An overview of PrepNet: abstract notions, frames and inferential patterns

Patrick Saint-Dizier
IRIT-CNRS
118, route de Narbonne
31062 TOULOUSE Cedex FRANCE
stdizier@irit.fr

Abstract

In this paper, we present the results of a preliminary investigation that aims at constructing a repository of preposition syntactic and semantic behaviors. A preliminary frame-based format for representing their prototypical behavior is then proposed together with related inferential patterns that describe functional or paradigmatic relations between preposition senses.

1 Introduction

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 and knowledge extraction 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.

This paper is based on a classification and preliminary results reported in (Saint-Dizier, 2005), (see www.irit.fr/recherches/ILPL/prepnet.html). This paper focusses on the frame structure and a classification of inferential patterns that organize preposition senses. We briefly sketch how the prototypical behaviors of prepositions, captured in the frames, can be enriched based on a corpus-driven methodology. In the future, we would also like PrepNet to be an open system in which external contributors can enter preposition descriptions of their own language.

2 Preposition usage characterization using Frames

PrepNet is structured according to three levels: the abstraction notion level, the frame level, and the language realizations level.

2.1 Architecture of PrepNet

In PrepNet, preposition senses are characterized by means of abstract notions which capture senses in a conceptual way. Abstract notions are then characterized by means of a set of syntactic frames, a semantic representation and shallow semantic restrictions. The semantic representation is a simplified version of the Lexical Conceptual Structure (LCS) [9]. Frames describe the structure in which the preposition is embedded (much larger than its maximal projection) so that appropriate constraints on usage can be stated. The abstract notion level can be viewed as an interlingua level, essentially conceptual. Representations are then further stratified to account for differences between prepositions. The lower level, (the realization level) deals with preposition realizations in different languages. To account for the large variety of realizations, we developed a multi-level partitioning approach, outlining usage norms, groups of exceptions, etc.

Within the PrepNet framework, we have de-

scribed so far 195 senses, using 65 primitives, based on English preposition names (on, near, with, etc.). The 65 primitives identified do 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). There are obviously decisions to make about sense distinctions and their encoding by means of primitives, one can argue on some choices and their ontological or cognitive status. A good test of this preliminary work will be the concrete use of PrepNet in the development of applications.

Work has been so far carried out on French, it is clear that some revisions and refinements may be needed when investigating other languages. For that purpose, we introduced some flexible means: abstract notions strata and a multi-level description of language realizations. However, we feel to have reached some level of stability for the abstract notion, and the for general architecture of the system. So far, descriptions are rather coarse-grained, with no underspecification.

An important issue is obviously the characterization of preposition uses over several languages. This is a major challenge because of the large variety of behaviors one can observed which are not necessarily only superficial, but which may involve different conceptual views. Another problem is that a number of languages use other syntactic categories, incorporation or morphology (such as case marks) instead of prepositions in some situations. So a multilingual analysis needs to be somewhat transcategorial or even syntactic.

2.2 Representation of abstract notions

Let us now concentrate on the representations we have settled. An entry at the conceptual level in PrepNet corresponds to an abstract notion [16]; it is composed of:

1. **a number, a name and a gloss**, that informally describe the semantics of the abstract notion at stake:
   [sense number], name from hierarchy above, 'gloss',

2. **a frame with constraints**, constraints are relatively 'shallow', these are further refined by corpus exploration and categorization at the realization level:
   X <ACTION> Y [sense number] Z,
   where X, Y and Z are the verb (the verb being noted as ACTION or STATE) arguments, this frame is followed by the specification of shallow constraints on the verb and on the arguments,

3. a **conceptual representation**, in simplified LCS form (in which we essentially keep the semantic field, for which we have developed a richer set of categories). At this level, only the semantics of the preposition is captured. This level introduces a decompositional approach to preposition meaning. This representation can be viewed as a kind of conceptual prototype.

