The VOLEM Project : a Framework for the Construction of Advanced Multilingual Lexicons

Ana Fernandez
Universitat autònoma de Barcelona
08193 Bellaterra, Spain
Ana.Fernandez@uab.es

Gloria Vazquez
universitat de Lleida, Pl. Victor Siurana, 1
25003 Lleida, Spain
gvazquez@dal.udl.es

Patrick Saint-Dizier, Farah Benamara, Mouna Kamel
IRIT, 118 route de Narbonne
31062 Toulouse cedex, France
stdizier@irit.fr, benamara@irit.fr, kamel@irit.fr

Abstract

We report in this short document the results of a Regional European project carried out on Spanish, Catalan, Occitan and French whose aim is to design a lexical knowledge base where syntactic and semantic descriptions have been normalized and are treated in a uniform way cross-linguistically. Besides the scientific aspects, one of the aims is to make less developed languages such as Occitan or Catalan accessible on the WEB to a larger audience.

1 Introduction

We report in this short document the results of a Regional European project carried out on Spanish, Catalan, Occitan and French whose aim is to design a lexical knowledge base where syntactic and semantic descriptions have been integrated into a uniform cross-linguistic framework. The goal is to establish syntactic and semantic correspondences between major and minor languages of the Pyrenees area, namely Spanish and French on the one hand and Catalan and Occitan on the other. One of the goals is to contribute to the integration of less developed languages in electronic communication and in machine translation.

The different groups involved in the VOLEM project had carried out in the past a number of descriptions for verbs and prepositions related to their syntactic behavior and to their semantic representations. One of the first stages of the project was to design a normalized and homogeneous representation that can accommodate most of the features relevant to each language. Although these languages are all Romance languages, it turns out that they exhibit major differences that make normalization a really complex task. For example, French requires a subject in any finite construction whereas Spanish and Catalan do not. Similarly, clitic formation and resultative stative constructions are quite different. This problem is known as the divergence problem. An in-depth study of diverges between Hindi and English is analyzed in [3] while a number of divergences between English, Spanish and German are reported in [4] and [5].

In this paper, we first present a number of methodological issues and then concentrate on the structure of the multi- lingual lexical knowledge base and on its implementation. So far, the description of about 1700 of the most common verb senses for each of these languages have been coded, while representations of 145 prepositions and prepositional constructs are available for French and under way for the other languages

2 Some methodological considerations

When one tries to analyse the needs of various research teams in terms of lexical resources, it turns out that they are very important, quite diverse, but rarely precisely identified, in particular for semantics, but this is also true to a lesser extent for syntax. If we go deeper, we then realize that important properties such as the granularity of descriptions, the theoretical anchoring and the data representations remain vague or refer to very clas-

sical structures such as feature structures. The conclusion is that it is the role of the resource designer to propose a coherent model for resources, with a clear theoretical background, and then to define lexical insertion rules whose role is to filter and to reformat data depending on the task considered, on the user's preferences, and theoretical and practical orientation.

2.1 The problem of sense delimitation for polysemic units

Let us now review a few questions we had to solve in this project. The first one is related to the abstract notion of word- sense and how a polysemous lexeme can be decomposed into several, more or less independent senses, see [19] and [20] among many others. Each sense needs to be treated separately, even if sense share some syntactic and semantic characteristics. A polysemic verb such as *cut* has at least 4 senses (physical separation, interruption of a process, etc.). These senses share a quite large set of semantic elements, e.g. the interruption of something (physical integrity, aspectual integrity, process integrity, etc.), but they also differ in a number of others, which need to be major characteristics. Some senses are narrow, while others are much larger. Some may be derived from more basic ones via metonymy, but got an independent status, e.g. by the virtue of frequent uses and autonomy. As can be noted, the picture is quite complex. Syntactic characteristics must exhibit major differences, like a different number of arguments, or arguments of very different semantic types.

The next question that arises concerns the nature and the form of the constraints used to delimit senses and to represent them. In particular, what are the meaning components at stake (e.g. causality, aspectual features, incorporated manners, etc.), and what are the syntactic criteria that helps making sense distinctions. The next question concerns senses and the relations between their primitive and derived uses, via metaphors [16], [17] or other transformations.

