Corpus-Derived Profiles: A framework for studying word meaning in text
Gregory Garretson

Syntagmatic relations such as collocation, colligation, and semantic preference are increasingly seen as an important part of word meaning. Growing interest in corpus-based and computational studies of word meaning calls for a unified approach to these relations. This thesis offers three components which contribute significantly toward such an approach: (1) the Corpus-Derived Profiles (CDP) framework, in which syntagmatic relations are studied by profiling words in corpora; (2) the implementation CenDiPede, a program for performing studies using the framework; and (3) a series of empirical studies of English nouns carried out using the framework.

The goal of the CDP framework is to define, interrelate, and automate analysis of the syntagmatic relations collocation, colligation, and semantic preference. It has two components: A "lexical profile" is a data structure containing information about a given word's relations in a given corpus; the "CDP query language" is a system for extracting information from and comparing profiles, allowing comparisons both of different words in the same corpus and of the same word in different (sub-)corpora. The Java program CenDiPede, which enables one to create and query lexical profiles, is offered freely for research under an open-source license.

CenDiPede was used to perform the three studies presented, each examining a different semantic relation using syntagmatic information.
The first study uses collocational information to study synonymy, focusing on "sort", "kind", and "type". It is shown that these nouns, though largely synonymous, are used in different contexts, operating on a cline with "sort" at one end, "type" at the other, and "kind" in the middle. The second study uses colligational information to investigate polysemy in the noun "time"; it is found that examination of the grammatical context of a token frequently suffices to predict which of several senses it corresponds to. The third study uses semantic preference information to investigate antonymy, demonstrating that algorithms based on semantic preferences can both select a noun's antonym from a set of candidates and identify asymmetries between antonyms. Further, a new model of antonymy is proposed, one which explains the unusual nature of nominal antonymy as mismatch between concept types and syntactic classes.