By shallow constraints, we mean (1) the use of a quite generic (or shallow) set of semantic types, (2) the use of generic verb classes largely derived from WordNet [8] and (3) the use of a number of semantic fields for LCS representations: poss (possession), temp (time), loc (localization), psy (psychological), comm (communication), epist (epistemic), abs (abstract), prop (property), perc (perception), and amount (quantity). We view these features as prototypes, around which uses are grouped. Other uses, such as metaphors or metonymies, will be derived by means of inference rule schemas, among which, type coercion for metonymies. We include in the examples below the synset and an example for French (which must be specified apart) with each frame given, so that they are easier to understand.

2.3 Examples: frames and strata

The first example below illustrates the main elements given above. The facet VIA of the 'spatial' family describes an action occuring via a passage. Classifications and distinctions are essentially made below from the identification of shallow selectional restrictions and language realizations. The generic case is numbered 1, 1.1 is a strata, below 1, that deals with a specific situation where the via is a narrow passage (French examples are here just for the illustration purposes, since they appear only at the realization level):

```
[1] : VIA - generic.
X  <ACTION> [1] Y
'An entity X moving via a location Y'
```

```
  X: concrete entity,
  ACTION: movement verb,
  Y: location
representation: X : via(loc, Y)
French synset: {par, via}
example: Jean entre par la porte.

[1.1] : VIA - narrow passage.
'An entity X moving via /
    an action that uses
    a narrow passage in an object Y'
X  <ACTION> [1.1] Y
  X: concrete entity,
  ACTION: perception verb,
  Y: location with a narrow passage
rept.: X : through(loc or temp, Y)
French synset:
    {a travers, au travers de, dans}
example: Jean regarde a travers la grille
     / dans les jumelles.
```

(litteral translations of examples: 1: John comes in by the door, 2: John looks through the gate / in binoculars). Sense [2] is analyzed as a specific case of [1], a particular stratification of meaning, that corresponds to different language realizations.

Preposition sense [1] has also other groups of strata associated. For example, it can be composite when 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 semantic representation has then an embedded functional structure:

```
X : via(loc, under(loc, Y)).
```

The description of [1] can then be associated with another group of strata as follows:

```
[1.2.1] VIA UNDER - generic
X  <ACTION> [1.2.1] Y
'An entity X moving via under Y'
X: concrete entity,
ACTION: movement verb,
Y: location with a passage under it
representation: X : via(loc, under(loc,Y))
French synset: {par dessous}
example: Jean passe par dessous le pont.
[1.2.2] VIA ABOVE - generic
   etc.
```

The second example below shows, for the abstract notion of 'front position', a distinction made on the semantic domain of the argument: localization on the one hand and psychological or epistemic on the other. This distinction is motivated by the emergence of two very distinct senses, characterized by two different synsets (linguistic realizations) and constraints in the frame. We have here two representations at the same level of abstraction. We then say that this abstract notion is polymorphic. Representation is as follows:

```
[3] : FIX LOC - in front of object
'An entity X located in front of
          another object Y'
X <ACTION/STATE> [3] Y
  X: concrete entity,
  ACTION/STATE: position verb,
                movement verb,
  Y: object
representation: X: opposite(loc, Y)
synset: {en face de, a l'oppose de}
ex.: Il habite en face de la mairie.

[4] : FIX LOC -
          front of psy or epist object
'Someone X against a law, an idea,
              or a principle Y'
X <ACTION/STATE> [4] Y
  X: human,
  ACT/STATE: psychological or epistemic verb,
  Y: abstract
rept.: X: front(psy or epist, Y)
synset: {contre}
example: Il proteste contre cette loi.
```

(translations of examples: 3: he lives in front of the town hall, 4: he demonstrates against this law).

The third example below introduces another type of stratification motivated by the expression of relations or constraints between arguments, which will determine different lexicalizations. This is, for example, the case for prepositions denoting instruments, as studied in [12]. In this work, we introduced two relations:

- the relation between the actor/agent and the instrument, with 3 levels: Undergo (no control on the instrument or its properties), Select (the actor has some control on the object, but it does not plan to do the action that happens, e.g. like in accidents), Control (the agent has full control on the instrument),

- the relation between the instrument and the Verb-object NP, with 3 levels: Be (the object has some intrinsic properties such that even being passive, it nevertheless participates to the action), React (controlled by the agent for a particular property, the object participates to the action via another property), Act (the instrument fully participates to the action).

