Constructing Verb Semantic Classes for French: Methods and Evaluation

Patrick Saint-Dizier
IRIT-CNRS
118 route de Narbonne F-31062 Toulouse Cedex France
stdizier@irit.fr

Abstract

In this paper, we study a reformulation, which is better adapted to NLP, of the alternation system developed for English by B. Levin. We have studied a set of 1700 verbs from which we explain how verb semantic classes can be built in a systematic way. The quality of the results w.r.t. semantic classifications such as WordNet is then evaluated.

1 Aims

Predicative forms are complex to describe; it is indeed necessary to describe in detail their syntactic behavior, the different meanings they may convey, preferably at different levels of granularity (e.g. argument structure, thematic grids, conceptual representations (Jackendoff 90)), and the relations between syntactic forms and meaning(s) (Levin 93), (Williams 94). It is also important to hierarchically organize these predicative forms so that syntactic and semantic descriptions are minimal and coherent.

Our work focusses on verbs and is primarily based on B. Levin's work (Levin 93) for English, where she shows that the syntactic behavior of verbs is in a large part predictable from some aspects of their semantics. By syntactic behavior, she means the way arguments are syntactically realized with respect to the predicate. This includes the description of the basic distribution of the arguments, the description of the other positions they may occupy (e.g. ergative and passive forms) and when they can be conjoined or deleted. These descriptions are called *alternations*. This work results in the creation of organized verb semantic classes, mainly based on their syntactic behavior (alternations may also include specific semantic restrictions). It is an extremely useful and detailed study of the syntactico-semantic relations between a predicate and its arguments.

We show here how the alternation system can be reformulated in a more NLP-oriented way, and develop for French a set of syntactic descriptions, called *contexts*, which share many similarities with alternations, and propose principles that help defining their form and contents. Next, we show how verb semantic classes can be constructed in a systematic way and evaluate them w.r.t. WordNet-like classifications. The implicit semantic conveyed by contexts is also analysed. The work presented here is applied to French, but can be transposed to other languages.

2 The context system

2.1 General approach and motivations

We have reformulated Beth Levin's notion of alternation into a more declarative one: the notion of *context*. A context is a frame where the category and some additional syntactic is used to describe a precise form and position the arguments of a verb may have in a sentence. Verb classes are then formed from verbs having similar sets of contexts.

Very briefly, compared to the alternation system, our approach avoids having to define a basic form from which alternations are produced and to have to explain what is the relation between a basic and an alternated form. Moreover, it avoids us to have to account for changes in meaning provoqued by alternations (e.g. by the adjunction of a preposition).

Defining contexts has led us to formulate a few principles:

- contexts should be of general purpose, this means that: exceptional forms should be avoided, only non-ambiguous and easy-to-use forms are acceptable, and theory-neutral descriptions should be used.
- contexts should minimally overlapp,
- they must only describe lexical properties; the scope of a context is usually a proposition
- as less semantic data as possible should be used, otherwise the classification will also be based on semantic criteria,

- the exact level of granularity of a context should be defined by feedback and retro-evaluation on verbs,
- consider generalizing two contexts into one, if their discriminatory power is low.

These principles allow us to partly automate the determination those of contexts which can be associated with a given verb (for example by corpora inspection). However, there will always remain quite a lot of manual work to check and improve the results, in spite of some promising research in this direction (Dorr et al. 95a).

As shall be seen below, the context system (which is not really a new concept) provides us with a very powerful tool for specifying and organizing the syntax and the semantics of verbs. Our contribution at this level is the way a context is defined, at what level of generality, with what formal means, and the way contexts are used to form verb classes.

From a methodological point of view, contexts for French have been defined from a transposition of some English alternations (about 1/3 of our contexts), from French syntactic descriptions, among which (Gross 75), from corpora and from our own intuitions of language. Context coverage has then been validated on corpora to ensure that we cover most of the syntactic behaviors of arguments w.r.t. predicates.

