# 4LEX: a Multilingual Lexical Resource

**Montserrat Civit**
CLiC, Univ. of Barcelona
Barcelona, Spain
mcivit@ub.edu

**Roser Morante**
Tilburg University
Tilburg, The Netherlands
R.morante@uvt.nl

**Antoni Oliver**
UOC
Barcelona, Spain
aoliverg@uoc.edu

**Joan Castelví**
CLiC, Univ. of Barcelona
Barcelona, Spain
joan.castellvi@ub.edu

**Joan Aparicio**
CLiC, Univ. of Barcelona
Barcelona, Spain
juan.aparicio@thera-clic.com

## Abstract

As the creation of computational (verb) lexicons is a huge time-consuming task, tagged corpora appear to be a very useful resource for inducing verb knowledge. In this paper we present a multilingual verb lexicon with syntactic and semantic information for four languages. For three of them (Catalan, Basque and Spanish) this lexicon is induced from syntactically and semantically annotated corpora created at the 3LB project; for Russian, the lexicon will be created manually. From this information, the correspondences between syntactic functions and semantic roles among the four languages will be set up.

## 1 Introduction

In this paper we present a methodology for creating a new multilingual lexical resource, 4LEX[1], a verb lexicon for Catalan, Spanish, Basque, and Russian. The lexicon will contain the information shown in table 1.

Two different methodologies will be used to obtain each monolingual lexicon: the Catalan, Spanish, and Basque lexicons, that will conform the 3LB-LEX resource, are obtained from syntactically and semantically annotated corpora that have been developped in the 3LB Project funded by the Spanish Government (Palomar et al., 2004). The Russian lexicon, that will result into the RUS-LEX resource, will be created manually following the same guidelines as the others.

| | Catalan, Spanish, Basque | Russian |
|---|---|---|
| morphological | – | case |
| syntactic | constituents and functions | functions |
| semantic | thematic roles, wordnet synsets | thematic roles |

Table 1: Information in 4LEX.

The goals of the 3LB project were to annotate three corpora (Catalan, Spanish, Basque) with semantic and syntactic information. As for the syntactic information, the Spanish and Catalan corpora were annotated with constituents and functions, whereas the Basque corpus was annotated with dependencies. The syntactic functions used for the annotation are: subject, direct object, indirect object, predicative, agent, prepositional complement selected by the verb, adverbial, and attribute. The annotation maintains the superficial order of constituents, and the quality of the annotation process is guaranteed by the inter annotator agreement tests that were performed (Civit et al., 2003) .

The semantic annotation consisted in assigning only one EuroWordNet synset to every noun, verb, and adjective. For this task a steady version of the Spanish, Catalan, and Basque versions of EuroWordNet was used

In Table 2 we show the size of every corpus. It

[1]This work is partially funded by the project CESS-ECE (HUM2004-21127-E).

has to be taken into account that Basque is an agglutinative language, and that many words, especially prepositions and articles, are suffixes attached to the stem.

| | words | sentences |
|---|---|---|
| Spanish | 100.000 | 4.000 |
| Catalan | 100.000 | 2.600 |
| Basque | 50.000 | 2.750 |

Table 2: Size of the 3LB corpora.

Concerning Russian, a syntactically annotated corpus is not available. The creation of RUS-LEX will be done manually, using a morphologically annotated corpus as a reference (Oliver, 2004). The corpus size is more than 16 million words.

The annotation of semantic roles in 4LEX will be done following the proposal of the PropBank project (Palmer et al., 2005). In 3LB-LEX the correspondence between syntactic functions and semantic roles of each verb sense will be established from the syntactic schemata of each verb, which are obtained from the corpus. The annotation of the corpus with semantic roles will be done automatically with a post manual validation. In the case of Russian, thematic roles along with syntactic and morphological (case) information will be annotated manually.

The reasons to create a lexicon of this characteristics are, on one hand, the interest it can have for multilingual studies, since it will allow to perform comparative studies about the syntactic and semantic behavior of the languages in the project plus English. Russian has been selected due to its verbal complexity and richness, in order to carry out contrastive studies on romance, slavic and germanic languages[2]. The compatibility in the annotation of semantic roles will facilitate this type of analysis. On the other hand, it will be possible to use the lexicon for automatic training of semantic–syntactic parsers that can be used for several tasks, like knowledge and information extraction.

In section 2 we present the extraction process of the verb lexicons for 3LB-LEX and the content of the resulting entries. In section 3 we present how we enrich the derived lexicons with information about semantic roles, which will be the basis for creating RUS-LEX. In section 4 we present 4LEX and finally, in section 5 we explore further applications of this resource.

