MEMORANDUM
RM-4828-PR
DECEMBER 1965

STUDIES IN INTER-SENTENCE CONNECTION

Kenneth E. Harper

PREPARED FOR:
UNITED STATES AIR FORCE PROJECT RAND

The RAND Corporation
SANTA MONICA • CALIFORNIA

MEMORANDUM
RM-4828-PR
DECEMBER 1965

STUDIES IN INTER-SENTENCE CONNECTION

Kenneth E. Harper

This research is sponsored by the United States Air Force under Project RAND—Contract No. AF 49(638)-1700—monitored by the Directorate of Operational Requirements and Development Plans, Deputy Chief of Staff, Research and Development, Hq USAF. Views or conclusions contained in this Memorandum should not be interpreted as representing the official opinion or policy of the United States Air Force.

The RAND Corporation
1700 MAIN ST • SANTA MONICA • CALIFORNIA • 90406

PREFACE

Automatic language data processing has been chiefly concerned with the analysis of written texts at, or below, the level of the sentence. The present study has a somewhat broader orientation. Rather than the individual sentence, a pair of adjacent sentences is taken as the object of analysis. Operating on the assumption that in many sentence pairs the second sentence is in some degree a continuation and development of the first, we have tried to identify and quantify some of the relationships obtaining between the members of the pair. The approach is empiric: We observe the characteristics of sentence pairs in a sample of written text. The long-range goal of this research is a better understanding of the ways in which sentences are strung together in coherent discourse.

SUMMARY

Portions of the Russian physics text processed at The RAND Corporation were subjected to systematic analysis to determine the extent of repetition in adjacent sentences. Recurrence of words in all pairs of contiguous sentences in the text (2,467 pairs) was recorded in a machine printout; sentence pairs for which word recurrence was not automatically recorded were visually inspected for other types of recurrence (through lexical stems, pronouns, synonyms, and paraphrases). The extent of the different types of recurrence is reported, and features of the recurring items are discussed. Sentence pairs characterized by nonrecurrence (12 percent of the total) are examined, and the relevance of inter-sentence recurrence and nonrecurrence to automatic syntactic analysis and abstracting is suggested.

STUDIES IN INTER-SENTENCE CONNECTION

1. INTRODUCTION

On the simplest level, a written composition may be regarded as an ordered sequence of elements (clauses, sentences, paragraphs, chapters, etc.). The order in which these elements are combined is often extraordinarily complex, as dictated by the author's fancy, the uneven progression of his thought, or his skill as a writer. Nevertheless, we may say that on some level each of the building blocks of ordered discourse should be well formed within itself, and should be arranged in a recognizable pattern with respect to other blocks of the same order. For example, an ill-formed (meaningless) paragraph can easily be constructed from a series of well-formed sentences. We note also that the relationship of a given block to neighboring blocks is a matter of importance. It is not enough that a sentence relate in a significant way to "some other" sentence in the discourse; its place in the pattern is normally determined by its relationship to nearby sentences.

We propose to adopt this simplified model of composition to study the inter-connection of sentences in written texts. If sentences are the constituent elements of paragraphs, in the ideal case each sentence will possess two attributes: continuity (the same subject matter is carried forward from preceding sentences) and development (new matter is added). We are concerned here only with the attri-

bute of continuity. According to the model, a given sentence finds its place in the structure of a paragraph by virtue of carrying forward subject matter from preceding sentences; a degree of sameness is, therefore, predictable. Further, if our previous remark about the interrelation of nearby blocks is valid, we should expect to find a considerable degree of sameness in adjacent sentences. The question then arises: In written texts, which are presumed to possess the attribute of continuity, is a degree of sameness observable between contiguous sentences?

How can the sameness of two sentences be measured? An author has innumerable ways of carrying forth the thread of his discourse. The most common of these, however, must certainly be the device of repetition: Lexical items in the first sentence are repeated in the second. On the level of the primer, the pattern may be as simple as

s. 1:	A is B.		s. 1: A is B.
s. 2:	A is C.	or	s. 2: B is C.
s. 3:	A is D.		s. 3: C is D.

