SP-2769

An Analysis of English Discourse Structure, With Particular Attention to Anaphoric Relationships

John C. Olney Dave L. Londe

10 November 1967

13

SP *a professional paper*

An Analysis of English Discourse Structure, With Particular Attention to Anaphoric Relationships

John C. Olney
Dave L. Londe

10 November 1967

SYSTEM
DEVELOPMENT
CORPORATION
2500 COLORADO AVE.
SANTA MONICA
CALIFORNIA
90406

SDC

ABSTRACT*

A brief account is given of a program system that attempts to recognize relaticnships constitutive of English discourse structure on the basis of syntactic and morphemic parallels. Results of operating the system on a sample text are presented. Attention is focused on the nature of discourse relationships, and more particularly on the problem of how to interpret discourse equivalences obtained by Harris's techniques.

*This paper was vrepared for presentation at the Tenth International Congress of Linguists, August 28-September 2, 1967, Bucharest, Romania

An Analysis of English Discourse Structure, With Particular Attention to Anaphoric Relationships*

John C. Olney
Dave L. Londe

In analyzing discourse structure, a revealing question to ask is: What conditions must be satisfied by a sequence of sentences, none of which is syntactically or semantically anomalous in a given language, if the sequence is to constitute a non-anomalous discourse in that language?** As Bever and Ross have pointed out ([1], p. 3), the sequence must exhibit some kind of connectedness, though clearly that is not a sufficient condition. A more general condition is that the sequence as a whole should be structured, or as we shall say, that the sentences constituting it should jointly exhibit thematic development. Whenever this more general condition is satisfied by a sequence, the sequence also exhibits connectedness. Bever and Ross suggest in effect that the task of giving a full specification of connectedness as a condition on non-anomalous discourses goes beyond linguistics, insofar as it would involve specifying concept and belief structures taken for granted by the authors of discourses. In our view the task of giving a full specification of thematic development also goes beyond linguistics. Nevertheless, we believe that useful, albeit partial, specifications of these notions can eventually be given in linguistic terms.

The principal objective of our research is to enable language processing programs to relate appropriately expressions occurring in different sentences of the same discourse, so that a more accurate representation of the intended information content of that discourse can be obtained. We have confined our attention to English discourses, and more particularly to carefully edited scientific articles, for reasons given elsewhere [7].

Initially we concerned ourselves exclusively with anaphoric relationships, which are the most prevalent manifestations of connectedness in the texts we have studied. One reason for focusing on anaphoric relationships is that, of all the important relationships constitutive of discourse structure, they are the easiest to analyze. That is, for a given discourse there is less uncertainty about exactly which expressions stand in anaphoric relationships than about

*The work reported herein was supported by SDC and Grant 1-R01-LM-00065-01, English Discourse Structure, for the Public Health Service, U.S. Department of Health, Education, and Welfare.

**A slightly different formulation of this question is given in [2].

any other frequently manifested discourse relationship. (but see Note 2 in Figure 2). Subsequently we broadened our concern to include non-anaphoric discourse relationships, in part because it became apparent that mechanical recognition of the latter can often contribute significantly to the mechanical recognition of anaphoric relationships.

At present we have a program consisting of three routines for recognizing discourse relationships: PAIRS, ANAPH, and DISCO. Since PAIRS has been described elsewhere [7], it will suffice here to point out that the syntactic information which it outputs for each text sentence is essentially a surface-structure description derived from the output of the Kuno Syntactic Analyzer [5]. In the PAIRS output, constituents (e.g., 'prepositional phrase'), functional relations (e.g., 'subject'), and dependency relations are marked. A few relationships belong to the deep structure are also marked; i.e., the subjects of active infinitives and participles are reconstructed, as are the objects of passive infinitives and participles.