## 2 Extraction of 3LB-LEX

For the extraction of the verb lexicon both sentences and clauses have been taken into consideration, and both verbs in the finite forms and verbs in the non finite forms have been included. In Table 3 we present statistic information of these corpora. The differences between Basque on one side, and Catalan and Spanish on the other are due to the fact that Basque lexicalizes less information in the verb, and it depends more on the complements[3].

| | Spanish | Catalan | Basque |
|---|---|---|---|
| Verb forms | 7.127 | 7.033 | 13.261 |
| Different verbs | 1.070 | 834 | 375 |
| Verbs appearing once | 392 | 304 | |
| Average occur. of a verb | 6,66 | 8,43 | |
| Average occur. of verbs appearing two or more times | 9,93 | 12,69 | |

Table 3: Number of verbs in the corpus 3LB.

The extraction of the verb form has been done taking the tree nodes including simple and compound verb forms, both the ones that belong to compound tenses and the ones that belong to verbal periphrases. In the compound forms the form which contains the lexical information is the last token in the array. The rest of elements of the lexical entry is composed by all the constituents having a syntactic function related to the verb.

**Content of the lexical entry**

For each occurrence of a verb in the 3LB corpora we induce a lexical entry that contains the verb form,

[2]These typological studies are especially interesting for the Department of Linguistics of the University of Barcelona, because it comprises a Section of Slavic languages, as well as Romance languages and Basque.

[3]For now we start from a moderate number of verbs and occurrences, because of the limitations imposed by the resources currently available, but a new project has started that will allow to increase the corpus until 500,000 words for Spanish and Catalan, and 350,000 for Basque. Consequently, the size of the verb lexicon will also increase.

the constituents that co-occur with it, and their syntactic functions. For example, for the occurrence of the Catalan verb *parlar* ('to talk') in the Cat3LB sentence: *En la segona de les conferències, programada per al dia 9 de juny, Robert Brufau parlarà dels edificis en què l'estructura juga un paper decisiu en l'arquitectura a partir d'obres diverses*[4] the entry shown in Table 4 has been induced.

| Func. | Text |
|---|---|
| CC | En la segona de les conferències, programada per al dia_9_de_juny, |
| SUJ | Robert_Brufau |
| verb | parlarà |
| CREG | dels edificis en què l'estructura juga un paper decisiu en l' arquitectura |
| CC | a_partir_d' obres diverses |

Table 4: Induced entry: Catalan *parlar* ('to talk')

As it can be seen in Table 4[5] the surface order of the syntactic functions is kept, so that it is possible to make studies about word order in relation to syntactic functions (subject and complements)[6].

Tables 5[7] and 6 show two entries induced from the Cast3LB corpus for two occurrences of the Spanish verb *mover* ('to move') that has two different frames (clauses with the verb *mover* appear in bold in the sentences).

The sentence in which *mover-1* appears is: **0* se enfocó desde muy antiguo hacia el transporte de vivos y bienes, del que *0* tenía ejemplo en las naves* **que movía el viento por aguas y mares con estimable eficiencia**[8].

The sentence in which *mover-2* appears is: *Este último partido entre dos equipos que una vez más están entre los mejores clasificados de las votaciones semanales de entrenadores y periodistas especializados – no hay clasificación oficial al no existir una liga nacional universitaria – resultó ser de los* **que mueven a la reflexión**[9].

| Func. | Text |
|---|---|
| CD | que |
| verb | movía |
| SUJ | el viento |
| CC | por aguas y mares |
| CC | con estimable eficiencia |

Table 5: Induced entry: Spanish *mover-1*

| Func. | Text |
|---|---|
| SUJ | que |
| verb | mueven |
| CREG | a la reflexión |

Table 6: Induced entry: Spanish *mover-2*

From the sentence in Eus3LB *Alderdiko kideei ez diegu agindu bat eman*[10] we have induced the entry for the verb *eman* ('to give') that appears in Table 7[11].

| Func. | Text |
|---|---|
| ncsubj | pro (elipsis) |
| nczobj | alderdiko kideei |
| verb-aux | ez-eman-diegu |
| ncobj | agindu bat |

Table 7: Induced entry: Basque *eman*('to give')

The 3LB lexicon, apart from syntactic information, will contain for each verb entry the synset of EuroWordNet. In this way, for those entries that are widely represented, this lexicon will allow to check if there exists variation in the syntactic–semantic

[4]'In the second conference, programmed for the 9th of June, Robert Brufau will talk about buildings in which the structure plays a decisive role in the architecture, based on several building works'.