2.2 Description of contexts

Contexts and the detailed criteria used to define them are presented in (Saint-Dizier 95). A context is a set of 'extended' distribution frames:

1. **a set:** a cluster of syntactic forms which must all be valid for a given verb-sense. A verb accepts a certain context if it accepts all the distributions the context is composed of. A distribution is a list of syntactic constructions (NPs, PPs and sentences); this list is ordered and corresponds to the way these constructions are linearly realized in the surface form as arguments or modifiers.
2. **'extended':** syntactic category distributions are expressed as a Type Feature Structure (written in Login (Aït-Kaçi and Nasr 86)). We have identified several types of constraints:

- *Local constraints on arguments or on the verb:* thematic roles (including those defined in (Pugeault et al. 94), from (Dowty 89, 91)), the verb subcategorization frame, the arity of the verb, and a few commonly-admitted selectional restrictions.
- *Introduction of syntactic forms:* coordination of arguments, introduction of reflexive pronouns and of a few modifiers.
- *Relations between arguments:* thematic grids, modifier-modifiee relation between arguments (e.g. noun complements), and expression of essential semantic relations: container-containee, and part-whole of various types.

Our descriptions are more declarative than alternations, however, it is clear that this formalism allows us to introduce some forms of constraints between basis forms (via constraints on the verb) and the form being described. Similarly, the use of clusters of descriptions permits us to relate two forms.

We have defined 70 contexts, including 'basic' contexts (corresponding to 'direct' realizations of argument structures) and non-basic ones. We have grouped the non-basic ones according to some similitudes into 17 subclasses. We have a total of 23 basic contexts (of general purpose) and 47 non-basic ones (there are 89 alternations in English). Non-basic contexts include the description of: middle reflexives, passives, inchoatives, place-subject inversion, introduction of the semi-auxiliary *faire*, support verbs with nominalization of the predicate (e.g. *crier - pousser un cri*), various forms of argument deletion, preposition change, reciproquals, body-part reformulations, means-instrument raising, reflexives, argument 'des-incorporation', perspective change, there insertion, etc.

For example, we have the famous English spray/load alternation, which also exists in French, which is described as follows:

```
context([dist(111,   % context ID is 111
  verb([]),  % no constraint on verb
  phrases([
   xp(syntax=>syn(cat=>n)), % distribution
   xp(syntax=>syn(cat=>p)),
   xp(syntax=>syn(cat=>n,
     type-prep=> [sur,dans]),
     semantics=>sem(thematic=>[[loc]]))]),
  constraints([]),
  ex([je,pulverise,la,peinture,sur,le,mur])),
              % I spray paint on the wall
        dist(111, verb([]),
  phrases([ xp(syntax=>syn(cat=>n)),
   xp(syntax=>syn(cat=>n),
       semantics=>sem(thematic=>[[loc]])),
   xp(syntax=>syn(cat=>p, type-prep=>[de]),
         semantics=>sem(thematic =>[[tg]],
          sem-type=>tsem(semp=>substance)))]),
  constraints([]),
  ex([je,pulverise,le,mur,de,peinture]))]).
              % I spray the wall with paint
```

(tg = general theme and loc = localization).

3 Construction of verb classes

3.1 Typology of the verb sample

The experiment presented here has been realized on a set of 1700 usual verbs which are the most frequently used in French. Our aim is to classify

3000 to 4000 verbs. The size of the sample considered so far is however sufficiently large to allow us to draw significant and precise conclusions.

It should be noticed that contexts are associated with a given word-sense, not with all the senses of a verb. Each sense of a polysemous verb is associated with a different set of contexts. The description of a verb is the following:

```
verb([verb],arity, [basic context number],
  [thematic grid],[prepositions],
  [list of contexts]).
verb([admirer],3,[20],[ae,tib,src],[pour],
   [50,51,61,102,150,171,180]).
```

(ae = effective agent, tib = incremental beneficiary theme, src = source). Contexts have been associated with verbs on the basis of a number of linguistic analyses of French (e.g. (Gross 75)), of already existing lexicons, and from corpora inspection and our own intuitions.