2.2 Dealing with multilingual aspects

In a multilingual environment, it is essential to establish relations (usually many-to-many) between senses which show a coherent level of decomposition. This problem is, for example, visible, and quite well resolved at a lexicographic level, in EuroWordNet [7]. Our approach considers much larger sense units and is therefore more decompositional, in terms of semantic unit identification, than EuroWordNet: we postulate the existence of a much smaller number of senses, which can undergo

different variations such as metaphors, sense shiftings, and metonymies. From our perspective, the relation with syntax and with conceptual representations is much easier to handle [15].

The counterpart is that we have to model the various forms of sense variation (metonymy, metaphor, etc.), but the gain is clearly a higher level of generalization, of linguistic and cognitive adequacy and of explanations of the phenomena.

2.3 Granularity of representations

Besides the obvious problem of sense delimitation, where we need to establish a strategy, we must, in correlation, investigate the problem of the granularity of the representations, probably constructing representations with different but coherent levels of complexity, depending on the needs. Each level must however be stable, internally coherent and unambiguous, and practically usable. It must also have some theoretical and practical anchoring that confers a real expressive power.

2.4 Designing incrementally coherent semantic representations

The final question, at this stage, is about the nature of the semantic representations, their functionalities and their adaptation to various application frameworks. This is largely a matter of theoretical and practical commitment. Due to a quite long descriptive tradition, we use the Lexical Conceptual Structure (LCS) [13], [14] which seems to be a good framework for representing information of a predicative nature. It can also incorporate different levels of generalizations, and different types of regular metaphors. Its different levels seem to correspond quite well to the different categorization levels observed when a child acquires his mother language.

Finally, from a more technical perspective, the LCS can be implemented in a logical framework that can support under-specification and λ-abstraction, two essential elements to deal with the construction of the semantics of sentences, following the principle of compositionality [22], [23].

2.5 Macro-organisation of the lexicon

At a more global level, the logical and conceptual structure of the lexicon reflects the theoretical vision on which the lexical description is based. If the introduction of ontologies is widely accepted, the introduction of types, underspecification, inheritance mechanisms and

facets are often at the basis of endless theoretical and practical discussions, at both linguistic and computational levels. For example, the numerous problems raised by a direct downward inheritance of properties according to the is-a relation has entailed a large number of ad'hoc constraints (such as blocking) here and there that make lexicons a little bit chaotic. It would have been much preferable to developed a more thorough analysis, identifying semantic markers, which would have better preserved the monotonicity of the macro- structure of the lexicon [2].

3 Syntactic descriptions

Four major types of information are encoded at the syntactic level: argument structure, subcategorization structure, selectional restrictions and alternations.

3.1 Argument structure

Argument structure specifies the number of arguments for each verb sense. This is much more controversial than it may seem at first glance. For example, the instrument in a number of constructs can be considered as an argument or as an adjunct. In:

open the door with a key

the instrument may be felt to be uninteresting, and would therefore fall in the modifier category. However, in:

open the door with a knife

this is no longer the case, and the modifier becomes much more salient, justifying, possibly, its incorporation as an argument.

Similarly, a construction describing a source and a destination can be analysed as having two arguments or a single, complex argument. The simpler case is related to movement verbs, as in:

John is leaving for school

where the target is known, and the source is contextually implicit. In that case, instead of having an argument for the implicit source (reconstructed in the semantic form) and another one for the destination, it seems preferable to have a single, complex argument typed as a trajectory. A final case must be cited for plural arguments, such as the object of verbs like *mix, agglomerate*.

In the VOLEM project, we had a rather conceptual analysis of this phenomena and were quite open to the incorporation of e.g. instruments as true arguments.

3.2 The subcategorization frame and the selectional restrictions

The subcategorization frame follows, where the nature of the arguments (NPs, PPs, S) is specified, with the type of the preposition expected for PPs (e.g. direction) and, whenever possible, the semantic type of each argument using selectional restrictions. Here again, the nature and status of selectional restrictions is subject to debates and it does not exist any norm around the world, in spite of numerous efforts, among which, most notably: EuroWordNet [7], and the ontologies defined at ISI (www.isi.edu) and around the Mikrokosmos project (www.nmsu.edu). So far, restrictions used are borrowed from technical terminologies, but general restrictions are not yet stabilized.

3.3 Alternations

Alternations are syntactic properties of predicates: they describe the way arguments can be arranged, moved, deleted, etc. around the predicate. They are largely language dependent, even if a quite large number of phenomena are shared among languages (like passive). Alternations can be specified at the lexical level, and can be encoded by the production of several subcategorization frames from the original one. They must not be confused with purely grammatical phenomena like Wh-movement or NP-raising, on which there is no lexical restriction a priori.

