Gregory Garretson
Boston University
gregory@bu.edu

The use of translation corpora for grammatical analysis: The case of ‘of’

Languages often use one structure for several expressive purposes. One English example is the preposition ‘of’, as in (1):

(1) a. the hypothesis of the researcher
b. the roof of the house
c. the mining of the ore
d. a bunch of daisies

Despite these expressions’ surface similarity, their underlying semantics are very different. Using a monolingual corpus, one may find many such tokens and sort them until semantic categories take shape, as was done with ‘of’-tokens in Garretson et al. (2002). However, a problem remains—the resulting categories are based on intuition. We might prefer to ground our analysis instead in observable surface distinctions. One such approach involves the use of a translation corpus (see e.g., Johansson 1998, Dyvik 2002).

When translated to Swedish, the above examples look like those in (2):

(2)	a. forskarens hypotes	literally	“the researcher’s hypothesis”
	b. taket på huset		“the roof on the house”
	c. brytningen av malmen		“the mining of the ore”
	d. en bukett tusenskönor		“a bouquet daisies”

In Swedish, a number of distinct forms are used. We may hypothesize that these encode different semantic categories. I focus on the following question: Can the fact that Swedish employs various formal options to represent these meanings illuminate semantic distinctions not easily seen in English? In addition, I seek to address a broader question: What role can translation corpora play in analyzing grammatical—not just lexical—categories?

The English-Swedish Parallel Corpus (Altenberg and Aijmer 2000) is an ideal resource for exploring such questions. 1500 randomly-chosen ‘of’-tokens from the English-original portion of the ESPC and their Swedish translations were coded for translation form and semantic category, using the 18-category taxonomy in Garretson et al. (2002). The monolithic English ‘of’-form was found to map onto 26 different surface forms in Swedish. In light of this one-to-many mapping, the English tokens were re-analyzed and the semantic categories adjusted, resulting in the creation of 6 new semantic categories. In this talk, I will present the methodology of the study, the results, and the adjusted ‘of’-taxonomy, and will discuss the validity of using translation corpora in analysis of this sort.

References:

Altenberg, Bengt, and Karin Aijmer. 2000. The English-Swedish Parallel Corpus: A resource for contrastive research and translation studies. In C. Mair and M. Hundt (eds), *Corpus linguistics and linguistic theory. Papers from ICAME 20, Freiburg im Breisgau, 1999*. Amsterdam: Rodopi: 15-33.

Dyvik, Helge. 2002. *Translations as Semantic Mirrors: From Parallel Corpus to Wordnet.* Presentation at ICAME 23, Göteborg, Sweden, May 22-26, 2002.

Garretson, G., B. Skarabela, and M. C. O’Connor. 2002. *Mapping out the English possessive: Using corpora to differentiate the senses of ‘of’.* Poster presentation at ICAME 23, Göteborg, Sweden, May 22-26, 2002.

Johansson, Stig. 1998. On the role of corpora in cross-linguistic research. In Johansson and Oksefjell (eds), *Corpora and cross-linguistic research: Theory, method and case studies*. Amsterdam: Rodopi. 1-24.