Syntactic and Semantic Frames in PrepNet

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

Abstract

In this paper, we present an in depth revision of the preliminary version of PrepNet, an lexical semantics database of preposition behaviors. This revision includes a much more detailed structure for the language realization level.

1 Aims

Describing the syntax and the semantics of prepositions, in a way similar to verbs (e.g. in FrameNet (www.icsi.berkeley.edu/ framenet/), or VerbNet (www.cis.upenn.edu/verbnet/)) or to nouns (as in WordNet and EuroWordNet) is obviously a very challenging, but necessary task. Prepositions turn out to be a very useful category in a number of applications such as indexing, knowledge extraction, textual entailment and question answering since they convey basic meanings of much interest like instruments, means, comparisons, amounts, approximations, localizations, etc. They must necessarily be taken into account—and rendered accurately—for effective machine translation and lexical choice in language generation. However, so far, prepositions have not been studied in NLP circles as extensively as nouns and verbs.

PrepNet (http://www.irit.fr/recherches/ILPL/Site-Equipe/ILPL.htm under revision) is a framework that aims at constructing a repository of preposition syntactic and semantic behaviors. PrepNet is structured in two levels:

- the abstract notion level: global, language independent, characterization of preposition senses in an abstract way, where frames represent some generic semantic aspects of these notions,
- the language realization levels that deal with realizations for various languages, using a variety of marks (postpositions, affixes, compounds, etc.). However, we will keep the term 'preposition' hereafter for all these marks.

We present here a revised version of a preliminary PrepNet. Besides some simplifications of the structure, innovations mainly concern the structures provided at the language realization level, motivated by the study of a few notions for a variety of languages. We provide the means to describe syntactic subcategorization forms, as well as semantic and pragmatic restrictions on uses. It seems that our view is now usable in a number of languages, but some are still resistant, like, e.g. Malay.

2 Related work

There is quite a lot of literature on prepositions in psycholinguistics circles, and some in AI and in cognitive sciences (Horno Cheliz, 82), (Cervioni, 91), (Lindstomberg 97), (Mari, 00), (Pesetzky, 82), (Talmy 76, 85), but less in NLP (Saint-Dizier, 06).

A quite old, but still of interest work (Spark-Jones et al. 85) proposes, via cases or roles, a structure for prepositions, and their relations to verbs.

The basis and the starting point of our research was developed about 10 years ago by Bonnie Dorr, it is accessible at:
www.umiacs.umd.edu/b̃onnie/
AZ-preps-English.lcs.
This is a very large database of preposition semantic

representations, characterized by their LCS representation and, sometimes, by a thematic grid. It was concieved for machine translation tasks, which explains some of its features. There are about 500 entries (compared to our 165 abstract notions), for probably all English prepositions. Finally, the Preposition Project offers a lexicographic view of prepositions in English.

3 Main features of PrepNet

Within the PrepNet framework, we have identified so far 195 preposition senses (which may have subsenses in particular languages), which can be represented in the Lexical Conceptual Structure framework on the basis of 65 primitives, based on English preposition names (on, near, with, etc.). These senses reflect the variety of primitive notions conveyed by prepositions. Abstract notion representations may be a composition of several primitives. Primitives are viewed here as linguistic macros, which can then be interpreted depending on the environment (e.g. Euclidean geometry for spatial prepositions, fuzzy logic axioms for the notion of approximation).

To elaborate an adequate formalism for the syntactic and semantic aspects of prepositions we want to encode in PrepNet, we have studied in depth preposition realizations in language around the abstract notions of theme and approximation (French, Spanish, Catalan, English, Thai) and instruments (for German, Italian, Spanish, French, Arabic and Berber, Thai, Bahasa Malaysia, Hindi, Urdu, Kashmiri, Bengali, and Filipino). This latter notion is much wider than the first two, and has a large variety of realization parameters, which greatly contributed to this second version of PrepNet. A multilingual analysis needs to be somewhat transcategorial and both syntactic and semantic.

