WORD, PHRASE AND SENTENCE

Rob't F. Simmons
Univ. of Texas, Austin

Among the relative verities of natural language processing are the facts that morphemes and words are primary semantic units, and that their co-ocurrence in phrases and sentences provides cues for selecting sense meanings. In this session, two psycholinguistic studies show some aspects of how human subjects process words while reading. A study of medical vocabulary shows that medical words are highly associated by co-occurrance in medical definitions. Another report shows the effectiveness of keyword identification and selection of prominent sentences to organize abstracts for retrieval.

A fifth study argues that analysis of existing natural language dictionaries can be expected to contribute importantly to what is needed for text understanding programs. The final study is an experiment with a sentence level translator applied to a large German-English translation task. These two studies are primarily concerned with analysis of language at the sentence level.

The most glamourous areas of natural language research are at levels above the sentence, concerned with dialogues and discourse, frequently disdainful of morphological or even grammatical analysis in their search for effective structures for understanding what the discourse is about. Scripts, frames, stereotypes, schemas are all studied in these areas; and often morphological and grammatical analysis is bypassed in favor of keyword scanning to extract some small relevant portion of the text to be bound as values for slots in these larger data forms.

This session reminds us that much can be accomplished with vocabulary analysis, with keyword scanning and statistical treatment of text and with semantic analysis at the single sentence level. Yet, with regard to most of the topics in this and other sessions, there is a strong sense of deja vu; the earliest natural language studies featured automatic extracting and information retrieval based on statistical, lexical and associational properties of keywords. Mechanical translation of sentences without regard for larger contexts marked the late sixties high point of MT research amid contemporaneous studies of the English dictionary and thesaurus. Competition among sentence parsing algorithms is an ACL tradition celebrated annually, while psycholinguistics has traditionally applied chronometric studies, and recordings of eye movements to measure this or that aspect of human linguistic processing throughout the period.

This is not to suggest that nothing new is happening; actually, the continued emphasis on these topics reveals that, though introduced early, they are still imperfectly understood. I believe science progresses in spirals; initial studies are accomplished and published supporting more advanced studies that build upon the findings of the earlier work. Superstructures of theory are constructed and more work is undertaken in this framework. Finally the initial studies are lost in years of accumulated literature, and perhaps some of the wildest theories begin to collapse. Then the field may suddenly show renewed interest in its beginnings and repeat its early studies with the added sophistication gained by experience. At this time the line of history spirals past the points it reached on earlier cycles. Hopefully, as in this session, the experience gained between cycles insures an upward progression rather than a profitless loop.