# COLLOCATIONAL GRAMMAR AS A MODEL FOR HUMAN-COMPUTER INTERACTION

W. Randolph Ford

Prism Associates
7402 York Road, Suite 301
Towson, Maryland 21204

Raoul N. Smith

GTE Laboratories, Inc.
40 Salvan Road
Waltham, Massachusetts 02254

Contrary to the long-held belief of transformational grammarians for communication in general, the majority of natural language sentences which people actually use in communicating with a computer, in an unconstrained mode, are not novel. As Thompson and Thompson 1981:660 observe: "monotony of structure is the rule rather than the exception in human-computer communication." Thompson 1981:41 reports in her study of such communications that 75 percent of the queries were wh-questions, 19 percent were commands, 5 percent were statements, and 1 percent were yes/no questions.

The repetitive feature of natural language is not a new concept. Similar observations have been made before. Damerau 1971 used collocation of lexical items as the basis for a Markov model in an experiment for text generation. Becker claims that "the wonderful feats of the human intellect... are based as much on memorization as on any impromptu problem-solving" (1975:62). He posits a phrasal lexicon consisting of six major categories of lexical phrases by which we "stitch together swatches of text that we have heard before; productive processes have the secondary role of adapting the old phrases to the new situations" (1975:60).

All of these approaches to natural language data rely heavily on the observation that many lexical items tend

co-occur. This surface co-occurrence is the result of what may at times be complicated syntactic and semantic interrelations of language units. Unfortunately, a systematic accounting of these interrelations has not been achieved in any linguistic theory. The thrust of our approach is that the more of language which can be handled lexically, the easier will human language be able to be modelled.

Actual data on the frequency of lexical collocation are very sparse. A study of word sequences in PANALOG text has shown a surprising amount of repetition of word sequences (Bienstock and Smith in preparation). (PANALOG is a system for passing messages among small groups in computer conferencing with telemail and calendar features. See Housman 1979. The data are of human-human communication and not human--computer communication and are thematically restricted. They therefore resemble Damerau's data.) A study of parts of the Brown English Corpus has been undertaken in order to get less thematically homogeneous material. In addition, Wizard of OZ experiments with unconstrained human-computer input will begin soon at GTE Laboratories in order to gather the more relevant human-computer data.

A PANALOG text of 16,133 words chosen for study. The longest string which occurred more than once was a seven word quote from Nietzsche. Two six-word strings occurred twice and at length five, one occurred three times and thirteen were repeated twice.

An interesting feature of these distributions is that the number of hapaxes (those strings occurring only once at a given string length) reaches a peak at length three (see Figure 1).

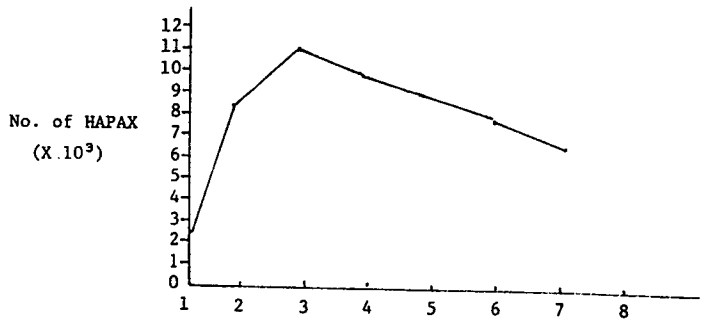


Figure 1.

This is a revealing measure of the amount of repetition in a text of this length. In particular, recurring two word strings account for 40.8 percent of the running text and recurring three word strings comprise 8.1 percent of the text.

The basic assumption of the frequent occurrence of lexical collocation in natural language texts, especially in human-computer communication, is the basis for the development of a new type of natural language processor. Ford, 1981, has constructed a natural language processing system for database updating, retrieval, and manipulation, which relies critically on the observation that real users tend to employ a very limited set of lexical string types in querying databases.

The Ford natural language processor consists of a two stage reduction algorithm for translating natural language inputs into basic functions which are then used to perform the query. The first stage of the reduction changes the input words to meaning representations using a list of lexical items and a meaning correlate list. The second stage takes as input strings of these meaning correlates and changes them into basic list.

409 numeric representations for words mapped down to 132 unique meanings and 1328 canonical sentence vectors mapped down

to 19 functions. This two stage reduction scheme worked efficiently enough to respond to 93.8 percent of the 1697 input queries, including ungrammatical ones from inexperienced users, with a response time of 1.5 seconds, operating in an environment of 90K 8-bit bytes. This compares very favorably to Thompson 1981 where only 67.7 percent of REL queries were correctly parsed with an average response time of 10 seconds. (Space requirements were not reported.) Similarly, Damerau 1981 and Petrick 1981 report a success rate for TQA of 65.1 percent inputs correctly parsed with the time required to process a sentence typically being 10 seconds.

The reason why the system works so well in terms of accuracy, speed, and small storage requirements is based on the two stage reduction technique which, in turn, is based on the fact that a great many inputs in human-computer communication are repetitious syntactically, semantically, and lexically. Repetition is a principal characteristic of human-computer communication.

REFERENCES

Becker, Joseph, 1975. "The Phrasal Lexicon," in Schank, Roger and Bonnie Nash-Webber, eds. Theoretical Issues in Natural Language Processing, ACL Workshop, Cambridge, MA, pp.38-41.

Bienstock, Daniel and Raoul N.Smith. In preparation. "Lexical Collocation in Three Types of Texts."

Damerau, Frederick J. 1971. Markov Models and Linguistic Theory. Janua linguarum series minor 95. Mouton: The Hague.

Damerau, Frederick J. 1981. "Operating Statistics for the Transformational Question Answering System", AJCL 7.1.:30-42

Ford, W. Randolph. 1981. "Natural-Language Processing by Computer - A New Approach", Unpublished Ph.D.dissertation, The Johns Hopkins University, Baltimore, Maryland.

Housman, Edward. "Computer Mediated Communication," Profile 3 (1979): 1-4.

Petrick, Stanley. 1981. "Field Testing the Transformational Question Answering (TQA) System," Proceedings of the Nineteenth Annual Meeting of the Association for Computational Linguistics, pp. 35-36.

Reiger, Chuck. 1977. "Viewing Parsing as Word Sense Discrimination," in Dingwall, William O. A Survey of Linguistic Science, Greylock Publishers.

Small, Steven. 1980. "Word Expert Parsing: A Theory of Distributed Word-Based Natural Language Understanding," University of Maryland Computer Science TR-954.

Thompson, Božena H. 1981. "Evaluation of Natural Language Interfaces to Data Base Systems," Proceedings of the Nineteenth Annual Meeting of the Association for Computational Linguistics. Menlo Park, CA: Association for Computational Linguistics, pp. 39-42.

Thompson, Božena H. and Frederick B. Thompson. 1981. "Shifting to a Higher Gear in a Natural Language System," Proceedings of the 1981 National Computer Conference Arlington, VA: AFIPS Press, pp. 657-662.