The syntactic information output by PAIRS is input to ANAPH and DISCO. Both accept two other inputs: the output of a suffixal analysis routine for words in the text and special dictionary information relating to pronoun replacement. For most expressions in the text that it recognizes as anaphoric, ANAPH outputs a list of 0 or more potential antecedents; for most antecedents an evaluation score is given, and the components of the score are indicated. Operating independently, DISCO outputs a list of pairs of expressions recognized as discourse equivalent on the basis of morphemic and syntactic parallels. Each equivalence is given an immediacy score (following Harris); portions of DISCO that have not yet been checked out also score each equivalence for various factors that may be said to confirm the equivalence. (One such factor would be that the equivalence reflects several independent parallels--see also [8], p. 7.) It is planned that certain high-scoring outputs from ANAPH and DISCO will be exchanged and the programs recycled.

Portions of the output of ANAPH and DISCO for the text reproduced in Figure 1 are shown in Figure 3. Comparison of the output of ANAPH with the intuitive anaphoric analysis given in Figure 2 reveals that ANAPH recognizes only about two-thirds of the anaphoric expressions while recognizing as anaphoric five expressions that are not. It found the correct antecedent*

*By 'correct antecedent' here we mean a preceding expression which either coincides with or is part of the expression for which a given anaphoric expression substitutes (see Figure 2). Admittedly, it is an oversimplification to speak of an anaphoric expression as a substitute. Essentially, an expression is anaphoric if: 1) it is understood in context in a sense more specific than can be obtained as a compositional function of its lexical sense(s) and those of words directly related to it syntactically (surface structure); and 2) its specialized sense can be obtained at least in part as a compositional function of the senses just referred to and the sense of some preceding expression (its correct antecedent).

SOCIAL DEPRIVATION IN MONKEYS[*]
(opening paragraph)

```
1  2   3      4        5            7          10       11  12    13
In AN OUTLINE OF PSYCHOANALYSIS, published posthumously in 1940   ,

  14     15   16  17   20  21   22 23   24       25      26  27   30   31
Sigmund Freud was able to refer to the common assertion that the child is

     32         33   34    35 36  37  40   41  42   43    44 45   46
psychologically the father of the man and that the events of his first

 47   50  51    52         53     54  55   56      57       60   61 62   63
years are of paramount importance for his whole subsequent life. It was  ,

64  65      66    67    70   71           72         73  74   75 76
of course   ,   Freud's own historic investigations  ,  begun a half-century

  77   100   101  102     103      104 105  106  107         110     111
before   ,   that first elucidated the role of infantile experiences in

112     113       114  115       116      117  120    121      122   123
the  development  of   the   personality  and  its disorders.  The central

  124      125   126  127    130   131     132 133  134  135  136  137    140
experience  of   this period  of childhood   ,  he  found   ,  is  the  infant's

  141     142  143   144        145    146  147  150  151   152   153   154 155
relation   to  his   mother.   Freud's ideas have now shaped the thinking of two

   156      157         160       161      162        163       164          165
generations of psychologists  ,  psychiatrists  and  psychoanalysts.  Much

   166 167  170    171 172  173    174      175  176   177         200
evidence  in support  of  his  deep insights have been accumulated  ,

    201        202    203       204     205 206   207     210     211         212
particularly  from clinical studies  of  the  mentally ill.  Contemporary writers

 213     214       215    216         217     220 221 222     223    224 225
stress inadequate  or inconsistent mothering as  a  basic  cause  of later

  226     227  230  231       232  233      234  235    236  237       240
disorders  such  as  withdrawal  ,  hostility  ,  anxiety  , sexual maladjustment

241   242        243  244  245          246     247        250     251
 ,   alcoholism  and  ,  significantly  ,  inadequate  maternal behavior!
```

[*] Scientific American, November 1962, Vol. 207, No. 5, pp. 136-146

Figure 1. Sample Text Processed by PAIRS, ANAPH and DISCO

Anaphoric Expression		Expression for Which it Substitutes
his	45	(27-30)
his	55	45
subsequent	57	subsequent to 45-47
Freud's	67	(14-15)
before	77	before 12
that	101	67-72
its	120	(115-116)
this period	126-127	45-47
he	133	(67)
his	143	140-141
Freud's	145	(133)
ideas*	146	ideas, specifically with regard to 104-121, and perhaps (122-144)
his	172	145
insights*	174	insights, inter alia, 104-121
later	225	57
disorders*	226	120-121