[5]The meaning of the syntactic function tags is: CC for adverbial complement, SUJ for subject, CREG for prepositional complement selected by the verb.

[6]Since in the 3LB corpora nodes have been added to the tree for the elliptic subject, this also appears in the extracted entries.

[7]CD is the tag for Direct Object.

[8]'Since ancient times it was directed towards the transportation of beings and goods, from which (he/she) had an example in the ships **that the wind used to move by waters and sees very efficiently**'.

[9]'The last match between two teams that once again are among the better classified in the weekly voting of trainers and specialized journalists – there is no official classification because it does not exist a university national league – turned out to be one of those **that cause reflection**'.

[10]'We did not give any order to the members of the team'.

[11]The meaning of the syntactic labels is: *ncsubj* for non sentential subject; *nczobj* for non sentential indirect object;*verbo–aux* for verb and auxiliary ;*ncobj* for non sentential object.

schemes for the different senses. Thus, this resource extends the information in EuroWordNet by adding syntactic–semantic schemes.

Since the 3LB corpora are annotated with complete syntactic trees including the already mentioned syntactic functions, the initial extraction of verb lexical entries with their corresponding arguments has been carried out in a completely automatic way.

Next, we present how the lexicon will be enriched with semantic role information.

## 3 Enrichment with semantic roles

From the induced entries presented in section 2 we will associate the corresponding semantic roles to each function, for each sense of each verb. This will be the basis to create RUS-LEX.

Table 8 presents the semantic roles that will be used in the semantic annotation process. The roles are equivalent to the ones used in the PropBank project (Palmer et al., 2005; PropBank Annotation Guidelines, 2002) and their use in 4LEX has the following motivation:

1. Applicability: this set of semantic roles has been defined for corpus annotation, and has been tested in the annotation of the PennTreeBank for the constitution of the PropBank. The semantic roles have been obtained from the analysis of verbal occurrences in a corpus and, consequently, they have been checked against a broad set of examples.

2. Flexibility: the annotation system is carried out both at the argumental and semantic role level. Thus, each argumental position can be specified with different semantic roles, and the other way round, one specific semantic role might appear in different argumental positions. So, it is possible to preserve, on the one hand, the degree of proximity of an argument in relation to the verb (this is indicated by numbering the arguments), and, on the other hand, the type of semantic relation that each argument establishes with the verb.

3. Standardization: this type of annotation allows direct comparison with PropBank and makes it possible to establish equivalencies with other proposals. As soon as 4LEX lexical entries are available, it will be possible to define the links between them, and to develop comparative studies for the following languages: English, Catalan, Basque, Spanish, and Russian.

The role tagset appears in Table 8. The numbered arguments are the ones that occupy argumental positions, while *ArgM* corresponds to adjuncts or more marginal elements in relation to the verb. *ArgA* is used in PropBank to indicate *volitional motion* as in sentences: *Upenn works John hard* and *Mr. Dinkins would march his staff out of board meetings*.

| **Role** | **Thematic roles** |
|---|---|
| Arg0 | agent |
| Arg1 | theme (TEM) / patient (PAT) |
| Arg2 | benefactive (BEN)/ instrument (INS) / atribute (ATR)/ final state (EFI) extension (EXT) |
| Arg3 | departure point (PDP) / benefactive (BEN) / instrument (INS) / atribute (ATR) |
| Arg4 | final point (PDD) |
| ArgMs | verb adjuncts |
| ArgA | external cause of an action |

Table 8: Semantic roles

The correspondence between syntactic functions and PropBank arguments appears in table 9.

| **Role** | **Syntactic functions** |
|---|---|
| Arg0 | subject (SUJ) / agent compl. (CAG) |
| Arg1 | direct object (CD)/ |
| | prepositional compl.(CREG) / |
| Arg2 | indirect object (CI) /adjunct (CC) |
| | prepositional compl. (CREG)/ |
| Arg3 | adjunct (CC) |
| Arg4 | adjunct (CC) |
| ArgA | subject (SUJ) / agent compl. (CAG) |

Table 9: Functions vs Roles

Table 10 presents the specification for ArgM or verbal adjuncts that we use for marking the 4LEX lexicons. Information relative to the negation (NEG), discourse connectors (DIS), and modal verbs (MOD) in the functioal tagging of ArgM appears in 3LB in the syntactic annotation; this is why

it will be omitted. The same criteria applies for Russian.

| | |
|---|---|
| LOC: place | CAU: cause |
| EXT: extension | TMP: time |
| FIN: goal | MNR: manner |
| ADV: general | DIR: direction |

Table 10: Specification of ArgMs

From the analysis of all occurrences of a sense of each verb we will generalise the information, and we will infere prototypical verb entries that will be used in the definition of the multilingual relationships. The same kind of entries will be created for Russian by checking occurences in a morphologically annotated corpus.