Alternations are less problematic to define for a given language. The difficulty arises when one tries to define a common set of alternations, as comprehensive and appropriate as possible from different languages. Originally, the Spanish groups had a quite abstract and semantic vision of alternations while the French team followed B. Levin's work [18] on English in a stricter way [21], with the elaboration of 51 alternations. To give an idea, among these, about 32 are shared with English.

The work on Spanish and Catalan alternations was organized around aspectual considerations: (1) opposition in the conceptualisation of two events and (2) aspectual opposition. While (1) allows for a detailed treatment of change of focus constructions, under-specification and infraspecification, (2) results in a clear distinction between resultative and middle constructions. These alternations were conceived as the phrasal expression of semantic contrasts, postulated as being cross-linguistically valid.

Our first task was then to establish a common set of alternations that can accommodate both visions of grammatical constructions. The next task was to establish

a comprehensive set of alternations. A large sample is given below (unusual constructions are ommitted). Main notations are as follows: caus: causative construction (agent subject, roughly), state: state verb with different realizations depending on the language, pas: passive (with an auxiliary like 'ser'), 2np indicates two argument NPs (one of which is the subject), pro denotes a pronominal construct, and faire (hacer) is a semi- auxiliary in Romance languages used e.g. to introduced agents. Alternations are represented as declarative structures, derived from the subcategorization frames; examples in Spanish or French are given and are followed by approximate glosses in English:

1. **Causatives:**
   caus-np: *Mara anda* (Mary walks)
   caus-2np: *Mara envìo un paquete, Fr: Marie envoie un paquet* (Mary sends a parcel)
   caus-np-pp: *Mara anda por el campo* (Mary walks on the road)
   caus-2np-pp:*Mara enva un paquete a Juan, Fr: Marie envoie un paquet à Jean* (Mary sends a parcel to John)
   caus-np-2pp: *Mara conversa con su profesor sobre poltica*

2. **Causatives including the semi-auxiliary faire / dejar/ hacer / ser...:**
   the semi-auxiliairy is noted as 'aux' to avoid any language dependent representation. **caus-aux-inf-2np:** *Ese comentaro hizo enfadar a Juan* (this comment made John laughts)
   caus-aux-inf-2np-pp: *La bruja hizo que el prncipe se convirtera en rana, Fr: La fée a fait que le prince s'est converti en grenouille* (the magician made the prince to become a frog)
   The semi-auxiliairy faire is used to introduce an agent in basic agentless constructs. These constructions involve infinitive subordinate clauses, similar alternations exist with a finite completive clause, notes as 'compl':
   caus-aux-compl-2np
   Ese commentario hizo que Juan se enfadara
   (That remark made John angry)
   caus-aux-compl-2np-pp
   La bruja hizo que el principe se convirtiera en rana.

3. **Anticausatives:**
   anti-pr-np: *La puerta se ha roto* (the door opened)
   anti-pr-np-pp: *Mara se ha sorprendido con su comentario* (Mary was surprised by his comment)

anti-np: *Los datos han variado* (data changed)

4. **Anticausatives with auxiliary dejar:**
   anti-aux-part-np: *Ha dejado confundida a Mara* ((he) let Mary confused by his comments), this construct is specific to Spanish that allows subjectless propositions.
   anti-aux-part-np-pp: *Ha dejado confundida a Mara con su comentario, Fr.: Il a laissé Marie confuse par son commentaire*
   anti-aux-adj-np
   Ha dejado sucia la cocina
   (He left the kitchen dirty)

5. **Passive with pronominal mark**
   pas-se-np: *Se han enviado los paquetes* ((someone) sent the parcels), this construction, subjectless, is proper to Spanish, in French, the 'empty' pronoun *on* is minimally required.
   pas-se-np-pp: *Se han subido las maletas al altillo*

6. **Passive with auxiliary**
   pas-aux-part-np: *Fueron enviados los paquetes* (parcels were sent)
   pas-aux-part-np-pp: *La tarta fue repartida entre los invitados* (the tart was shared among the guests)

7. **Resultative States**
   resul-aux-part-np:
   La mesa està limpia
   result-aux-part-np-pp:
   El sobre ya està enviado a su destino
   (The envelope has already been sent to its destination)
   result-pr-adj:
   está sucio
   (it is dirty)