Explanation of Intuitive Anaphoric Analysis

1) The number(s) in parentheses immediately to the right of each entry in the left-hand column indicates the ordinal position of the word(s) held to constitute an anaphoric expression and correspond to the octal numbers written over the words (in the text, as reproduced in Figure 1. The number(s) in the right-hand column are enclosed in parentheses if the expression(s) so designated cannot replace the whole anaphoric expression unless further changes are made in the sentence containing the anaphoric expression. In other words, when the sequence numbers are enclosed in parentheses, the expression so designated suggests but is not identical to the expression for which the anaphoric expression substitutes. An anaphoric expression referred to in the right-hand column should be understood as representing the expression for which it substitutes directly or indirectly. When a portion of the expression for which an anaphoric expression substitutes is not directly suggested by a preceding expression of the discourse, that portion is written out on the right-hand column.

2) Entries in the left-hand column that are marked with an asterisk are not considered to be clear cases of anaphora. Exactly which expressions in a text are anaphoric and exactly what they substitute for cannot be determined unless the intended information content of the text is specified (which in turn presupposes that the background knowledge assumed by the author has been specified) and the portion of this information content which each sentence in the text is to carry has been decided upon.

Figure 2. Intuitive Anaphoric Analysis of the Opening Paragraph of Social Deprivation in Monkeys

Syntactic Head of Each Expression Recognized as Anaphoric	Syntactic Head of Each Expression Recognized as an Antedecent Candidate[1]	Equivalences[2] Obtained by Morphemic Parallels	Equivalences[2] Derived From Equivalent Governors

child (30)
man (37)
His (45)
his (55)
	man (37:4), father (34:3), child (30:3), Freud (15:3)	psychoanalysis (5)$=_{.5}$ psychoanalysts (144);
		child (30) $=_{.5}$ childhood (131); disorders (121)$=_{.5}$
			sexual (237)$=_{1.5}$ inadequate;
			infantile (107)$=_{1.5}$ central (123);
			sexual (237)$=_{1.5}$

years (47)
subsequent (57)
that (101)
	investigations (72), course (65)	disorders (226);
		experiences (110)$=_{.5}$ experience (124);
		infantile (107)$=_{.5}$
			maternal (250);
			inadequate (214) $=_{2.5}$ maternal (250)

personality (116)
its (120)
	personality (116:5), development (113:3), role (105:4), course (65:3), life (60:1) ...	infant's (140); mother (144)$=_{.5}$ mothering (217);
		inadequate (214)$=_{.5}$ inadequate (247)
			Equivalences[2] Obtained by Coordination
			that (26)$=_{1}$ that (41);
			personality (116)=1 disorders (121);

period (127) he (133) his (143) mother (144) his (172)	Freud's (67:7) Freud's (145:2), infant's (140:1)	Equivalences[2] Obtained via Predicate Nominative child$=_{1}$ father (34); it (61)$=_{1}$ investigations (72); experience 124$=_{1}$ relation (141)	psychologists (160)$=_{1}$ psychiatrists (162)$=_{1}$ psychoanalysts (164); withdrawal (231)$=_{1}$ hostility (233)$=_{1}$ anxiety (235)=1
insights (174) later (225)	Equivalences[2] Derived from Equivalent Dependents mothering (217)$=_{1.5}$ behavior (251)		maladjustment (240)$=_{1}$ alconolism (242)$=_{1}$ behavior (251)

Figure 3. Portions of the Output of ANAPH and DISCO

(1) The antecedent candidates found by ANAPH for the non-pronominal expressions and their scores have not been shown because these results are still too inaccurate to be worth presenting. No score is given for 'that' (101), because ANAPH does not yet have rules for scoring antecedent candidates of relative pronouns. The components of the scores given for other pronouns (separated from their respective sequence numbers by a colon) are not shown because of space limitations.