The mapping of syntactic function into argument-thematic roles will be done manually for each verb entry. For example, after the analysis of all occurrences of the verb *parlar* in Catalan, we build the following entry:

| Parlar | |
|---|---|
| SUJ | Arg0 |
| CREG | Arg1 |
| CC | Arg2 |
| CC | ArgM-TMP |
| CC | ArgM-ADV |

With this prototypical verb entry we will annotate later all the occurrences of this verb in the corpus. Since there is ambiguity in the correspondence between functions and thematic roles (CC for this verb), there will be a manual postprocess for disambiguating all those cases. The final result will be:

| | |
|---|---|
| CC-ArgM-TMP | En la segona de les conferències, programada per al dia_9_de_juny , |
| SUJ-Arg0 | Robert_Brufau |
| REL | parlarà |
| CREG-Arg1-de | dels edificis en què l'estructura juga un paper decisiu en l' arquitectura |
| CC-ArgM-ADV | a_partir_d' obres diverses |

The prototypical entries for the verb *mover* will be:

| mover-1 | *change of position* |
|---|---|
| SUJ | Arg0 |
| CD | Arg1 |
| CC | Arg2 |
| CC | Arg3 |
| CC | Arg4 |
| CC | ArgM-LOC |
| CC | ArgM-MNR |

| mover-2 | *obliged action* |
|---|---|
| SUJ | Arg0 |
| CD | Arg1 |
| CREG | Arg2 |

From this entries we will annotate the occurrences of the verb:

| | |
|---|---|
| CD-Arg1 | que |
| REL | movía |
| SUJ-Arg0 | el viento |
| CC-ArgM-LOC | por aguas y mares |
| CC-ArgM-MNR | con estimable eficiencia |

| | |
|---|---|
| SUJ-Arg0 | que |
| REL | mueven |
| CREG-Arg2-a | a la reflexión |

In the case of Basque, the prototypical entry for the verb *eman* would be:

| Eman | |
|---|---|
| ncsubj | Arg0 |
| ncobj | Arg1 |
| nczobj | Arg2 |

And the annotated sentence will be like this:

| | |
|---|---|
| ncsubj-Arg0 | pro (elipsis) |
| nczobj-Arg2 | alderdiko kideei |
| REL | ez-eman-diegu |
| ncobj-Arg1 | agindu bat |

The grid for the Russian sentence *Ivan otkryl dver' kl'učom*[12] would be as follows[13]:

| | |
|---|---|
| $\text{Arg0}_N$ | Ivan |
| REL | otkryl |
| $\text{Arg1-TEM}_A$ | dver' |
| $\text{Arg2-INS}_I$ | kl'učom |

[12] Ivan opened the door with the key

[13] Subindexes stand for case information: $_N$ for Nominative; $_A$ for Accusative, and $_I$ for instrumental.

## 4 4LEX: Multilingual Lexicon

So far the construction of the monolingual lexicons has been described. In this section we present three prototypical cases of comparison between Catalan and Russian, to illustrate the creation of the multilingual lexicon.

In the first case, for a verb such as *obrir*, *otkryt'* (*open*) the argumental and syntactic structures are identical:
Russian sentence: *Ivan otkryl dver' kl'učom*;
Catalan sentence: *El Joan va obrir la porta amb la clau* (English: *John opened the door with the key.*):

| | |
|---|---|
| SUJ | Arg0 |
| REL | (verb) |
| CD | Arg1-TEM |
| CC | Arg2-INS |

In the second case, for verbs such *omplir*, *napolnit'* (*fill*) we find identical argument structure in both languages, but Russian presents a syntactic structure that does not exist in Catalan (the impersonal one). In this case, two syntactic configurations in Russian map to only one Catalan structure.
Russian: *Rabočie napolnili jamu vodoj*
Catalan: *els treballadors van omplir el forat amb aigua* (English: *the workers filled the whole with water*)

| | |
|---|---|
| SUJ | Arg0 |
| REL | (verb) |
| CD | Arg1-TEM |
| CC | Arg2-INS |

The impersonal Russian construction $jamu_A$ $napolnilo$ $vodoj_I$, whose literal traslation is $the\text{-}whole_A$ *filled* $with\text{-}water_I$, and whose syntactic–thematic grid is:

| | |
|---|---|
| CD | Arg1-TEM |
| REL | (verb) |
| CC | Arg2-INS |

does not exist in Catalan. However, the passive sentence *jama napolnilas vodoj* (*the whole was filled with water*) will have the same semantic grid in Catalan: *el forat es va omplir amb aigua.*

| | |
|---|---|
| SUJ | Arg1-TEM |
| REL | (verb) |
| CC | Arg2-INS |

This is only one of the possibilities of impersonal constructions in Russian. Here we only highlight the most representative for the comparison.

