

Defining verb semantic classes for French and their semantic characterization

Patrick Saint-Dizier

IRIT-CNRS, 118 route de Narbonne, 31062 TOULOUSE Cedex France

KEYWORDS: lexical semantics, lexical organization, computational linguistics.

E-MAIL: stdizier@irit.fr
FAX NUMBER: 61 55 62 58
PHONE NUMBER: 61 55 62 44

Abstract

In this document, we first show how verb semantic classes can be formed from syntactic descriptions (called contexts) realized in a way close to B. Levin's work. The verb class system we obtain for French is radically different from the system obtained for English. We explain why and what kind of additional criteria (thematic grids) should be used to get a class system which has the good properties of the English one. The different softwares that permit this classification are presented. Finally, we show that verb semantic classes is a powerful means for organizing the various semantic information related to verbs, useful in NLP applications. We show how various, coherent semantic or conceptual systems can be associated with these verb classes: selectional restrictions, thematic grids, aspectuality, and the Lexical Conceptual Structures (LCS).

1. Introduction

This work is primarily based on B. Levin's work (Levin 93) for English, where she shows that the syntactic behavior of verbs is essentially predictable from some aspects of their semantics. By syntactic behavior, she means the way arguments are realized in a syntactic form with respect to the predicate, how they can move (e.g. to define ergative or passive forms) and when they can be deleted. These movements and deletions are called alternations. It is important to note that the operations described by alternations are strictly lexically-based (therefore they do not include Wh- or NP-movements).

Although B. Levin's work corresponds to a certain level of granularity in the linguistic description which has a certain degree of stability, we do not think that it really introduces a new level with a theoretical status in syntax or in semantics. Rather, we consider this work as a very useful, practical and relatively comprehensive one, which can form

a good perspective and a good practical basis for the extraction and for the organization of lexical data.

We have substantially reformulated B. Levin's system in order to avoid a number of difficulties inherent to her approach, which do not seem to us to be central to her system (as she recognizes it in recent discussions), and to make the system more declarative and more usable in NLP systems. Instead of considering the alternation system of B. Levin, we have reformulated this notion into a more declarative one: the context system. Then, we can assign to each verb-sense the set of syntactic contexts it accepts. By syntactic context we mean a set of distribution frames (a frame may include several distributions) where the category and some additional syntactic and semantic information is used to describe the nature and the form of the possible complements and the subject of a verb. This is presented in (Saint-Dizier and Marrafa 96) (and also at the last ACH-ALLC conference). We now show how this system is used to construct semantic verb classes and how semantic information can be associated with these classes.

2. Verb classes

Verb classes are formed out from verbs having similar sets of contexts by specific software in Prolog that we have defined. For about 1700 usual verbs of French, we obtain about 400 verb classes (This is not a surprising result). A verb with different senses appears in different classes. We, in fact, consider that a class could define the 'boundaries' of a verb-sense and could therefore introduce a 'data-driven' way of defining polysemy for verbs.

Classes produced are very different from the English ones. We have, on the one hand, very large classes of verbs, where verbs are not necessarily semantically related (subsets of semantically related verbs are observed), and on the other hand, very small verb classes, with verbs having very close meanings. We also get classes with different types of verbs, often accepting few contexts. In general the more contexts a class is based on, the better its quality is. Good results are obtained with classes associated with at least 5 contexts. The global semantic relatedness is about 61%, which is not very good. This rate has been computed from a WordNet classification (Fellbaum 90) that we have realized following the principles of WordNet for English and using exactly the same verb-senses. This is due to the fact that in French we have less contexts than there are alternations in English. It is therefore not surprising that they form a less discriminatory system w.r.t. verb class formation. Also, our contexts do not contain semantic information (as those of B. Levin's do).

We have realized a detailed analysis of the results

in order to be able to improve the classification system, so that we can obtain, on an as automatic as possible way, semantic verb classes of a quality equivalent to those obtained by B. Levin for English. We will consider 2 principles: the taking into account of a few exceptions (which slightly improves the quality of semantic classes), and the introduction, step by step, and on a global basis of very general purpose semantic criteria.

We have first conducted a second experiment where, similarly to B. Levin, we allow exceptions in classes. To be more precise, we allow verbs in a given class to have one context more or less than the set of contexts established as the norm of the class. This flexibility can be adjusted to only those contexts estimated to be less important. Besides allowing the grouping of verbs which have very close syntactic behaviors, this flexibility also allows us to weaken the effect of possible judgement errors in the construction of the list of contexts associated with every verb (in some cases, it is not very easy to say if a verb definitely accepts a context or if it is just a marginal use). In this case, we obtain slightly better results, in particular for the smaller classes which are now larger and have a size which can really be considered as that of a class. Larger classes have not changed much. We now need to elaborate criteria for decomposing these classes into smaller units where verbs are semantically related.

There are two basic sources of information which can be considered to decompose these latter classes: selectional restrictions and theta-grids. The most reliable and relevant system are theta-grids, for which we have defined relatively fine-grained theta-roles (Pugeault et al. 94), inspired from (Dowty 91). We have decomposed these larger classes into smaller units based on the contents of thematic grids: e.g. the presence of a localization (spatial, temporal or abstract), of a means or of an incremental beneficiary theme, etc. We then obtain a set of verb semantic classes comparable to the results obtained for English by B. Levin. Notice that thematic grids alone (i.e. without contexts) would not have been sufficiently discriminatory to construct verb semantic classes, since there are hundreds of verbs with e.g. a volitional agent and a general theme).