(2) DISCO has rules which derive additional equivalences, as follows if A$=_{n}$ B and B$=_{m}$ C, where n and m are immediacy scores (see p.4), then A $=_{n+m}$ C, provided certain constraints are satisfied; if A$=_{n}$B, where A and B are single words, then C$=_{n}$ D, where C and D are the longest phrases of which A and B, respectively, are the syntactic head. Owing to space limitations, none of these equivalences are shown here.

*The brace indicates that ANAPH has grouped these words together on the assumption that they have the same antecedent, an assumption which of course is false in the case of 133 and 143.

for one-third of the anaphoric expressions but did not always give the highest score to that antecedent. It should be pointed out that, for other text samples, our rules for recognizing anaphoric expressions and for preferring the correct antecedents of personal pronouns work at about the 80% level. By making simple, relatively non-ad hoc refinements to ANAPH we could get it to recognize all the anaphoric expressions in the text reproduced in Figure 1 except for 'disorders' (226). However, no comparably simple refinements would enable us to avoid picking up the short noun phrases introduced by generic 'the' and 'his.' (In texts we have examined, about half of the short noun phrases introduced by 'his' prove to be anaphoric.) What we should perhaps do is work out a scoring system for recognizing anaphoric expressions as well as for preferring their correct antecedents.

The output of DISCO is rather difficult to evaluate. Many of the rules used by this program are closely similar to rules used by Zellig Harris in producing discourse analyses of texts. There are some significant differences between our rules and his; for example we use phrasal environments as well as sentential environments and dispense with his notion of 'optimal transform.'* Even if our rules were not significantly different from Harris's, evaluation of the equivalences would still be difficult because little guidance is afforded either by Harris's specific criteria (summarized in [2], p. 69) or by his general criterion: "...the classes [of discourse-equivalent expressions] set up ... [should be] such that their regularity of occurrence will correspond to some relevant semantic interpretation for the discourse." ([4], p. 7). Presumably this latter criterion should be interpreted in the light of Harris's rather unclear claim that the correlation of formal features of discourses analyzable by distributional methods within each text "...with a particular type of situation [i.e., the one in which the discourse occurred] gives a meaning-status to the occurrence of these formal features"[3], p. 3). Harris can hardly have intended that discourse equivalent expressions must also be semantically equivalent, since it is often just in those parts of a text where the author is being most informative that expressions which are not semantically equivalent will be asserted to be equivalent (e.g., in Figure 1, 122-131 and 137-144) or will appear in a parallel construction (e.g., 235 and 247-251). We have provisionally adopted the following criterion, which is admittedly vague (but see below): The procedures for establishing discourse equivalences should yield a grouping of the expressions in a connected discourse that will approximately correspond to a grouping of the entities they denote or express, such that those entities closely related in the conceptual framework of the discourse are grouped together.

*We agree with Bierwisch's criticism of this notion, as given in [2].

Our criterion has the effect of allowing a pair of expressions which stand in an anaphoric relationship to be discourse equivalent.[*] We shall say that such equivalences have an anaphoric interpretation, and it is by working out comparably specific interpretations for other kinds of discourse equivalences that we shall attempt to provide a more useful criterion for deciding when a given pair of expressions should be accepted as discourse equivalent. Recently, composition teachers have begun to give detailed descriptions of the various roles a sentence can play with a paragraph--e.g., to define. particularize, concede, support, contract, etc. (see [6].) From the writer's perspective, each of these roles may be viewed as a particular method of thematic development. By analyzing a corpus of scientific articles, we have tentatively identified about 50 methods of thematic development. One of these methods (which we have called 'limitation') is exemplified by several pairs of expressions in the text reproduced in Figure 1, viz., 122-144 is a limitation of 61-121, and 122-131 is a specific limitation of 107-110: 211-251 is a limitation of 137-144 and 145-164, and 211-212 and 214-217 are specific limitations of 155-164 and 137-144, respectively; and 201-210 is a limitation of 165-177. We suggest that the equivalences $110 =_{.5} 124$ and $144 =_{.5} 217$ should be interpreted as reflecting a relationship of limitation between the expressions specified above in which they occur. We hope ultimately to show that a given pair of expressions should be accepted as discourse equivalent just in case their equivalence can be interpreted as reflecting one or more instances of a specified set of generalized relationships that may be regarded as constitutive of connectedness and thematic development.