In this situation, we consider we have 4 strata (numbered below 5.1 to 5.4), that correspond to different lexicalizations, as illustrated for French:

```
[5] : MANNER - MEANS - Instrument
'Someone X doing an action Y
          using instrument Z.'
X <ACTION> Y [5] Z
  X: human,
  ACTION: verb of change,
  Y: object
  Z: instrument
rept.: X: by-means-of(_, Z)
5.1: +Be +Undergo - synset: {grace a}
5.2: +Be +Select - synset: {par}
5.3: +Select +React - synset: {avec}
5.4: +Act +Control -
         synset: {avec, au moyen de}
```

(litteral translations of the synsets: thanks to, by, with, with + by means of).

As can be noted, PrepNet frames are aimed at being prototypical, with usage constraints based on shallow types, which allows to have a priori a number of exceptions. These frames remain conceptual: we view them as a kind of prelexical level. Frames have been defined from several sources: general semantics considerations (such as thematic roles or semantic categories), dictionaries, and corpus data. Our claim, based on feedback from corpus analysis, is that frames, with the aim of being prototypical, reach a certain level of stability and granularity, from which we can study preposition semantics in more depth, via a stratified approach. Stratification allows for some flexibility when dealing with several languages. Their relevance and usability has been tested by lexicographers, as reported in [2].

3 Populating preposition frames via corpus

The second stage of our work aims at associating, mainly with the lower-level frame strata, a number of corpus observations, so that usage restrictions proper to each preposition under that strata can be further analyzed and described. Exceptions to the general rules can also be observed and reported at this level. Standard usages are essentially treated from corpora observations while low frequency cases are analysed via dictionaries or by substituting a preposition by another one of the same class. An interesting approach, which could be considered here is the one adopted in the preposition project (TPP) accessible at:
`http://www.clres.com/prepositions.html` ,
with the development of a well-designed lexicographic method.

Via a semi-automatic bootstrapping method, we collected corpus occurrences of prepositions. We focussed in a first stage on 14 prepositions (and their different senses): *contre, vers, pour, dans, sous, sur, depuis, apres, autour de, par, des, pres de, aux environs de, aux alentours de* (litt.: against, towards, for, in, on, from, after, around, by, as soon as, near, around) which are among the most frequent ones in French.

Since prepositions have very diverse behaviors, we had to manually analyze them. One of the goals is to promote those uses that correspond the best to the specifications given in the frame. Considering in particular the restrictions on the object Y (the subject X is often much more autonomous), we then introduced a multi-level partitioning of realizations, according to a much more fine-grained set of restrictions, essentially so far on Y. We can then capture a whole spectrum of usage norms. This lower level of description, much more accurate, is called the **realization level**. Each set of realizations that correspond to a usage norm is represented by a kind of synset, where the prototypical preposition(s) is(are) specified with appropriate selectional restrictions; it is also associated with frequency measures.

The final step is to establish links between these norms and the other usages found in corpus. Exceptions are so far just listed when found. For derived uses such as metaphors, dedicated operations such as type coercion (modelled by means of rewriting rules on restrictions) need to be developed. This latter level is called the **derived realization level**.

From the corpus, we extracted usages which are prototypical, leaving the others for a later analysis. The motivation is to make a much more detailed categorization of uses. Categorization is characterized by several types of constraints: semantic types, syntactic constraints such as case, type of the verb, etc. We report below our study of the approximation facet, considered for the 3 main semantic fields in which it is used: time, quantity and location. Prepositions studied are *vers, autour de, aux environs de, aux alentours de*, which all have a sense that expresses the idea of

approximation (around, about, near). So far, these restrictions remain somewhat informal, and the abstract notion studied here is among the simplests. This abstract notion is quite regular: it has very few derived uses and metaphors, and it has a relatively high autonomy w.r.t. the verb. Our aim is here to show the principles of the stratification. More complex cases will be shortly available on our web site at:
`www.irit.fr/recherches/ILPL/prepnet.html`.
Distributions observed are given below in Fig. 1. Fig. 2 shows the restrictions observed for the same notion in Spanish (thanks to Silvia Puig Roura). This notion is quite simple, it just suggest research directions.