4 Preposition usage characterizations using Frames

4.1 General overview of abstract notions

In PrepNet, preposition senses are characterized by means of abstract notions which capture senses in a conceptual and language independent way. These abstract notions have been defined from corpus analysis and bilingual dictionaries (in particular the German-French Harrap's, which has an excellent conceptual approach to translation). From corpora, focussing on French, Spanish and English, we have studied every preposition lexeme, its distributions and constraints, and we have identified manually the different meanings they convey. We have then formed groups of senses out of which the abstract notions emerged. This took about 1.5 man/year of work, preposition being not as numerous as e.g. verbs or adjectives (about 50 to 100).

The first level, the abstract notions, is organized as follows:
- a first level characterizes a **semantic family**, of a level roughly comparable to thematic roles: localization, manner, quantity, accompaniment, etc.,
- a second level accounts for the different **facets** of a semantic family, e.g. source, destination, via, and fixed position for the localization family,
- a third level characterizes, roughly speaking, the **modalities of a facet** when appropriate. For example, the facet *manner and attitudes* is decomposed into 3 modalities: *basic manner, manner by comparison and manner with a reference point*. Due to space limitations, this latter level will not be much developed here.

Abstract notions are the following (revised from (Saint-Dizier, 05)):

- **Localization** with facets:
  - source, - destination, - via/passage, - fixed position.
  From an ontological point of view, all of theses facets can, a priori, apply to spatial, temporal or to more abstract arguments.

- **Quantity** with facets:
  - numerical or referential quantity, - frequency and iterativity, - proportion or ratio.

- **Manner** with facets:
  - manners and attitudes, - means (instrument or abstract), - imitation, agreement or analogy.
  Imitation: *he walks like a robot; agreement: he behaves according to the law*,

- **Accompaniment** with facets:
  - adjunction, - simultaneity of events (co-events), - inclusion, - exclusion.

Adjunction : *flat with terrace / steak with French fries / tea with milk*, Exclusion: *they all came except Paul*.

- **Choice and exchange** with facets:
  - exchange, - choice or alternative, - substitution.
  Substitution : *sign for your child*, Choice: *among all my friends, he is the funniest one*.

- **Causality** with facets :
  - cause, - goal or consequence, - intention - purpose.
  Cause: *the rock fell under the action of frost*.

- **Opposition** with two ontological distinctions: physical opposition and psychological or epistemic opposition, e.g.: *to act contrary to one's interests*.

- **Ordering** with facets:
  - priority, - subordination, - hierarchy, - ranking, - degree of importance.
  Ranking : *at school, she is ahead of me*.

- **Instrument** (see below),

- Other groups: **- Theme, - in spite of, - comparison**.
  Theme: *a book concerning dinosaurs*.

4.2 Representation of abstract notions

An abstract notion is characterized by:

1. **a name and a gloss**, that informally describe the abstract notion at stake,

2. a **conceptual representation**, in simplified LCS form,

3. **inferential patterns** and presuppositions, which will not be developed here.

4.3 Representation of the language level

While we have a unique, language independent, structure for abstract notions, we have a set of descriptions for each language. At this level, we may also have semantic subdivisions whenever relevant, called strata. Let us study here the direct usages, i.e. those which are not a priori metaphorical or in any other form of meaning shift. Our approach, however, integrates the possibility to describe these shifts either directly or via rules.

The language level descriptions include at the moment the following features:

- syntactic frames: the syntactic subcategorization frames the preposition heads (possibly in conjunction with other predicates like a verb), with some statistics on usage frequency. Frames are a little bit complex since they need to take into account (1) elements not completely headed by the preposition but which nevertheless play a major role (the verb and the 'external argument' of the preposition) and (2) the structure the preposition heads (in general an NP, possibly an S). This is realized by corpus inspection. In addition, alternations prepositions may undergo are given and some grammatical movements (such as fronting).

- semantic and domain restrictions: each argument in the frame may have selectional restrictions. These allow us to identify different language realizations. Restrictions may also be related to domains and not to arguments.