8. **Reflexive:**
   refl-pr-np: *La niña se peinó* (the girls combs herself)
   refl-pr-2np: *La niña se peinó el pelo* (the girls combs herself the hair)
   refl-pr-np-pp: *Juan se vendió a su enemigo* (John sold himself to his ennemies)

9. **Reciprocal:**
   rcpr-pr-np: *Juan y Ana se escribieron* (John and

Ann write each other)
rcpr-pr-2np: *Juan y Ana se escribieron una carte*
(John and Ann write each other a letter)

10. **Stative constructions**
    state-2np: *Juan tiene un libro* (john has a book)

11. **Small Clause introduction**
    caus-2np-pred: *Jean nomme Edith ministre* (John appoints Edith minister)

12. **Impersonal with "il" and "there"**
    imp-inch-np-pp :
    Il apparait un bateau à l'horizon
    (It appears a boat at the horizon), The impersonal pronoun 'il' (it) is specific to constructions in French: it is used to introduce an empty subject, since French requires a subject in any finite clause.

13. **Diathesis specific to Spanish**
    These are related to subjectless passive forms.
    pas-pr-pp:
    Se ha conversado con los dirigentes
    ((someone) talked to the leaders)
    pas-pr-2np:
    Se ha bajado al enfermo a la primera planta
    (The patient has been taken to the first floor)
    pas-pr:
    Se Mintio
    (Someone lies)

The merging of the French, Catalan and Spanish alternations is not just the union of them. We had, roughly, the following different situations:

- similar, or almost similar constructions were kept intact, with minimal generalizations,

- alternations proper to only one language were kept as such, There are not so many, but another criteria worth considering is the contrastive frequency of use,

- groups of French alternations were covered by a single Spanish alternation or vice-versa, in that case, the generic alternation was kept, possibly with some restrictions (such as: selectional restrictions, thematic role distributions, aspectual factors, etc.)

- some other groups of alternations, like the anticausatives have been completely reformulated,

making a compromise between the need to stay at a syntactic level (French system) and to incorporate semantic aspects (Spanish and Catalan systems).

Some minor cases have been left out while some others have been generalized (like the use of semi-auxiliaries: *faire, hacer, dejar*). The result is of much interest and matches perfectly with the expectations of each language and partner. We feel that these alternations should also be appropriate for a number of other languages, at least the Romance family. It would be of much interest to evaluate the overlap with other languages like the Germanic languages, or the different language families in India [3]. There are obviously more or less 'idiosyncratic' constructions, specific to a small number of languages which cannot be re-used.

In addition and in correlation, we have analysed the semantic contents of these alternations [11] : a construction is in general not neutral and it conveys some semantic contents, not included in the basic subcategorization frame. For example, passive constructions change the focus, the conative construction makes the result of the action uncertain (*to cut at the meat*), etc.

The above set of alternations can be used in a number of theoretical frameworks (HPSGs, TAGs) or applied ones. An implementation in TAGs is being carried out using the TAG approach and a meta- grammar compiler, within the INRIA GeNI project (at www.loria.fr) This is of much use, e.g. in knowledge extraction where it is essential to clearly identify the nature and the position of the different arguments.

4 The syntax semantics interface

Our goal being to represent meaning, we now investigate different forms of semantic representations. The simplest is the assignment of thematic roles [10], [6] to predicate arguments, and possibly to modifiers if one wants to tag them in texts, since they may convey essential information. This first level of semantic representation can be viewed as a model of the world, with agents, directions, means, manners, etc. We allow several roles to be assigned to each argument. For example, the subject of give, is both the agent and the source of the transfer of possession. The same strategy as for alternations has been applied here to build a common set of thematic roles, which are in our project:

- *Initiators*: *agents* or *causal theme*, which causes an event to happen, with or without explicit volition.

- *Themes*: undergo actions, they are either *holistic* (not affected) or *incremental (affected) beneficiary*

or victims. At the extreme, we consider incremental themes of *creation* and *destruction*. Finally, we also have a *theme of consequence*.

- *Localizations*: either *spatial, temporal or abstract* at various degrees. We have the classical distinctions between *source, position, direction* (not necessarily reached) and a complex structure: *trajectory* that includes source, via and destination into a single cluster for e.g. verbs of movement.