We have extended our classification system in order to take into account these basic semantic criteria. We then get much better classes, which can be hierarchically organized (an evaluation has been carried out and is given in (Saint-Dizier 95b).) A global result of semantic relatedness of 89% is obtained. There still remains some work to be done by hand to go beyond this rate, in order to deal with exceptional verb behaviors and with verbs which have been misplaced for various reasons. We have identified a few methods for these

manual tasks, so that they can be reproduced for other subsets of verbs, of general purpose and related to technical domains, and also for verbs of different languages.

3. Semantic characterization

The last point of our research is to show that the verb semantic class system is a very powerful system for organizing lexical semantics data, besides the well-known essential lexical semantics relations such as *isa*, *part-of*, *synonymy* and *oppositions*. These relations are essentially paradigmatic whereas the relation introduced by verb classes is essentially analytic.

In a first stage, we have constructed verb semantic classes that can be grouped into families such as: the 'say', 'manage', 'transfer of possession', 'movement' and 'psychological' verb families (Saint-Dizier 95). These families are more general than those introduced in B. Levin's classification. With our verb sample, we have identified so far 16 verb families. These families can be further decomposed into sub-families according to precise criteria related to the semantics of the class. For example, we have the sub-family 'transfer of possession with retro-transfer of money' a subclass of the family of transfer of possession. A subfamily contains several verb semantic classes. This decomposition introduces an *isa* hierarchy, represented as a directed graph. We can then associate with every family a set of semantic information (such as selectional restrictions, fragments of thematic grids and LCS-based primitives) which is inherited by the subfamilies and the verb semantic classes. Then, at the level of each subfamily (and next, verb semantic class), we can associated more precise or additional semantic information. This is not an easy task: it requires the introduction of various semantic elements related to different ontological dimensions. So far, we have successfully decomposed and described the family of transfer of possession (the 'say' family is in progress). This work has then been confirmed in most of its aspects (and extended, e.g. to identify prototypical verbs) by psychological experiments carried out on a quite large population of children (10 years old) and adults (students) (Dubois and Saint-Dizier 96).

The introduction of fragments of LCS representations (represented by typed-lambda expressions where underspecified elements are in the scope of a lambda-expression) is of much interest. First it shows us how a verb sense can be constructed from more generic and abstract senses specified at the level of families and subfamilies. Next, it introduces methodological elements to define additional primitives, central to a family, in the spirit of those already defined in the LCS. Finally, it shows the limits of the LCS and opens research directions on a concrete basis.

4. Applications

In terms of applications, the first point to mention is the classification programmes we have defined and the methods defined around them, their use in different situations and languages, and the methods to use for the tasks which cannot be done on an automatic basis.

The second point is the creation of a quite large lexicon for French of predicative forms with a graphic interface that allows users to navigate in the semantic verb classes and to have access to the different semantic forms. This lexicon is associated with parametrized procedures that allow the construction of a sublexicon from that lexicon, specifically designed for an application. This lexicon can also be used for teaching purposes.

The last point is the use of this lexicon in information retrieval applications where we want to be able to extract structured information (predicates and arguments) and to be able to represent it in a generic way (using LCS, for example, combined with a few ontological elements). Such an application has been successfully carried out (Pugeault et al. 94) for the French National Electricity company) and has been evaluated in depth.

Saint-Dizier, P., Marrafa, P., Constructing a knowledge base for describing the general semantics of verbs, in vol. 6 of 'Research in Computing for the Humanities', Oxford Univ. Press, 1996.

Speas, M.J., Phrase Structure in Natural Language, Kluwer Academic Press, 1990

Bibliography (Short)

- Dorr, B., Machine Translation: a view from the lexicon, MIT Press, 1993.
- Dowty, D., Thematic Proto-roles and Argument Selection, Language, vol. 67-3, 1991.
- Fellbaum, C., The English Verb Lexicon as a Semantic Net, International Journal of Lexicography, 3/4, 278-301, 1990.
- Grimshaw, J., Argument Structure, MIT Press, 1990.
- Jackendoff, R., The Status of Thematic Relations in Linguistic Theory, Linguistic Inquiry 18, 369-411, 1987.
- Jackendoff, R., Semantic Structures, MIT Press, 1990.
- Levin, B., English Verb Classes and Alternations, the University of Chicago Press, 1993.
- Napoli, D.J., Predication Theory, Cambridge University Press, 1990.
- Pugeault, F., Saint-Dizier, P., Monteil, M.G., Knowledge Extraction from Texts: a method for extracting predicate-argument structures in texts, in proc. Coling 94, Kyoto, 1994.
- Saint-Dizier, P., Viegas, E., (Eds.), Computational Lexical Semantics, Cambridge University Press, 1995.
- Saint-Dizier, P., A semantic classification of French verbs based on B. Levin's approach, research report, IRIT, 1995b.