- pragmatic aspects: prepositions convey a number of pragmatic factors such as: stress (identifying a new focus), polarity, illocutionary force, formal character, etc. These are mainly given to restrict usages.

4.4 Basic case: the VIA notion

The facet VIA of the 'localisation' family describes a movement via a passage. The abstract notion frame is defined as follows:

```
VIA
'An entity X moving via a location Y'
representation: X : via(loc, Y)
```

This frame reads as follows: after the name and the gloss of the notion, we find a simple, LCS-based, semantic representation where X is the external argument, and loc specifies the domain: localization (other cases are assumed to be derived by metaphor).

Let us now consider the language level. In French, the by-default associated synset is {*par, via*} . X and Y are restricted respectively to concrete entities and location, the verb is restricted to inherently directed motion (in B. Levin's terminology), as in:

passer par la porte, transiter par la Belgique, Paris-Strasbourg via Nancy.
Syntactic frames examples are (given informally, in readable format):

```
[X(np,subj), Verb, Y(np,obj,optional),
preposition, Z(np,obj2)],
```

[Y(compound NP, +loc), preposition, Z(np, +loc)], etc.

Next, we also have a specific case (a strata) where the passage is narrow. In that case, the synset is {*à travers, au travers de, dans*}, and Verbs are either inherently directed motion or perception verbs (*regarder dans le téléscope, regarder à travers la grille*).

4.5 Compound forms: via under

The abstract notion VIA has a few strata that correspond to compound notions that express a more precise trajectory like via under, via above. For example, in French, to express this notion, the preposition *par* is combined with a fixed location preposition such as *dessous, dessus* etc. to form compounds such as: *par dessus, par dessous* (via under, via above). The frame structure remains the same, except that the semantic representation has then an embedded functional structure:

```
VIA UNDER
'An entity moving via under a location'
representation: X : via(loc, under(loc,Y))
```

At the language realization level, the French synset is: {par dessous }.

5 A more complex case: dealing with instruments

The study of the abstract notion of instrumentality, as reported in (Kawtrakul et alii, 06), has led us to revise and largely improve the language level formalism. Let us report here some of its main facets. In this work, 12 languages from 5 linguistic families are studied: Thai, Malay, Hindi, Urdu, Kashmiri, Bengali, German, Spanish, French, Italian, Arabic and Berber. Filipino has been recently considered.

Besides basic syntactic frames, of much interest is that most of these languages use other forms than 'prepositions' to realize these abstract notions: postpositions, various kinds of affixes, verb compounds, etc. A number of the languages studied have an instrumental case.

Our investigations tend to show that we can have on the one hand a stable abstract frame that represents the semantic and some pragmatic aspects of the abstract notion and, on the other hand, at the language realization level, a description of the behaviors of 'prepositions' in the various languages. We do not establish any direct connection between two preposition realizations in two languages. The relation, in terms of translation, is established via the set of restrictions imposed on each lexicalization that corresponds the best to the restrictions imposed on the argument Y. The impact of the other elements (arguments and verb) remains to be explored. Each language has an independent description.

5.1 Representing the abstract notion

The generic frame for instruments is as follows:

```
INSTRUMENT
'An actor X uses an object Z (the instrument)
to reach a goal E'
X, Y : by_means_of(E, Z)
```

In this frame, an event E is introduced to refer to the goal, e. g. 'cut bread' as in *John cuts bread with a knife*. Note also that in the semantic representation, the instrumental expression has wider scope over the event: this reflects the fact that most adjuncts have scope over propositions (predicate and its arguments).

In terms of restrictions, a major difficulty is the prototypicality of instruments. At a conceptual level, it is quite difficult to characterize what is a prototypical instrument for a given action. Each event has its own prototypical instrument, making corpus studies extremely large, probably unfeasable. For the time being, we leave the type of the instrument largely open. However, at the language realization level, we may have some useful restrictions, as will be seen below.