- Finally, we have isolated roles such as *quantity* (numerical), *accompaniement, instrument: basic or effector* and *identification* (for proper nouns).

In the lexical knowledge base, each verb is a priori assigned a thematic grid (a set of roles for each argument). This is a by default assignment which can vary to a certain extent in concrete situations, depending on the semantics of the argument, in particular, objects. Verbs but also prepositions are assigned a thematic grid. Prepositions are essentially relations, often with two arguments. They relate an argument of the verb (usually the subject) with the NP they dominate [1].

For example, a verb like *eat* has the following description:
EAT: verb, arity 2,
subcat: np(+hum), np(+eatable),
thematic grid: [agent, incremental victim theme]
This verb is simple and rather stable. A verb like *devour* has the same description, except that it undergo several metaphors, such as *devour books/ novels*, where the type shifting between eatable and object with intellectual contents must be accounted for.

A verb like *give* is not as straightforward: it can be assigned a priori the following thematic grid:
[[agent, source], [holistic theme], [incremental beneficiary theme, destination]],
but the beneficiary theme can become a victim if the object given is a punishement. Therefore, the lexical thematic grid is a kind of by-default assignment.

About instruments, the situation is more complex: in:
Ann eats cereals with a spoon,
the spoon is just a tool for Ann to realise the eating, while in:
Mary cuts the bread with a knife,
Mary excerts a certain force on the knife, but it is the knife that does the cutting. In that case, the knife is an instrument of type effector.

5   Using the LCS as a conceptual, semantic representation

We now evaluate the adequacy of the LCS [13], [14], [4], [5] to represent the fundamental meaning of predicative forms. The LCS is basically designed to deal with information with a predicative content. To get a more comprehensive representation for verbs, it is necessary to additionally consider other systems that complement the LCS such as attribute-value pairs, lexical semantics relations and inferences [2], [7].

5.1   Some methodological considerations

In the VOLEM project, we evaluate the use of the use of LCS in lexical resources, and then, in composition in propositions, according to several dimensions:

- its global expressiveness and ease of use in applications, and its linguistic adequacy,

- its ability to be augmented or impoverished for users needing particular forms or willing to encode specific information (e.g. aspect),

- its ability and flexibility to account for sense variations such as metaphors and metonymies,

- its integration into syntactic frameworks,

- its ease of integration with other paradigms, as those cited above,

- its ease of implementation, in particular in logic, using underspecified structures,

- its ease to be combined when processing sentences, according to the compositionality principle.

Points 2, 4 and 5 are reported in [23], they are not developed here.

The LCS verb description is carried out by families, as specified in WordNet [8], [9] : e.g. verbs of possession (subclasses include: transfer, with various directions, exchange, etc.), verbs of movement, verbs of consumption etc. These classes may be quite large, but considering subclasses, with a more specialized meaning, greatly contributes to the homogeneity and to the coherence of the descriptions. Conversely, we can also better identify contrasts between semantically close verbs and make explicit their differences in meaning.

LCS representations have been shown to be a well-designed approach to design interlingua forms [4], [5]. We can thus construct a database of LCS forms and add

a pointer from each verb sense to one of these representations, for each verb in the different databases we have constructed.

5.2 Semantic representations for verbs

We cannot go into all the details of our description method. After several control and evaluation steps, we got a quite satisfactory set of LCS forms. Underspecification (not shown here) has been kept to a reasonable complexity in order to have readable forms. Here is a quite simple but comprehensive example with comments, shown for French and Spanish with some English glosses. All the data are merged for the sake of readability, and some examples are omitted to make the example simpler:

Common part to the three languages:
thematic grid:
$[inic(ag, tc), th]$
(e.g. agent or causal iniciator, theme)

LCS representation (interlingua) to which the verbs point (LCS 37):

$$[_{event} CAUSE([_{thing} I],$$
$$[_{event} BECOME_{+char,+ident}([_{thing} J],$$
$$[_{state} STATE])])$$

LCS: Literally: I (subject) caused an object J to undergo a change of state in its ontological universe, BECOMING (achievement) STATE.

