

Christiane Fellbaum  
Princeton University

Abstract for "Processing Text-Technological Resources"  
(Recognition of Thematic Structures: Methods, Resources, and Applications)

### **Towards a better lexical resource for NLP**

Many Natural Language Processing applications use statistical methods but in addition avail themselves of lexical resources, which may reflect symbolic approaches and the linguistic intuitions of their creators.

One widely used, handcrafted, broad-coverage lexical resource is WordNet . Its principal shortcoming is the sparsity of the connections among its synsets and words, which limits the information that is available for a given concept and hampers word sense discrimination. In particular, WordNet has few syntagmatic links, and it is here that statistical analyses of corpora can provide complementary information.

Given WordNet's relational structure, which lends itself well to quantitative analyses of semantic similarity, we might wonder whether important connections that are intuitively obvious are missing. For example, WordNet has no way to link between members of such pairs as *Thanksgiving* and *turkey*, *dollar* and *green*, *chopstick* and *Chinese restaurant*. Purely statistical corpus analyses could find some, but not all such intuitively related pairs, and would moreover identify spurious ones.

We performed an experiment aimed at increasing WordNet's connectivity by identifying syntagmatic links among synsets in ways that did not introduce biases and limitations inherent to traditional, systematic, introspectively defined relations. We collected human ratings that reflect the associative strength of linguistically expressed concepts. First, we semi-automatically determined a set of 5,000 highly frequent and salient concepts from WordNet ("CoreWordNet"). Using a specially designed interface, students were presented with randomly selected pairs of CoreWordNet synsets and asked to rate the strength with which the concept expressed by the synset that was presented first *evokes* that expressed by the second synset. We thus obtained directed and weighted ratings of similarity for concept pairs. Compared the results with standard measures of semantic similarity, we found that our evocation method captures similarities that elude these measures.

The results are potentially beneficial for the construction of more powerful lexical resources. And importantly, they raise questions as to the nature of semantic relations, semantic similarity, and human conceptual organization.

(Joint work Jordan Boyd-Graber, Daniel Osherson and Rob Schapire.)