On a more advanced level, writers try to avoid the appearance of being repetitious and are motivated to use other devices for achieving continuity. Under these circumstances, and in view of the unequal status of sentences as units of exposition, lexical repetition ould appear to be an inadequate measurement of the sameness of two sentences. Nonetheless, because the criteria for measurement are objective, we

propose to use this standard as a means of attacking the larger problem of inter-sentence connection.

The importance of recurring words in sentence pairs may be illustrated by an example. Consider the first two sentences of Lincoln's "Gettysburg Address."

> Fourscore and seven years ago our fathers brought forth on this continent a new nation, conceived in liberty, and dedicated to the proposition that all men are created equal.
>
> Now we are engaged in a great civil war, testing whether that nation, or any nation so conceived and so dedicated, can long endure.

Some of the recurring items in the two sentences are

Sentence 1	Sentence 2
(Fourscore and seven years) ago	Now
our (fathers)	we
(a new) nation	(that) nation
conceived (in liberty)	(so) conceived
dedicated (to the proposition...)	(so) dedicated

This comparison reveals five instances of recurrence: an adverbial time expression, a pronominal repetition, and three lexical items. On other levels we can observe more important "echoes," because the sentences are deliberately built on the principle of contrast. Among these, we may note that: (1) there is great similarity in syntactic structure in the two sentences, complemented, however, by the difference in tense ("brought forth," "conceived," etc. are contrasted to "are engaged" and "testing");

(2) the relatively short time span of "fourscore and seven years" in the first sentence is contrasted to "long endure" in the second; (3) "our fathers" were doing something constructive, whereas "we," a later generation, are self-destructive. Such considerations convince us that there is an enormous interplay between continuity on the one hand (parallel structures, repeated words, the fact that the subjects of the two sentences are genetically related) and development on the other hand (time has passed, and "we" are behaving differently). It must be true that the reader or listener is so absorbed in the comprehension of these complexities (i.e., in following the thought, or even in savoring the style), that the fact of word recurrence goes unnoticed. These lexical recurrences may be regarded, however, as some lower level tactical device available to the composer of paragraphs. We propose to determine the extent to which this device is actually employed in written texts.

At this point, it may be interesting to cite a small-scale experiment with a parlor game called "Paragraph." The pieces making up the game are not letters, as in "Scrabble," but sentences; the objective is to construct well-formed paragraphs, rather than words, from the pieces. The player has the task of unscrambling a given number of sentences (say, twelve) that originally comprised a given number of paragraphs (say, two). (The original paragraphs are chosen from written texts, or composed, so that a unique

correct ordering exists.) The first task in the game is to sort the sentences into the proper paragraphs--an easy task if the subject matter of the two paragraphs is different. The task of ordering the sentences within each paragraph is more difficult, especially because the player has no knowledge of the larger context in which the paragraphs were embedded. Experiments suggest that the game is playable, as correct solutions are given by most players. A brief analysis of the solutions showed that the correct sequencing of sentences within a paragraph was highly correlated with the repetition of words (or root morphemes) in a potential sentence pair.

To what extent do writers, consciously or unconsciously, use the rather simpleminded device of word repetition in the process of joining together the sentences of a paragraph? The following is an attempt to quantify this compositional tactic (which we shall hereafter call "recurrence"). (Beyond the fact that we are dealing with inter-sentence relationships, out study has little connection with Zelig Harris' work in discourse analysis.[1]) In this study, the body of Russian physics text processed at The RAND Corporation was subjected to analysis by a computer.[2] The text, contained on magnetic tape, is subject to automatic search at very high speeds; the tedious task of comparing the words in sentence pairs over a span of more than 400 pages of text was accomplished in approximately thirteen minutes.