5.2 Dealing with language realizations

Let us now present a variety of language realizations that motivated the different facets of the formalism we have developed. In each case, we have a by-default synset of marks, and more restricted sets of marks for specific cases, related in particular to the semantic type of the instrument, but also to pragmatic effects or domains of discourse.

The different language variations presented below have been elaborated in several steps via cor-

pus and dictionaries. A first set of utterances was collected by using the various prepositions in each language and by analyzing the usage restrictions observed. Then we constructed a second set of utterances to confirm the analysis, by attempting to find counter examples. Counter examples always exist, but they must remain marginal for the analysis to be confirmed. Analysis was done independently for each language in order to avoid any influence. Much more data can be found in (Kawtrakul et alii, 06).

Let us now present language realization levels for a few quite diverse languages.

French by-default synset: `[avec, par, au moyen de, grâce à, à l'aide de, à travers]`.
some syntactic structures:
`[X(subj) Verb(action) Y(obj) preposition Z(np or S)], ['utiliser' Z 'pour' Verb(action, infinitive) Y]`, etc..
More informally, other syntactic properties are: the instrument is in general an adjunct to the VP, it therefore has the properties of such types of adjuncts. It undergoes the alternation: 'Characteristic property of instrument' (Levin 86) and a few other movements, e.g.: fronting. Finally, it cannot be inserted between the verb and the object.
Usage restrictions: introduces a focus on the instrument (in particular to focus on non prototypical instruments): `au moyen de`.
polarity: positive: `grâce à`

German by-default synset: `[mit, mit Hilfe von, mittels, durch, anhand, kraft, dank, per.]`
Of interest here are the domain and pragmatic restrictions on preposition usages, e.g.:
domain: juridical, psychological: `kraft, anhand`
formal usage: `mittels`
focus: `mittels, mit Hilfe von`
instrumental manner: `durch`.

Hindi by-default synset: `[se, me, ke karaan, ke dwaraa, kar, karaan, dwara]`
the syntactic frame encodes here postpositions, case marks and the SOV form: `[X(subject, ergative), Y(object, accusative), Z(adjunct), postposition, Verb(action)]`.
This form is a priori very regular, Z is an NP or an S. Y and Z can occasionally be permuted. Let us note some interesting usage restrictions:
instrument type: concrete: `se`
instrument type: abstract: `dwAra`
instrument type: means of transportation: `me`
involvement of instrument: agentive: `ke dwara`, causal `ke karaan`
instrumental path: `me, se`.

Concerning other Northern India languages, Urdu has about the same distribution and distinctions, while Kashmiri and most notably Bengali have some more distinctions, with the use of a large number of prefixes and suffixes, which are expressed in the subcat frame by means of features. Thai is relatively straightforward, prepositions may be even omitted.

Filipino is close to Malay and Indonesian in terms of structure, except that it is basically a VSO language. It has also a large number of marks to capture the notion of preposition: `[ng, kay, kina, sa]`, as in:
Pinalo niya ang aso ng patpat (litt. hits he the dog with a stick).
The syntaxtic structure is therefore: `[Verb, X(subj), Y(obj), preposition, Z]`.

So far, the formalism we have elaborated allows us to encode syntactic frames, restrictions, case marks, prefixes and suffixes as well as postpositions. However, languages of the malayo-polynesian family raise additional problems which are not so easy to capture. Let us just, for the sake of illustration, survey a few aspects here for Malay and Filipino.

Malay has three ways to introduce instruments: preposition + NP, affixes and compounding. Affixed words are built from stems which are instrumental nouns, this allows for the construction of the equivalent of PPs, based on the prototypical use of the instrumental noun. The most common being: prefixes: beR- (e.g. from kuda, horse, berkuda, on horseback), meN- (e.g. from kunci, key, mengunci, lock with key), prefix + suffix: meN- + -kan (e.g. from paku, nail, memakukan, to fasten with nails), and with suffix -i (e.g. from ubat, medicine, mengubati, by means of medicine). At the moment, we feel it may be confusing to add derivational morphology considerations (with many restrictions) into syntactic frames, probably an additional means would be necessary.