Precise place: a place explicitly given (Paris, the campus), by reference: via another point (to the left of, north of). Precise event: event with clear boundaries. Global quantity: a rounded number (e.g. 15 kg opposed to 15.675 kg). As the reader can note, lexicalizations of the approximation notion are relatively homogeneous, with some variations, depending on the semantic field (temp, amount or loc).

To end this corpus study section, let us note that one of our aims is to better capture preposition uses but also to better make explicit their roles in knowledge extraction. For that purpose, we aim at developing a set of annotations schemas. These may be quite diverse: we can annotate separately the preposition-verb or the preposition-NP relations or both in one relation. We can also annotate the frame strata itself, to focus on the semantic contents more accurately. Frequency measures are at this level of much interest to e.g. raise annotation, and, therefore, interpretation, ambiguities.

4 Inferential patterns

Given the description of abstract notions by means of conceptual representations, it is then of much interest to investigate the inferential patterns they may be associated with. Besides the inferential forms proper to the semantics of an abstract notion (e.g. the level of intentionality in notions dealing with instrumentality) and its associated restrictions at the origin of the categorization, a number of patterns operating over sets of semantically related notions can be formulated. Since we are just starting this investigation, we simply report here preliminary ideas. A number of projects in psycholinguistics have elaborated such inferential patterns. Our objective is to evaluate how they can be formulated and used in NLP tasks.

4.1 Non-branching proportional series for the approximation abstract notion

To futher refine and structure the approximation abstract notion (applied to any type of physical value: weights, time, space, money, and to a number of abstract domain values interpreted metaphorically) it is of much interest to integrate the corresponding primitives along a non-branching proportional series, well-known in lexical semantics, used here to provide a kind a partial ordering among notions.

Three primitives are used to organize this notion: Within, Near and Around. According to what we found in tourism corpora, Within is relatively vague and is defined to refer to quite large areas. Near indicates relative proximity while Around both indicates a close proximity and the possibility of something lying all around a reference point. We thus have the structure:
Within $>$ Near $>$ Around.
Consequently, we have the patterns:
$Around \Rightarrow Near.$
$Near \Rightarrow Within.$
This proposal needs further elaboration and testing, in particular from a psycholinguistic point of view.

Furthermore, a few prepositions do not simply denote an area of approximation, but a precise point in such an area. These are represented by the composition of two primitives: At+Near or At+Around. For example, *aux environs de* is represented by:
X : at(loc,near(loc,Z)), when interpreted in the localization domain. More generally, prepositions that denote approximation are realized as follows in French:

abstract notions and their lexicalizations	
abstract notion	lexicalizations
X : around(T, Y)	près de, proche de
X : at(T, around(T, Y))	vers, aux alentours de
X : near(T, Y)	près de, aux alentours de
X : at(T, near(T, Y))	aux environs de
X : within(T, Y)	dans, etc.

T is any appropriate LCS type such as: temp,

Detailed usages for the approximation facet		
frame	nb annotated	restrictions on Y
vers Y(+temp)	45	precise time, date, event
autour de Y(+temp)	12	any form of precise date
aux environs de Y(+temp)	15	any date or event
aux alentours de Y(+temp)	23	any form of date
vers Y(+amount)	55	any global quantity: fare, weight, number, size, except time
autour de Y(+amount)	18	any global form of quantity
aux environs de Y(+amount)	21	same as for vers Y
aux alentours de Y(+amount)	23	same as for vers Y
vers Y(+loc)	74	any precise area, direct or by reference
autour de Y(+loc)	18	any form of altitude or precise place
aux environs de Y(+loc)	32	any location defined directly
aux alentours de Y(+loc)	23	any precise location

Figure 1: Approximation: realization level for French

Detailed usages for the approximation facet in Spanish	
frame	restrictions on Y
cosa de Y(+temp)	precise time, date, durative event
hacia Y(+temp)	any form of precise time, date, durative event
cerca de Y(+temp)	any hour in the day or any event
alrededor de Y(+temp)	any form of date, time or event
cosa de Y(+amount)	any global quantity: fare, weight, number, size, and time
hacia Y(+amount)	any global form of number, distance or size
cerca de Y(+amount)	same as for cosa de Y
alrededor de Y(+amount)	same as for cosa de Y
cosa de Y(+loc)	any global notion of altitude
hacia Y(+loc)	any global form of area, altitude or place
cerca de Y(+loc)	same as hacia de Y, place must be precise
alrededor de Y(+loc)	same as hacia de Y

Figure 2: Approximation: realization level for Spanish

loc, amount, etc. Lexicalizations in the above chart are mainly for illustration purposes, it is clear that notions remain abstract and that preposition choice does heavily depend on the verb and on the arguments, subject, object as well as oblique.