Spanish lexical database:
Spanish verb: cerrar (to close)
Sense number: 75

Alternations + examples:
caus-2np:
El viento cerró las ventanas de golpe
(the wind closed the windows)
caus-aux(hacer)-compl-2np:
El golpe de aire hizo que la puerta se cerrara
(the wind made the door closed)
anti-pr-np:
La puerta se ha cerrado
(someone closed the door)
pas-pr-np:
A las 7, se cerraron las puertas
(At 7, doors are closed)
pas-aux(ser)-part-np:
La caja ha sido cerrada

(the door was closed)
anti-aux(dejar)-part-np:
Ha dejado cerrada la puerta
(He left the door closed)
result-aux(estar)-part-np:
La puerta está cerrada
(the door is closed)

Catalan lexical database:
Catalan verb: tancar (to close)
Sense number: 75
The Catalan uses are very similar to the Spanish ones.
Alternations + examples:
caus-2np:
El vent va tancar les finestres de cop
(the wind closed the windows)
caus-aux(hacer)-compl-2np:
El cop de d'aire va fer que la porta es tanqués
(the wind made the door closed)
anti-pr-np:
La porta s' ha tancat
(someone closed the door)
pas-pr-np:
A les 7 can tancar les portes
(At 7, doors are closed)
pas-aux(ésser)-part-np:
La capsa ha estat tancada
(the door was closed)
anti-aux(estar)-part-np:
La porta està tancada
(He left the door closed)
result-aux(deixar)-part-np:
Ha deixat tancada la porta
(the door is closed)

French lexical database:
French verb: fermer (to close)
Sense number: 75
Alternations + examples:
caus-2np:
Le vent ferme les fenêtres d'un coup
(the wind closed the windows)
caus-aux(hacer)-compl-2np: is not acceptable: * *Le coup de vent fait que la porte se ferme*
(the wind made the door closed)
anti-refl-np:
La porte s'est fermée (d'un coup)
(approx.: someone closed the door)
pas-pr-np:
A 7 heures, les portes ferment
(At 7, doors are closed)

pas-aux(être)-part-np:
La porte est fermée
(the door was closed)
anti-aux(avoir)-part-np:
Il a laissé la porte fermée
(He left the door closed)
anti-pr-np: (middle reflexive)
La porte se ferme facilement
(The door closes easily).

As can be noted, the French alternations are quite different, as expected. Let us now consider a verb with two senses, we can then see that alternations and LCS representations are substantially different:

(1) ahogar: (to kill someone by preventing him from breathing) sense 665
French : étouffer
example : *El asesino ahogó a su vctima*
l'assassin étouffe sa victime.

Alternations for Spanish:
caus-2np:*El asesino ahogó a su vctima* (the murderer suffocates his victim)
anti-pr-np: *La muchacha se ahogó* (the child suffocates)
pas-ser-part-np: *La muchacha fue ahogada* (The child was sufficated.)

thematic grid: [inic(ag), tiv]

LCS representation:

$$[_{event} CAUSE([_{thing} I],$$
$$[_{event} GO_{+char,+ident}([_{thing} J],$$
$$[_{path} BECOME_{+char,+ident}([_{state} DEAD],$$
$$[_{manner} NOT([_{event} GO_{+loc}([-thing AIR],$$
$$[_{path} INSIDE_{+loc}([_{thing} LUNGS])])])])]$$

(2) ahogar: (to oppress, psychological verb) sense 666
French: oppresser
Example: *Este ambiente enrarecido me ahoga / Cette atmosphère raréfiée m'étouffe.*
Alternations:
caus-2np: *Este ambiente enrarecido me ahoga/* (this athmosphere oppresses me)
anti-pr-np: *El atleta se ahogaba*
(the athlet is oppressed).

thematic grid : [inic(tc), tiv]

LCS representation:

$$[_{event} CAUSE([_{thing} I],$$
$$[_{event} BECOME_{+char,+ident,+psy}([_{thing} J],$$
$$[_{state} OPPRESSED])])]$$

The by-default semantic field is +psy (psychological), while the ontological category +char,+ident deals with the evolution of inherent properties of objects.

5.3 Semantic representations for prepositions

The same task is being carried out for prepositions. It is comprehensive for French at the moment (with a total of 85 prepositions and about 50 prepositional phrases). LCS for prepositions are much simpler to write and quite direct, with the use of 65 low-level primitives. Looking at prepositions in parallel with verbs is particularly useful since it allows us to better balance the descriptions between the verb (i.e.: we must include in the verb strictly what is in the verb semantics) and the preposition. Preposition semantics is most useful in a number of applications where e.g. elements such as instrument, manner, localization need to be identified and extracted. In machine translation, the treatment of prepositions is crucial but extremely difficult. A research report can be obtained from the authors.