3.2 A simple verb classification

We have carried out a simple classification where a verb class contains all the verbs which accept exactly the same set of contexts. This is not the classification method adopted by Beth Levin: her verb classes are constructed from subsets of alternations, intuitively selected, which are sufficiently selective to allow for the characterization of a set of semantically related verbs. Exceptions are allowed in order to effectively gather all the verbs which are intuitively semantically related. Her classification method, based on a large number of linguistic analyses involving some subtle semantic criteria (e.g. intentionality), can only be carried out manually and is therefore not adapted to our approach.

We obtain a total of 953 classes. We get a large number of classes with just one element (about 77%), this is not surprising, however, since contexts can be combined in a large number of ways. 56% of the verbs appear in classes with at least 2 elements, and 33% of them are in classes with at least 5 elements.

This number of classes is quite large compared to Beth Levin's results (about 200 classes), however, our classes have been constructed on a strict equivalence class basis, without any exceptions, and all the contexts have been taken into account. We have an average of 1.8 verbs per class. A similar result was also obtain by (Gross 75), on a different basis (including morphology) and with more criteria (about 200).

A very informal study of the progression of the number of classes tends to indicate that the increase of the number of new classes is not linear, but progressively decreases. It seems that beyond 2500 verbs almost no new verb class should be created, defining about 1100 to 1200 classes. But this is clearly too much.

3.3 Evaluation of the semantic relatedness of verb semantic classes

The overall quality of the verb classes are studied in detail in (Saint-Dizier 95). With the same set of verb-senses, we have carried out a classification similar to the classification proposed in WordNet. Besides the main categories presented in (Fellbaum 93), we have added two classes: aspectual verbs and verbs expressing causality. We have then subdivided these main categories according to different types of properties or constraints following as much as possible those defined in WordNet. In our current classification, we consider 198 hierarchically organized classification criteria, instances of the *is-a* (or troponymy) relation, the depth of the decomposition is 3 (Saint-Dizier 96). We therefore get 198 verb classes (called WN classes) for levels 1 to 3. For example, a three level decomposition is for *movement* verbs (level 1), *directed motion*, *local motion*, etc. (level 2) and *upward motion*, *downward motion*, etc. (level 3).

If we now compare the degree of overlapp between the classes (with at least 2 elements) formed above from syntactic contexts (called VS classes) and those of WN, we get the following results:

WN level	(2)	(3)	overlapp VS/WN
1	17	120	54%
2	75	41	47%
3	106	18	32%

(1): number of WN classes, (2): average size of a WN class at this level.

Classes where verbs are associated with at least 5 contexts are of a much better quality (semantic relateness with WN classes above 64%) than those under 5. The best classes contain an average of 4 to 7 verbs, larger classes (above 10 elements) are often of a lower quality or may contain several subsets of semantically related verbs: in a large number of classes with more than 8 elements we found 2 or 3 subsets of classes of WN. These classes are often formed from a small number of contexts (1 to 3), which explains their low semantic relatedness rate.

Globally, these results aren't very good. If we want to explore in more depth the cooperation between syntax and semantics, and if we want to be able to construct verb semantic classes on a rigourous basis, it is necessary to develop methods that improve the quality of VS classes (considering that syntactic criteria are the most 'rigourous' ones *a priori*). The first approach, which is the simplest, is to make the classification more flexible by allowing exceptions: a verb in a class may have one more or one less context than the norm of the class. This approach gives however very bad results, with an overlapp VS/WN rate below 35%. To improve that rate, exceptions should depend

on the VS class, but this is extremely subjective and hard to carry out. The second type of solution consists in analyzing the implicit semantics conveyed by contexts and to form classes from sets of contexts, on the basis of their implicit semantics. Then all the verbs accepting exactly an *a priori* given set of contexts will belong to the same VS class, even if they accept many other contexts.

4 Analysis of the semantics conveyed by contexts

Some contexts are quite general and are not related to precise semantic notions, while others convey clearly identifiable meaning components.