4.2 Inferential patterns in the spatial domain

The spatial domain is particularly rich in inferences. Let us focus here on a relation that structures abstract notions hierarchically (as in taxonomies) with in view some forms of concept relaxation in mind (as analyzed in more depth in Benamara, this workshop). For example, if someone is looking for *a hotel in front of the sea at a certain location*, and if there is none, then he can be proposed, via concept relaxation, *a hotel 'close' to the sea*.

More generally, preposition abstract notions that denote spatial orientation of an object with respect to another object (*adjacent to, in front of, along, against, behind,* etc.) can be generalized to a notion that expresses e.g. close location (*near, close to, in the vincinity of,* etc.). This generalization, which obviously cannot be used in any context, needs to be restricted w.r.t. the objects it applies and their physical properties. Nevertheless, given appropriate domain or general purpose restrictions, we can formulate the following patterns:
X:front(loc,Y) $\Rightarrow$ X:near(loc,Y).
X:against(loc,Y) $\Rightarrow$ X:near(loc,Y) etc.
This second pattern requires that Y does not need X to be against it (like a ladder against a wall would), but is an entity that has its own autonomy in this respect.

A similar situation can be observed with the notion of contact which can be relaxed to near but with no contact. In this class fall generalizations such as: on $\Rightarrow$ above.

5 Perspectives

This preliminary investigation was aimed at identifying difficulties and at organizing the research. The global architecture looks an interesting approach, and the main organization seems to have reached a certain degree of stability: abstract notion definitions seem to be quite stable, but the status of strata needs further investigations. The multi-level approach to language realizations seems a good direction to accomodate the diversity of realizations. It needs however a much larger testing on a number of languages and a more clear method to organize sets of realizations

This preliminary study has, obviously, a number of perspectives. Our first aim will be to develop in depth preposition descriptions for French, followed by multilingual work, on Spanish and Catalan, and then on English and German. An idea to test is to have a kind of open system, where, following guidelines, linguists can enter descriptions and uses of the prepositions of the language they study. Another aspect is to make data accessible in a variety of ways via the Internet.

At a more theoretical level, we plan to study in more depth the relations with the verb, for which we have developed quite a lot of descriptions in the past for French. Obviously, it is of much interest to investigate relations with VerbNet and FrameNet. An interesting point is also compositionality, since some PP may have wider scope over the verb in semantic representations. Another point is the use of inferential patterns in various applications such as question-answering and data integration. Finally, also of interest are the integration, via lexical descriptions, verb-particle constructions [18] and collocations.

Acknowledgements I thank Silvia Puig Roura for her contribution, in particular for the Spanish data.

References

Baker, M.C., (1988), *Incorporation: A Theory of Grammatical Function Changing*, Chicago University Press.

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., (1993), *Machine Translation, a view from the lexicon*, MIT Press.

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.

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

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

Mari, A. (2000), *Polysémie et Décidabilité. Le cas de avec ou l'association par les canaux*, Thèse de Doctorat, EHESS, Paris, collection Langue et Parole, L'Harmattan, 2003.

Pesetsky, D., (1982), *Paths and Categories*, MIT doctoral dissertation.

Pustejovsky, J., (1991), *The Generative Lexicon*, Computational Linguistics, vol. 17, MIT Press.

Pustejovsky, J., (1995), *The Generative Lexicon*, MIT Press.

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., (2005) *Verb-particle Constructions in the WWW*, in P. Saint-Dizier (ed), *Syntax and Semantics of Prepositions*, Kluwer Academic, to appear.

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.