For example, the representation of *contre* (approximately 'against') can be organized as follows:

- The direct usage is a physical object positioned against another one:
  $\lambda K [_{place} NEXT-TO_{+loc,c:+}([_{thing} K])]$
  where NEXT-TO indicates a physical (+loc) proximity, while contact is encoded by c:+ (Jackendoff 90). The subject I is against K, and the agonist force exerted by I on K is balanced by the antagonist force exerted by K on I. The physical contact is the most visible; it is in the foreground, while the forces view (or facet) is rather in the background. In French, our analysis is that *contre* describes a position, not a path. It is important to note that the idea of movement, if any (as in: *push the chair against the wall*) comes from the verb, not from the preposition.

- *Contre* can also be used to express opposition: *to swim against the current* or metaphorically in the epistemic or psychological domains as in: *to argue against a theory/ a practice*. The primitive AGAINST is used to capture the fundamental idea of antagonistic forces:
  $\lambda K [_{place} AGAINST_{+psy \vee +epist,c:-,ta:+}$
  $([_{thing} K])]$.
  In that case, the physical contact no longer exists

8

(c:-), while the agonist/ antagonist force is present and in the foreground (noted ta:+, (Jackendoff 90), slightly simplified here).

- *Contre* has also a sense that expresses notions like 'goal for a certain protection or defense': *medecine for cought*. It is represented as follows:
  $\lambda\,X\,[_{event}\,FOR_{+char,+ident}([_{event\vee thing}\,X\,])]$

- Then we have the notion of exchange: *I substitute my hors d'oeuvre against a desert*, representation is as follows:
  $\lambda\,X,\,\lambda\,Y\,[_{path}\,EXCH_{+poss}([_{thing\vee event}\,X\,],$
  $[_{thing\vee event}\,Y\,].$

- Finally, we have the senses related to the expression of the ratio or the proportion: *9 out of 12 came*:
  $\lambda\,X\,[_{amount}\,AGAINST_{+comm}([_{amount}\,X\,])].$

In addition, prepositions receive a thematic grid. For example *contre* has the following grid: [theme, position].

6   The database

The database architecture is organized as follows: each language has its own language dependent database, while a number of multilingual resources, common to all languages, are stored in a dedicated, interlingua database.

In itself the databases have nothing deeply original. Our goal is simply to make the data accessible on the WEB for users that wish to learn or just know something precise about the languages studied.

We have a database for each language that contains everything proper to the language: the subcategorization frame, the selectional restrictions, the preposition type, if any, and the alternations. The semantic data : thematic grids, the LCS and the main WordNet semantic classes are shared and are stored in a separate database.

Each verb sense in the language dependent databases points to an LCS, a thematic grid and a WordNet class via a dedicated key. This allows us to factor out common representations to several verbs in each language and over several languages. These language independent database is an interlingua database [3], [4]. It is organized around the verbs of the 3 languages coded so far. A sense.subsense number is assigned to each entry, and the lexicalization is given for each language. An entry corresponds to the maximal sense distinction decomposition w.r.t. the languages studied. If there is more than one translation in one or more of these languages, corresponding most probably to a more subtle sense distinction than in the other language(s), then entries are duplicated, with different subsense numbers,

to avoid confusions, since translations are in general many-to-many relations. For example we have sense 630 (to approach from something) that translates in Spanish as *acercar*. It is decomposed into 3 subsenses since Catalan has 2 lexicalizations, with slight differences in meaning that Spanish does captures by a verb (but by an adverb), these are: *apropar, atansar*. Similarly for French we have two corresponding lexicalizations: *approcher, rapprocher*. *Atansar* translates either into *approcher* or *rapprocher* while *apropar* only translates into *approcher*. These latter verbs have slightly different LCS and possibly alternations. We then have the following structure in the interlingua database:
630.1 acercar, apropar, approcher
630.2 acercar, atansar, approcher
630.3 acercar, atansar, rapprocher.

To establish the correspondences, each entry of a verb for a given language starts by a list of sense.subsense numbers. There are several elements in that list when a verb corresponds to several subsenses.

One may criticize this decompositional approach, arguing that it my become very large when new languages are added, but we think that it has a quite reasonable limit, quite rapidly reached, since our 'standard' languages do not exhibit so many subtle sense distinctions, with the granularity level that we consider in this generative approach. Results would be quite different (see EuroWordNet) with a lexicographic view.