First, there are contexts which convey very precise meaning components, which are not taken into account, for various reasons, in WordNet classifications. For example, the context of the form '*pousser* + nominalization of verb' is associated with verbs of sound emission: painful sounds for humans and any sound for animals; verbs which accept the 'dans/en-de preposition change' convey an idea of putting something into something else (*bourrer le tuyau de papier, bourrer le papier dans le tuyau*).

Next, a second type of context conveys meaning components which can directly be associated with WN criteria. We have carried out a detailed analysis of the correlations between WN criteria and contexts. There are 19 non-basic contexts (out of 47), which can very clearly be associated with 1 or 2 WN criteria. For example, context 91, (je fais atterir l'avion ('I make land the plane')), is at 90% associated with verbs of body care. Context 151, (alternation 2.13.4 in Beth Levin: Les grimaces de Jean terrifient Sophie), is associated at a rate of 60% with psychological verbs. This is studied in detail in (Saint-Dizier 96).

5 Perspectives

The semantic characterization of contexts should allow us to construct verb semantic classes on a stronger basis, and with a clear method. We have carried out preliminary experiments on transfer of possession verbs which confirm this hypothesis. Besides these results, it is of much interest to study how WN and VS classification systems can cooperate and can contribute to defining the syntax and the semantics of verbs, in a quite comprehensive and fine-grained way. It should be noted that we consider that the syntax-based approach (VS) is the most stable and the most formal approach, it should therefore be the central element of our classification strategy. WN criteria are extremely useful, but they remain nevertheless somewhat intuitive and less connected to language realizations.

Our ultimate goal, from this perspective, is to associate with families of verb classes, verb classes and possibly individual verbs, hierarchically organized semantic representations, under the form of partially instanciated LCS-based semantic representations (a sucessfull experiment in this direction has been carried out for English by (Voss and Dorr 95), and also by ourselves on verbs of transfer of possession) and ontological knowledge.

Acknowledgements I thank Bonnie Dorr, Martha Palmer, Beth Levin, Doug Jones and Palmira Marrafa for discussions that helped improving this research. Many thanks also to Alda Mari who carried out parts of the syntactic descriptions of verbs.

References

Aït-Kaçi, H., Nasr, R., LOGIN: A Logic Programming Language with Built-in Inheritance, journal of Logic Programming, vol. 3, pp 185-215, 1986.

Dorr, B., Garman, J., Weinberg, A., *From Syntactic Encodings to Thematic Roles: Building Lexical Entries for Interlingual MT*, Machine Translation, 9-3, Kluwer Academic, 1995

Dowty, D., On the Semantic Content of the Notion of Thematic Role, in G. Cherchia, B. Partee, R. Turner (eds), *Properties, Types and meaning*, Kluwer, 1989.

Dowty, D., Thematic Proto-roles and Argument Selection, Language, vol. 67-3, 1991.

Fellbaum, C., *English Verbs as Semantic Net*, Journal of Lexicography, 1993.

Gross, M., *Méthodes en syntaxe*, Masson, Paris, 1975.

Jackendoff, R., *Semantic Structures*, MIT Press, 1990.

Levin, B., *English verb Classes and Alternations: A Preliminary Investigation*, Chicago Univ. Press, 1993.

Pinker, S., *Learnability and Cognition*, MIT Press, 1993.

Pugeault, F., Saint-Dizier, P., Monteil, M.G., *Knowledge Extraction from Texts: a method for extracting predicate-argument structures from texts*, in proc. Coling 94, Kyoto, 1994.

Saint-Dizier, P., Verb Semantic Classes in French, IRIT research report, December 1995, (revised and extended May 1996).

Saint-Dizier, P., *Semantic verb classes based on 'alternations' and on WordNet-like semantic criteria: a powerful convergence*, in proc. workshop on Predicative Forms, univ. of Toulouse, August 1996.

Voss, C., Dorr, B., Toward a Lexicalized Grammar for Interlinguas, Machine Translation, 9-4, Kluwer Academic, 1995.

Williams, E., *Thematic Structure in Syntax*, Linguistic Inquiry monograph no. 23, MIT Press, 1994.