Finally, we would like to give an example of the potential usefulness of exchanging selected outputs from ANAPH and DISCO and recycling the programs. Words 45 and 55 were correctly grouped together by ANAPH as having the same anteecedent (see Figure 3). By our criterion of discourse equivalence these words are equivalent. From this equivalence DISCO can derive the equivalences $47 =_1 60$ and $45 - 47 =_1 55 - 60$. But this latter equivalence is exactly what ANAPH would need in order to find the correct antecedent of 'subsequent.' Admittedly, it is not clear that this equivalence can be formally recognized as having an anaphoric interpretation just on the grounds that it was derived from an anaphorically derived equivalence; semantic information showing the relationship between the appropriate senses of 'years' and 'life' may be required. (In this connection it is worth noting that ANAPH and DISCO jointly have a significant potential for syntactic and semantic disambiguation (cf. [8], p. 14).)

[*] Harris also treats such pairs of expressions as discourse equivalent (cf. [4], p. 23).

REFERENCES

[1] Bever, T. G., and Ross, J. R. Underlying Structures in Discourse. Proceedings of the Conference on Computer-Related Semantic Analysis Las Vegas, 1965.

[2] Bierwisch, M. Review of Discourse Analysis Reprints. Linguistics, April, 1965. 61-73.

[3] Harris, Z. Discourse Analysis. Language, 1952, 28, 1-30.

[4] Harris, Z. Discourse Analysis Reprints. The Hague: Mouton & Co., 1963.

[5] Kuno, S. Some Characteristics of the Multiple-Path Syntactic Analyzer. Language Data Processing. Cambridge, Mass.: The Computation Laboratory of Harvard University, 1964, pp. C-6-1--C-6-8.

[6] Larson, R. Sentences in Action: A Technique for Analyzing Paragraphs. College Composition and Communication, March 1967, 16-22.

[7] Londe, D., and Olney, J. PAIRS: A Postprocessor for the Kuno-Oettinger English Analyzer. Santa Monica: SDC document (in press).

[8] Olney, J., and Londe, D. A Research Plan for Investigating English Discourse Structure, with Particular Attention to Anaphoric Relationships. Santa Monica: System Development Corporation document, TM(L)-3256, November 22, 1966. (This document is available only with permission of the author.

DOCUMENT CONTROL DATA - R & D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1. ORIGINATING ACTIVITY *(Corporate author)*

System Development Corporation
Santa Monica, California

2a. REPORT SECURITY CLASSIFICATION
Unclassified

2b. GROUP

3. REPORT TITLE

An Analysis of English Discourse Structure, With Particular Attention to Anaphoric Relations

4. DESCRIPTIVE NOTES *(Type of report and inclusive dates)*

5. AUTHOR(S) *(First name, middle initial, last name)*

John C. Olney
Dave L. Londe

6. REPORT DATE
10 Nov 67

7a. TOTAL NO. OF PAGES
10

7b. NO. OF REFS
8

8a. CONTRACT OR GRANT NO.
Grant 1-R01-LM-00065-01, English Discourse Structure

b. PROJECT NO.

c.

d.

9a. ORIGINATOR'S REPORT NUMBER(S)
SP-2769

9b. OTHER REPORT NO(S) *(Any other numbers that may be assigned this report)*

10. DISTRIBUTION STATEMENT

Distribution of this document is unlimited.

11. SUPPLEMENTARY NOTES

12. SPONSORING MILITARY ACTIVITY

13. ABSTRACT

A brief account is given of a program system that attempts to recognize relationships constitutive of English discourse structure on the basis of syntactic and morphemic parallels. Results of operating the system on a sample text are presented. Attention is focused on the nature of discourse relationships, and more particularly on the problem of how to interpret discourse equivalences obtained by Harris's techniques.

14 KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
Anaphoric relationships						
Discourse structure						
Syntax						
Morphemic parallels						
Linguistics						