2. PROCEDURE

Our text sample consisted of fifty-six different articles, by as many writers; a total of 2,467 sentence pairs (47,000 running words in all) were examined. The machine procedure resulted in a comparison of identification numbers* of all nouns and adjectives in every pair of adjacent sentences within each paragraph, i.e., the first and second sentences, the second and third, etc. For a portion of the text (Corpus 3, Table 1), recurrence across paragraphs was also recorded (i.e., the last sentence of each paragraph was also compared with the first sentence of the following paragraph, in addition to the comparisons described above). For each pair of sentences, an automatic matching routine was applied, resulting in a printout of the nouns and adjectives found in both members of a sentence pair. The machine search was limited to nouns and adjectives, since we knew from hand sampling that recurrence of verbs is negligible; the recurrence of adverbs, prepositions, and the like was not considered interesting.

The data for intra-paragraph recurrence is given in Table 1, which shows the number of sentence pairs characterized by one (and only one) of the following:

1. Noun/adjective recurrence, i.e., recurrence of at least one noun or adjective, as recorded in the machine search.

---

*That is, serial numbers of words in the glossary of the text.

2. "Other" recurrence, consisting of recurrences not recognized in the machine search. Data for this kind of recurrence was obtained by visual inspection of sentence pairs for which noun/adjective recurrence was not recorded during the machine search. We use the term "other" to signify recurrence of any one of the following types: root morphemes (e.g., "compute" and "computation"), technical symbols (not distinguished in the keypunching process), pronouns, and synonyms. The first two of these types are potentially machine recognizable, since they are graphic in nature; the latter types are not presently machine recognizable. The relative frequencies of these recurrence types were not distinguished; the data are presented merely as an indication of the extent to which "other" recurrence replaces noun/adjective recurrence.

3. Nonrecurrence.

Table 1

RECURRENCE, BY SENTENCE PAIRS

Type	Corpus 2	Corpus 6	Corpus D	Corpus 3	Total
Noun/adjective recurrence	276	216	448	832	1772 (72%)
"Other" recurrence	57	97	100	254	394 (16%)
Non-recurrence	38	34	86	158	301 (12%)
Total	371	347	634	1115	2467

In addition to the data summarized in Table 1, the following facts may be noted: (1) The average length of all sentences was twenty words; on the average, eleven of these twenty words were nouns and two were adjectives. (2) The volume of recurrence is in no way reflected in Table 1. On the average, slightly less than two nouns or adjectives recurred in each sentence pair; a large number of sentence pairs were characterized by the recurrence of three or more words. We note also that a count was not kept of the supplementation of noun/adjective recurrence by "other" recurrence.

3. DISCUSSION OF RESULTS

3.1. Recurrence of nouns or adjectives is found in approximately 70 percent of all sentence pairs. (Recurrence of nouns alone was observed in 62 percent of all pairs.) The extent of recurrence, as reflected in this data, is considerably greater than the author would have predicted. If to noun/adjective recurrence is added recurrence through root morphemes and technical symbols (not detailed in Table 1), the data show that approximately 80 percent of all sentence pairs are characterized by machine-recognizable recurrence. It is estimated, on the average, that more than two items of these types recur in every sentence pair.

Other studies will be necessary before it can be determined that the observed degree of graphic recurrence is a stylistic peculiarity of scientific text, or of Russian scien-

tific text. We note merely that a visual examination of ten pages of English text (on the rise of business corporations in the United States) revealed approximately the same extent of recurrence. It seems likely that recurrence through synonymy, paraphrase, and the like will increase in nontechnical texts. The language employed does not appear to be a significant factor in recurrence.

3.2. Of the 509 sentence pairs not characterized by graphic recurrence (20 percent of the total), 208 were characterized by recurrence through pronouns or synonyms. Recurrence through pronouns was rather restricted in our text because of the low incidence of personal names. As noted above, the limited use of synonyms is probably characteristic of texts in which the authors are not seeking stylistic variety.

3.3. Nonrecurrence is observed in approximately 12 percent of all sentence pairs. The most common situations for nonrecurrence were the following.

1. The first sentence of a paragraph referred to a diagram or table accompanying the text; succeeding sentences specified, in turn, the different elements in this point of reference. In such instances, the paragraph becomes a kind of single long sentence in which the elements of a set are conjoined; point-to-point development (progression) is suspended. Paragraphs of this type are not unusual, and it seems potentially useful to be able to identify strings

of sentences characterized by this kind of nonrecurrence. The researcher in automatic abstracting or information retrieval should take advantage of the fact that such strings have a different informational content than ordinary strings.