Similarly, Filipino has also a large number of marks that play the role of prepositions, one of which is viewed as an anteposed particle, working as a determiner.

6 Perspectives

Although we have stabilized semantic notions and some formalisms for the representation of preposition behaviors over a number of languages, PrepNet is still in a development stage. The descriptions over various languages are huge tasks. Our method for the future will be to proceed by notion and study a variety of languages to have a better grasp at the semantic distinctions and language realizations, as we did for instruments. For each case, a dedicated method is often required. Descriptions are encoded in XML, so that data can be easily shared.

Although the study of prepositions is an interesting topic in itself, it is of much interest to investigate how this work can be integrated into larger frameworks such as FrameNet or VerbNet, and this is one of our major prospective.

FrameNet says little about prepositions, but it has a few frames such as Accompaniment which are of interest. The roles defined in FrameNet are more accurate than the abstract notions of PrepNet, which aims at a relatively generic description. Those roles are related to a variety of situations which are not necessarily introduced by prepositions. However, a preliminary, exploratory, task could be to attempt to classify FrameNet roles under the main abstract notions of PrepNet.

VerbNet uses a quite detailed list of thematic roles which have some similarities with the top nodes of PrepNet abstract notions hiererachy. In a VerbNet frame, the syntax slot could be enriched by preposition type (abstract notions) restrictions. Similarly, predefined primitives such as location or direction are really close to our semantic representations in LCS, but they are used in a different manner. PrepNet has additional primitives to handle argument and non argument structures (e.g. approximation).

References

Cannesson, E., Saint-Dizier, P. (2001), *A general framework for the representation of prepositions in French*, ACL01 WSD workshop, Philadelphia.

Carmen Horno Chéliz, M. del, (2002), *Lo que la preposición esconde*, University of Zaragoza press.

Cervioni, J., (1991), *La préposition: Etude sémantique et pragmatique*, Duculot, Paris.

Dorr, B., Olsen, M.B., (1997), *Deriving Verbal and Compositional Verbal Aspect for NLP Applications*, proc. ACL'97, Madrid.

Dorr, B. J., Garman, J., and Weinberg, A., (1995), *From Syntactic Encodings to Thematic Roles: Building Lexical Entries for Interlingual MT*, Machine Translation, 9:3-4, pp.71-100.

Fellbaum, C., (1993), *English Verbs as Semantic Net*, Journal of Lexicography, vol. 6, Oxford University Press.

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

Kawtrakul, A. et alii (2006), A Multilingual Analysis of the Notion of Instrumentality, proc. EACL workshop on prepositions, Trento.

Levin, B., (1993), *Verb Semantic Classes: a Preliminary Investigation*, Chicago University Press.

Lindstromberg, S. (1997), *English Prepositions Explained*, John Benjamins.

Saint-Dizier, P., (2005), *PrepNet: a Framework for Describing Prepositions: Preliminary Investigation Results*, IWCS05, Tilburg.

Spark-Jones, K., Boguraev, B., A note on a study of cases, research note, dec. 85.

Talmy, L. (1976), *Semantic Causative Types*, In M. Shibatani (ed.), Syntax and Semantics 6: The Grammar of Causative Constructions. New York: Academic Press, pp. 43-116.

Talmy, L., (1985), *Lexicalization Patterns: Semantic Structure in Lexical Forms*, in *Language Typology and Syntactic Description 3: Grammatical Categories and the Lexicon*, T. Shopen (ed.), 57-149, Cambridge University Press.

Villavicencio, A., (2006) *Verb-particle Constructions in the WWW*, in P. Saint-Dizier (ed), *Syntax and Semantics of Prepositions*, Kluwer Academic.

Wierzbicka, A. (1992), *Semantic Primitives and Semantic Fields*, in A. Lehrer and E.F. Kittay (eds.), Frames, Fields and Contrasts. Hillsdale: Lawrence Erlbaum Associates, pp. 208-227.