Translation-driven mapping of semantic fields: the case of Dutch and French inceptive verbs

Lore Vandevoorde  Gert De Sutter  Koen Plevoets
Ghent University
{Lore.Vandevoorde, Gert.DeSutter, Koen.Plevoets}@UGent.be

The (semi-)automatic retrieval of semantically similar words has become of increasing importance to lexical semantics and lexical variation studies (e.g., Peirsman et al., 2010), which has led to the advent of vector-based approaches like Latent Semantic Analysis (Landauer and Dumais, 1997), first and second order bag-of-words models (Manning and Schütze, 1999) and the behavioral profiles method (Divjak and Gries, 2006, 2009). These models are generally characterized as distributional, which means that they capture word meaning in relation to their context in large corpora. The present corpus-based study will put forward the use of translational data and translation corpora as another reliable way of retrieving semantic relationships and mapping them in semantic fields, complementary to the Semantic Vector Spaces. In computational linguistics, the use of more than one language to identify lexical relationships has already proved to be a successful way of resolving problems of ambiguity (Dagan et al., 1991) and as a cross-lingual solution for word sense disambiguation (Lefever, 2012). Drawing on Dyvik’s (2004, p. 311) assumption that “semantically closely related words ought to have strongly overlapping sets of translations”, overlapping sets of translations should commensurably reveal the semantic relations between translations, between translations and their source language items and, more importantly, between the source language items themselves. The present corpus-based study will map out the semantic field of the Dutch inceptive verb BEGINNEN and its most salient French translation COMMENCER by creating semantic fields through what we will call “back-and-forth translation”. The method is carried out as follows: first, all translations of a given (set of) lexeme(s) in a large corpus are checked manually. Then, inversely, all translations of these translations back into the initial source language (we will call this “back-translations”) are looked up. These so-called back-translations enable us to access the structure of the semantic field of the initial (set of) lexeme(s) via the first-order translations, thus guaranteeing a verification of the initial set of lexemes as well as a broadening of the semantic field without any (word class) restriction imposed by the initial selection of lexemes.

We used translational data extracted from the Dutch Parallel Corpus, a ten-million-word parallel and comparable corpus, balanced with respect to five text types and four translation directions (Macken, et al., 2011). In order to generate the semantic field of BEGINNEN, a concise set of near-synonyms was selected, consisting of beginnen, aanvangen, een aanvang nemen, starten, van start gaan and aanwatten. We based our selection on lexicographic data and inter-annotator substitution testing. The French translations of this set of onomasiological variants of BEGINNEN (n=528) were manually checked, returning a total of 17 different translations. Then, the 17 translations were inversely queued from the corpus as source-language lexemes. The Dutch translations of this set (n=1563) yielded 47 translations back into Dutch. The French variants of COMMENCER (entamer, démarrer and débuter) were submitted to the same procedure: manual checking (n=253) returned 8 different translations that were subsequently translated back into French, returning 31 back-translations (n=1393).
The resulting frequency tables were analyzed with the technique of correspondence analysis (Greenacre, 2007; Lebart et al., 1998). Correspondence analysis arrives at a lower-dimensional representation of the row and column categories, analogous to a Semantic Vector Space. More specifically, the multidimensional data structure of the 6 variants for BEGINNEN with 17 French translations were approximated in 2 dimensions, thus mapping out the semantic field of BEGINNEN. The 47 back-translations were subsequently projected onto this space as so-called “supplementary points”. The rationale of this approach is that the projected back-translations do not reveal onomasiological but semasiological clusters. The same procedure was followed for the variants of COMMENCER.

The Dutch back-translations of BEGINNEN and the French back-translations of COMMENCER were plotted into two separate graphs, depicting their respective semantic fields. We observe that both graphs show lexemes clustering together. Figure 1 shows that most lexemes are in the plot’s origin, e.g. beginnen [to begin], meteen [right away], ten eerste [firstly], aanvang [onset]. This cluster can consequently be interpreted as the prototypical center, consisting of lexemes with the basic meaning of the inceptive category, viz. “start of a general process”. A second cluster appears slightly to the right of the central cluster, consisting of lexemes like starten [to start], lanceren [to launch], op gang brengen [to bring about], starter [starter], actief [active]. They generally refer to the “starting up of a business, a company, a medical treatment or an interpersonal relation”. More to the right, we find opzetten [to set up], invoeren [to establish], instellen [to set up] and in werking treden [become effective], mostly referring to a “rule or legislation becoming effective”. The outlying cluster (bottom left of the origin) consists of aanvragen [to request], een aanvang nemen [to commence], sluiten [to close] and ingaan [to take effect], commonly appearing in texts provided by governmental instances and usually referring to a “lease or hire agreement taking effect”. When looking at Figure 2 for COMMENCER, we find a fairly similar graph to the one of BEGINNEN with most lexemes around the origin (e.g. se lancer [to dive into], commencer [to begin], départ [start] and d’abord [first]).

As appears from our results, translational data are an interesting source for the bottom-up identification of a semantic field’s structure and for the differentiation of prototypical meanings from peripheral ones. Although we are not able yet to compare our translational approach with the distributional approaches mentioned above, the creation of semantic fields on the basis of translational data appears to have several advantages. Firstly, it does not require complex annotating techniques making it far less time consuming than some other (distributional) methods. Secondly, our translation-driven approach yields a semantic field with different word classes, which broadens the structure of the generated semantic fields, a strategy not often adopted by distributional models. Finally, our method provides an opportunity for a straightforward, cross-linguistic comparison of semantic fields.
**Figure 1: Semantic field of BEGINNEN**

**Figure 2: Semantic field of COMMENCER**
References