2. Continuity and development were achieved in a given sentence through word association. Here, the connection between sentences is supplied by the reader, on the basis of a larger context; the reader's background knowledge, in both a broad and narrow (technical) sense, is called into play, as is his ability to recognize analogy, generalization, and the like. In effect, a special kind of recurrence can be observed here, but since it is not achieved by repetition of graphemes or by use of synonyms, we have called it nonrecurrence. For example, the following word pairs possess different kinds and degrees of association, which can be recognized by the initiated reader but not by the literally-minded computer:

Sentence 1	Sentence 2
fil'm (film)	izobraženie (image)
naprjaženie (voltage)	vol't (volt)
veščestvo (substance)	rastvor (solution)
fotografija (photography)	fokusirovka (focusing)
éffekt (effect)	javlenie (phenomenon)
svet (light)	diffrakcija (diffraction)
ob"jasnit' (to explain)	zaključenie (conclusion)

Without attempting to categorize these and other kinds of word association, we note the following: (a) associated word pairs contribute enormously to inter-sentence coherence, both as a complement to and as a replacement of graphic recurrence; (b) sentence pairs provide a good source of data for establishing word associations (thesauric word clusters). For the researcher concerned with the problem of establishing networks of associated words, the sentence pair provides a special, restricted instance of co-occurrence.

3.4. Inter-paragraph recurrence is as common as intra-paragraph recurrence (cf. the data for Corpus 3 with the data for other corpora in Table 1). This finding, which contradicted our expectations, suggests either (a) that the transition from paragraph to paragraph is not unlike the transition from sentence to sentence, or (b) that the division of these texts into paragraphs was performed in an arbitrary fashion. We have no grounds for asserting that our texts are "compositionally" deficient; the problem needs further study.

3.5. Recurrence is not merely the effect of the greater frequency of words in an article. From the statistical point of view, there are two types of recurring words: those that are frequent in the article, and those that are infrequent. The latter should recur rarely, assuming random distribution of the words of the article. The matter appears to be more complicated, because of the

"local" phenomenon of recurrence: On the average, two of the thirteen nouns and adjectives of a given sentence will be repeated in the following sentence. This fact appears to have more bearing on recurrence than the overall frequency of a given word in an article.* We conclude that recurrence of infrequent words should provide "local context" cues for the researcher in information retrieval.

3.6. The relative position of recurring words in the structure of each sentence in a pair is notably similar. "Relative position" in this context refers to the level of a word in a sentence or clause, i.e., the number of nodes distant from the clause head (the predicate) in a dependency analysis. Thus, a noun subject or object of a verb in Sentence 1 was likely to become the subject or object of the verb in Sentence 2; if a noun was deeply buried in the dependency tree of Sentence 1, it was likely to be equally buried in Sentence 2.

This conclusion is based on a sample of 172 sentence pairs, for which the clause level of recurring nouns was determined by visual inspection. (No distinction was made between "independent" and "dependent" clauses.) The results are given in Table 2. The figures show the

---

*A sample was taken of the nouns and adjectives recurring in two articles, each about forty sentences in length. It was found that 61 different words recurred, and that 23 of these occurred in the respective article four times or less.

number of times that recurring nouns appearing at a given clause level in Sentence 1 appeared at a given level in Sentence 2. Thus, there were 22 instances in which the recurring word appeared at clause level one in the first sentence (i) and at clause level two in the second sentence (j).

Table 2

RECURRING NOUNS BY CLAUSE LEVEL

Sentence i \ Sentence j	Level 1	Level 2	Level 3	Levels 4-7	Total
Level 1	27	22	7	5	61
Level 2	17	16	9	8	50
Level 3	5	13	9	2	29
Levels 4-7	3	10	4	15	32
Total	52	61	29	30	172

Table 2 suggests the following observations:

1. A majority of recurring words appeared at clause level one or two in both sentences of a pair.

2. There is a strong tendency for recurring words to appear at the same clause level in the two sentences. In one-third of all instances, the difference in level was zero; in 70 percent of all instances, the difference in level was one or zero. A chi-square test of association was applied, yielding a sample value (37.206) considerably

larger than the upper 1 percent point of the distribution with nine degrees of freedom. In other words, if tables of this kind were constructed at random with the row and column totals shown in Table 2, less than one in one hundred would show as great a degree of association as is shown in Table 2. This association is principally revealed in the corners of the table: The upper left and lower right corners contain larger numbers than would be expected by chance, and the upper right and lower left contain fewer.

