Collocation in CIIL Corpus: Implications for Language Teaching and NLP

Thennarasu, S.

1. Introduction

Collocations are a feature of natural languages that are not well addressed by current Tamil language teaching and current models used for NLP. Language is full of word combinations that occur more frequently than expected (Joachim Wagner, 2008). The scholar Graeme Kennedy has stated that Palmer (1933) is one of the most influential English language teaching specialists of the 20th century, who adopted the term *collocation* for recurring groups of words. He defined a collocation as "a succession of two or more words that must be learned as an integral whole and not pieced together from its component parts" (e.g. Tamil. *talaimai aluvalakam* 'chief office', *mālai vēlai* 'evening time'etc.). Palmer went so far as to suggest that even a "selection of common collocations... exceeds by far the popular estimate of the number of single words contained in an everyday vocabulary." The possibility that there are many more collocations to learn than there are words in a language perhaps helps explain why learning a language usually takes so long in comparison with other complex learning tasks.

Palmer's (1933) pioneering work on collocations in English language teaching was paralleled in different branches of the language sciences. Among a number of scholars who took account of the phenomenon of collocation, Firth (1957) emphasized the importance of both linguistic collocation and situational context for the description of languages in his maxim, "You shall know a word by the company it keeps".

Corpus-based evidence was used by Sinclair (1991) to support what he called the *idiom principle* in language learning and use (characterized by the use of routinized combinations of words in speech and writing), and to highlight the neglect of collocations in the theory and practice of English language teaching.