing γ in every place in which β contains a segment that replaces x_i is not a sentence, then γ is said to belong to the co-occurrence difference on the x_ith place between α and β.

Note that the co-occurrence difference is not symmetrical.

Thus (10) and (12) enter the set $\{\langle 1, 2\rangle, \langle 1, \text{that } 2\rangle\}$ at $\langle 1, 2\rangle$. (11) and (13) enter the same set at $\langle 1, \text{that } 2\rangle$. The co-occurrence difference on the second place between (12) and (10) is substantial and contains *painting*, *swimming*, etc. The second place co-occurrence difference between (10) and (12) contains nearly every sentence.

The structural strangeness in relating some sentences (like (10) and (12)) that enter a set of related sequences at the same sequence may be explained by studying their co-occurrence differences. As a rule, greater strangeness goes together with larger sets of co-occurrence differences. Moreover, the co-occurrence differences at various places may differ in size, and in this way "show" the pattern of strangeness between the sentences.

Secondly, besides studying their co-occurrence differences the intuitive unrelatedness of two pairs of sentences (like (10), (11), as against (12), (13)) may be traced to the fact that the sentences of the first pair are each congrammatical with another sentence via some new set of related sequences whereas

[5] The notion of co-occurrence was introduced in [4]. Here it is used in a slightly different way. For, firstly, only the co-occurrence difference and not a positive co-occurrence is used here, and secondly co-occurrence was in [4, p. 285], relative to grammatical categories like noun, adjective, etc., whereas here it is relative to a set of related sequences.

there is no sentence which is congrammatical with a sentence of the second pair via this particular set of related sequences. E.g., (10) is congrammatical with

(14) John thinks hence she is pretty

via sequences ⟨1, 2⟩ and ⟨1, hence, 2⟩ but

(15) John likes hence painting

is not a sentence.

Hence (10) enters a different set of related sequences than (12).

(10) enters, for example, the set

(A) ⟨1, 2⟩, ⟨1, that 2⟩, ⟨1, hence, 2⟩

(12) enters, among others, the set

(B) ⟨1, 2⟩, ⟨1, that, 2⟩, ⟨1, his, 2⟩.

It is important that (10) does not enter (B) and that (12) does not enter (A). In the study of a language it is useful to record not only what can occur, but also what quite definitely cannot occur. Therefore it may be helpful to introduce into a set of sequences also negative components. E.g., to characterize (10) more fully we may add *not* ⟨1, his, 2⟩ to (A).

Now, a set the elements of which are either sequences or negatives of sequences and which is such that all sequences that take part in it (occurring positively or negatively) together constitute a set of related sequences is called a battery of transformations.[6] In other words a battery of transformations is composed of sequences that form a set of related sequences after all the negatives of sequences are taken as positive occurrences of the same sequences. For a negative of a sequence $\langle x_1, x_2, \cdots, x_m\rangle$ I shall write

$$*\langle x_1, x_2, \cdots, x_m\rangle\,.$$

(10) enters at ⟨1, 2⟩ the battery

(A′) ⟨1, 2⟩, ⟨1, that, 2⟩, ⟨1, hence, 2⟩, ⟨2, and, 1⟩, ⟨1, that, 1, that, 2⟩, *⟨1, his, 2⟩, *⟨1, the, 2⟩, *⟨a, 2, is what, 1⟩.

On the other hand (12) enters at ⟨1, 2⟩ the battery

(B′) ⟨1, 2⟩, ⟨1, that, 2⟩, ⟨1, his, 2⟩, ⟨1, the, 2⟩, ⟨1, that, 1, 2⟩, ⟨a, 2, is what, 1⟩, *⟨1, hence, 2⟩, *⟨2, and, 1⟩.

(A′) and (B′) differ in many components. This fact constitutes the basis for a second explanation of the feeling of strangeness between (10) and (12), the first explanation having been found in the large co-occurrence differences between (10) and (12). That is, although two sentences enter a battery of transformations at the same sequence, one of them may enter another battery which the second sentence does not enter.

There are intrinsic connections between co-occurrence differences of two sentences and the batteries of transformations that they enter. About these connections I would like to propose two hypotheses.

Hypothesis 1. If two sentences α and α' that enter the same battery A of transformations at the same sequence X have a substantial co-occurrence difference on the kth place, then there are batteries of transformations B and

[6] A battery of transformations is a concept that is a generalization of two (different) notions of transformations, one found in [4], the other in [1] and [2].

B' such that α enters B at X and α' enters B' at X, α does not enter B' nor α' enters B, many sentences resulting from α by putting at the kth place elements of the co-occurrence difference enter B at X, and many sentences obtained from α' by putting at the kth place elements of the co-occurrence difference do not enter B'.

Hypothesis 1 claims, roughly speaking, that if there are substantial differences in what can be substituted at a given place in two sentences, then the two sentences are subject to different batteries of transformations. A battery (B) reflects a large co-occurrence difference.

To state the second hypothesis I need a definition of a refinement.

A sequence X' is a simple refinement of a sequence X if and only if X' is obtained from X by replacing an element by one or by several elements.

A sequence X' is a refinement of the sequence X if and only if there is a finite sequence of sequences which starts by X and ends by X' and is such that each sequence in it is a simple refinement of the preceding one.

A battery B is a refinement of a battery A if and only if for every sequence in A there is a refinement of it in B obtained by a uniform process of replacements.

Hypothesis 2. If α and α' enter a battery A of transformations at the same sequence X and if the co-occurrence difference between α and α' as they enter A is substantial, then there is a battery B such that B is a refinement of A and the co-occurrence difference between α and α' as they enter B is substantially smaller than in the first case.

In other words, refining the segmentations and adjusting the batteries one can considerably reduce a co-occurrence difference.

In this paper batteries of transformations play a similar role to the role played in other approaches by sequences of grammatical categories of successive segments in a sentence. One can therefore take numerals occurring in sequences of a battery as grammatical categories. A grammatical category is, then, relative to a battery. If Hypotheses 1 and 2 are true, then by building suitable batteries of transformations one can obtain more suitable grammatical categories. This does not mean, however, that there exists a uniform method of discovering suitable batteries for a language. There is no substitute for insight. But the methods described here are in many ways close to actually grasping the sentence structures in a language. Moreover, the methods are general enough to abstract from peculiarities of languages on which much of grammatical work is modeled. There is an effort here to present grammatical studies of a language as something open, changing, progressing. There are infinitely many batteries of transformations for a language. We change our grammar, our global view of the language, by shifting our attention from one finite set of batteries to another.

Bibliography

1. N. Chomsky, *Three models for the description of language*, IRE Transactions on Information Theory vol. IT-2 (1956) pp. 113–124.

2. ———, *Syntactic structures*, 's-Gravenhage, Mouton and Co., 1957.

3. Nelson Goodman, *Graphs for linguistics*, this volume, pp. 51–55.

4. Zellig S. Harris, *Co-occurrence and transformation in linguistic structure*, Language vol. 33 (1957) pp. 283–340.

5. H. Hiż, *Types and environments*, Phil. Sci. vol. 24 (1957) pp. 215–220.

6. S. Leśniewski, *Grundzuge eines neuen Systems der Grundlagen der Mathematik*, Fund. Math. vol. 14 (1929) pp. 1–81.

7. Malcolm W. Pownall, *An investigation of a conjecture of Goodman*, University of Pennsylvania Doctoral Dissertation, 1959.

8. A. Tarski, *Logic, semantics, metamathematics*, Oxford University Press, 1956.

University of Pennsylvania,
Philadelphia, Pennsylvania