In the database, data is represented using XML, and an interface is been created to query the databases using xslt. In a first stage, a simple interface has been developed in html. We plan to have a simpel interface, with basic queries offered for casual users, and a more elaborated interface for 'professionals' who want to go deeper into the informational structures, make statistics on any type of data over the languages and create his own lexical database from the data he wishes to use. Access to the databases should be available shortly, http paths can be obtained from the authors.

7   Conclusion

The database with the comprehensive description of about 1700 verb forms for Spanish, Catalan, French, and to a lesser extent Occitan will be available soon. The main aim of this project was to define common frameworks to represent syntactic as well as semantic and conceptual information in an homogeneous way from a priori different previously established experiences and descriptions. English is essentially used to provide pointers

and glosses. This work is original in the sense that such a detailed syntactic and semantic description has never been reached in the past, especially within multilingual environments.

8 References

[1] E. Cannesson, P., Saint-Dizier, "Defining and Representing Preposition Senses: a preliminary analysis", ACL workshop on WSD, Philadelphie, July 2002.

[2] Cruse, A., *Lexical Semantics*, Cambridge university Press, 1986.

[3] S. Dave, J. Parikh, P. Bhattacharyya, "Interlingua Based English Hindi Machine Translation and Language Divergence", to appear in *Journal of Machine Translation*, v.17, 2002

[4] B. Dorr, M. Katsova, "Lexical Selection for Cross-Language Applications: Combining LCS with Word-Net", 3rd conf. Machine Translation, Lahorne, PA, 1998.

[5] B. Dorr, "Large-scale Dictionary Construction for Foreign Language Tutoring and Interlingual Machine Translation", *Journal of Machine Translation*, 12(1), 1999.

[6] Dowty, D., "On the Semantic Content of the Notion of Thematic Role", in G. Cherchia, Dowty, D., Thematic Proto-roles and Argument Selection, *Language*, vol. 67-3, 1991.

[7] EuroWordNet, *general document, version 3*, P. Vossen (ed.), univ. of Amsterdam, 1999.

[8] C. Fellbaum, "English Verbs as Semantic Net", *Journal of Lexicography*, vol. 6, Oxford University Press, 1993.

[9] Fellbaum, C. . "A Semantic Network of English Verbs", in C. Fellbaum (ed.) *WordNet: An Electronic Lexical Database*, Cambridge, MA, MIT Press, 1997.

[10] C. Fillmore, "The Case for Case", in *Universals in Linguistic Theory*, E. Bach and R.T. Hams (eds.), Holt, Rienhart and Winston, New York, 1968.

[11] Goldberg, A., *Constructions: A Construction Grammar Approach to Argument Structure*, University of Chicago Press, 1994.

[12] Gruber, J., *Studies in Lexical Relations*, MIT doctoral dissertation and in Lexical Structures in Syntax and Semantics, North Holland , 1967.

[13] Jackendoff, R., *Semantics and Cognition*, MIT Press, Cambridge, 1983.

[14] Jackendoff, R., *Semantic Structures*, MIT Press, 1990.

[15] Gayral, F., Saint-Dizier, P., "Peut-on couper à la polysémie verbale ?", *actes TALN 99*, Cargse, 1999.

[16] Lakoff, G., Johnson, M., *Metaphors we Live by*, Chicago University Press, 1980.

[17] Lakoff, G., Johnson, M., *Philosophy in the Flesh*, Basic Books, 1999.

[18] Levin, B., *Verb Semantic Classes: a Preliminary Investigation*, Chicago University Press, 1993.

[19] G. Numberg, "Transfer of Meaning", *Journal of Semantics*, vol. 12, 1995.

[20] Pustejovsky, J., *The Generative Lexicon*, MIT Press, 1995.

[21] P. Saint-Dizier, "Alternations and Verb Semantic Classes for French", in *Predicative Forms for NL and LKB*, P. Saint-Dizier (ed), Kluwer Academic, 1998.

[22] P. Saint-Dizier, "Underspecified Lexical Conceptual Structures for Sense Variations", *Workshop on Lexical Semantics IWCS*, Tilburg, 1999.

[23] P. Saint-Dizier, G. Vazquez, "A Compositional Framework for the Semantics of Prepositions", *IWCS4*, Tilburg, 2001.