Finally, in the third case, that of the verb *recordar*, *vspomnit'*[14] (*remember*) there is no direct mapping between arguments. In Russian the subject of such verb is not an agent but a Experiencer (the passive is not allowed), while in Catalan the subject is indeed the agent.

For Russian the grid would be:

| | |
|---|---|
| SUJ | Arg1 |
| REL | (verb) |
| CC | Arg0 |

In Catalan, the grid is:

| | |
|---|---|
| SUJ | Arg0 |
| REL | (verb) |
| CD | Arg1-PAT |

So, the relationship between arguments should be explicitely declared.

## 5 Conclusions and future research

We have presented 4LEX, a multilingual lexical resource that contains information about the syntactic structure, the argumental structure, and the semantic roles of those verbs that occur in the 3LB corpora and their Russian equivalents. Figure 1 shows the general architecture of the system.

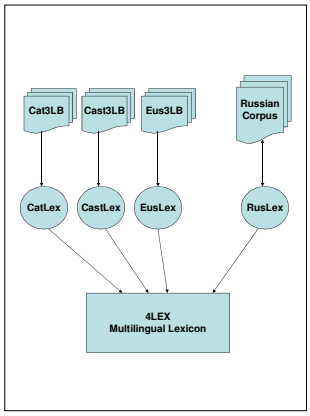


Figure 1: 4LEX

This multilingual lexicon is intended to be a basic resource for the development of Machine Translation systems, IE and IR systems, as well as a source of information for human translators.

[14] We follow (Babby, 1998) in this analysis.

Secondly, the mapping between argument structure and syntactic structure in 4LEX can be used to tag parallel corpora. With this lexicon it is possible to carry out syntactic–semantic annotation of parallel corpora, even if only information for one language is available. Besides, it will also be possible to carry out theoretical studies about the diatheses and their interlinguistic behavior.

Finally, it will be possible to use 4LEX to make interlinguistic comparisons between the languages of the project: we will compare the sematic roles of the same verb in all languages, and how the semantic roles are syntactically expressed. This information will allow us to make inferences about the interlinguistic correspondences between syntax and semantics. This is not a parallel corpora framework but a multilingual lexicon created with the same methodology. Our hypothesis is that languages might share the same argument structure and semantic roles, but they might still differ substantially in how the arguments are syntactically expressed. There is substantial theoretical literature about this topic, which will be interesting to contrast against empirical data. The development of 4LEX guarantees the consistency of the argument and thematic role system tagset. Since the semantic roles match the roles in the PropBank project, it will possible to extend the comparison to English.

## References


L. Babby. 1998. Voice and diathesis in Slavic. *Workshop on Comparative Slavic Morphosyntax.* Spencer, Indiana.

M. Civit, A. Ageno, B. Navarro, N. Bufí, and M.A. Martí. 2003. Qualitative and Quantitative Analysis of Annotators' Agreement in the Development of Cast3LB. *Proceedings of the 2nd Workshop on Treebanks and Linguistic Theories (TLT03), Växjö, Sweden*:21-32.

A. Oliver. 2004. Adquisició lèxica i morfosintàctica a partir de corpus sense anotar: aplicació al rus i al croat. *Tesi Doctoral, Universitat de Barcelona*.

PropBank Project. 2002. PropBank Annotation Guidelines. "http://www.cis.upenn.edu/ ace/".

M. Palomar, M. Civit, A. Díaz, L. Moreno, E. Bisbal, M. Aranzabe, A. Ageno, M.A. Martí, and B.Navarro. 2004. 3LB: Construcción de una base de datos de árboles sintáctico-semánticos para el catalán, euskera y castellano. *Procesamiento del lenguaje Natural* 33.

M. Palmer, P. Kingsbury, and D. Gildea. 2005. The Proposition Bank: An Annotated Corpus of Semantic Roles. *Computational Linguistics* 21(1):245-288.