Preprint No. 62

COLLECTION OF ABSTRACTS OF PAPERS

INTERNATIONAL CONFERENCE

ON

$\overset{*}{\text{C}}\overset{*}{\text{O}}$MPUTATIONAL $\overset{*}{\text{L}}\overset{*}{\text{I}}\overset{*}{\text{N}}\overset{*}{\text{G}}$UISTICS

⇓

COLING

1969

KVAL
Forskningsgruppen
för
KVANTITATIV
LINGVISTIK

RESEARCH GROUP FOR QUANTITATIVE LINGUISTICS

Address: Fack Stockholm 40, SWEDEN

Abstract

AN APPLICATION OF COMPUTER TECHNIQUES TO ANALYSIS OF THE VERB PHRASE IN HINDI AND ENGLISH: A Preliminary Report

Dr. L.M. Khubchandani and W.W. Glover

Authors worked on the Project at Poona, India with the facilities of the computer CDC 3600-160A installed at the Tata Institute for Fundamental Research, Bombay.

The Project uses two sets of data: a corpus of verbal phrases drawn from a modern Hindi play and a completeparadigm of English sentences generated from the kernel"he eats it". The computer was programmed to group into classes the words occurring in identical contexts, and substitute in the data corpus for these words a reference to the class where they have been put. The classification and substitution thus produced suggested phrase patterns, with the filler class for each tagmeme defined as the class represented in the particular slot of the pattern.

The results obtained with a criterion for classification of "identical context one-deep on both sides" were quite satisfactory. In Hindi 25 classes were formed from the corpus of 65 phrases. Atleast one word was classified in each 37 (62%) of the phrases and all words were classified in 3 phrases (15%). With an increased sample of similar data these percentages would be expected to increase. In English 24 patterns were obtained and 15 classes were formed from the full paradigm of 112 sentences.

However, as some of the classes contained grammatically dissimilar members, the criterion was altered to "identical context two-deep on both sides". The results with this criterion appear less promising in Hindi. The data sample was extended to 248 phrases of three words or more. The machine discovered 223 patterns and 13 classes, and in only 29 patterns (13%) one word was replaced by a class reference. This criterion, however, enjoyed some success in analysing the English paradigm which is, of course, highly restricted data. With the full paradigm, the machine discovered 30 patterns and 19 classes. 18 of them are quite homogeneous in membership and the sentences generated by the patterns using these classes are all legitimate.

It is felt a useful basis for further investigation would be to refine the broad classes formed with a "one-deep" criterion in a subsequent run through the same or new data.