The fact that recurring words often tend to appear at the same clause levels was a surprise, because it is perfectly easy to construct sentence pairs on a radically different pattern. (We note also that if simple grammatical transformations had been applied, the relative position of recurring words would have been more nearly identical.) Further investigation of this phenomenon is needed. The relative position of different types of recurring elements should be studied. A possible application can be foreseen in the problem of identifying extra-sentence antecedents of pronouns. Our data suggest that the search for the antecedent of a pronoun should begin with nouns whose clause level is approximately the same as the pronoun's.

3.7. The implications of recurrence for automatic abstracting or information retrieval should be studied. We have noted above that passages characterized by non-

recurrence might receive special handling by abstractors (3.3); that the information content of a paragraph might vary according to the frequency with which recurring words occur in an article (3.5); and that the clause level of recurring words may be a useful clue in determining the antecedents of pronouns (3.6). It has been suggested that only the words high on a dependency tree of a sentence be selected for purposes of automatic indexing and abstracting;[3] we have found (3.6) that most of the recurring words in sentence pairs would be preserved if such syntactic pruning were employed. In general, however, we have not established that recurring words are important for information retrieval purposes. Our findings suggest that the phenomenon of recurrence is (a) frequent, (b) machine recognizable to a very large degree, and (c) of potential use to researchers in automatic information processing.

3.8. Discourse synthesis should rely heavily on the principle of recurrence. Although the generation of paragraphs remains a distant goal of automatic abstracting, we may predict that recurrence will be a necessary (although not sufficient) condition for joining sentences together. The principle appears to be valid, if for no other reason than that humans observe it in such a pronounced way.

REFERENCES

1. Harris, Zelig, "Discourse Analysis," Language, Vol. 38, 1952, pp. 1-30.

2. Harper, K. E., D. G. Hays and B. J. Scott, Studies in Machine Translation: Bibliography of Russian Scientific Articles, The RAND Corporation, RM-3610-PR (AD 410445), June 1963.

3. Doyle, Lauren B., The Microstatistics of Text, System Development Corporation, SP-1083, February 1963.

DOCUMENT CONTROL DATA

1. ORIGINATING ACTIVITY	2a. REPORT SECURITY CLASSIFICATION
THE RAND CORPORATION	UNCLASSIFIED
	2b. GROUP

3. REPORT TITLE

STUDIES IN INTER-SENTENCE CONNECTION

4. AUTHOR(S) (Last name, first name, initial)

Harper, Kenneth E.

5. REPORT DATE	6a. TOTAL NO. OF PAGES	6b. NO. OF REFS.
December 1965	21	3

7. CONTRACT or GRANT NO.	8. ORIGINATOR'S REPORT NO.
AF 49(638)-1700	RM-4828-PR

9a. AVAILABILITY/LIMITATION NOTICES	9b. SPONSORING AGENCY
	United States Air Force Project RAND

10. ABSTRACT

A systematic analysis of portions of the Russian physics text processed at RAND, to determine the extent of repetition in adjacent sentences. Recurrence of words in all pairs of contiguous sentences in the text (2467 pairs) was recorded in a machine printout; sentence pairs for which word recurrence was not automatically recorded were visually inspected for other types of recurrence (through lexical stems, pronouns, synonyms, and paraphrases). The extent of the different types of recurrence is reported, and features of the recurring items are discussed. Sentence pairs characterized by nonrecurrence (12 percent of the total) are examined, and the relevance of inter-sentence recurrence and nonrecurrence to automatic syntactic analysis and abstracting is suggested.

11. KEY WORDS

Machine translation
Russian language
Computer programs
Language
Linguistics