

Towards the Automatic Merging of Language Resources

Silvia Neculescu[†], Núria Bel[†], Muntsa Padró[†], Montserrat Marimon^{*}, Eva Revilla[†]

[†] IULA

Universitat Pompeu Fabra
Roc Boronat 138,
08018 Barcelona

^{*} Universitat de Barcelona^{*}

Gran Via de les Corts Catalanes, 585
08007 Barcelona

E-mail:

nuria.bel	muntsa.padro	silvia.neculescu	eva.revilla}	@upf.edu	montserrat.marimon@ub.edu
-----------	--------------	------------------	--------------	----------	---------------------------

Abstract

Language Resources are a critical component for Natural Language Processing applications. Throughout the years many resources were manually created for the same task, but with different granularity and coverage of information. To create richer resources for a broad range of potential reuses, information from all resources has to be joined into one. The high cost of comparing and merging different resources by hand has been a bottleneck for merging existing resources. With the objective of reducing human intervention, we present a new method for automating merging of resources. We have addressed the merging of two verb subcategorization frame (SCF) lexica for Spanish. The results achieved, a new lexicon with enriched information and conflicting information signalled, reinforce our idea that this approach can be applied for other task of NLP.

1. Introduction

The production, updating, tuning and maintenance of Language Resources for Natural Language Processing is currently being considered as one of the most promising areas of advances for the full deployment of Language Technologies. The reason is that these resources that describe, in one way or another, information about the characteristics of a particular language are necessary for language technologies to work. For many technologies –Machine Translation, Parsing, Information Extraction, etc.– this particular information is stated in the form of a lexicon that registers how words are used and combined within that language. In other cases, the technology induces this information from a corpus of texts annotated with explicit information about these relations. Thus, the demand of both annotated corpora and lexica has augmented in the last years.

Although the re-use of existing resources such as WordNet (Fellbaum, 1998) in different applications has been a well known successful case, it is not very frequent. The different technology or application requirements, or even the ignorance about the existence of other resources, has provoked the proliferation of different, unrelated resources that, if merged, could constitute a richer repository of information augmenting the number of potential uses. This is especially important for under-resourced languages (perhaps for all but English), which normally suffer the lack of broad coverage resources. The research reported in this paper was done in the context of the creation of a gold-standard of subcategorization frames of Spanish verbs to be used in lexical acquisition (Korhonen, 2002). We wanted to merge two hand-written, large scale Spanish lexica to get a new richer and validated one. Because subcategorization frames contain rich and structured information, it was considered a good scenario for testing language resource merging methods.

Several attempts of resource merging have been

addressed and reported in the literature. Hughes et al. (1995) report on merging corpora with more than one annotation scheme. Ide and Bunt (2010) also report on the use of a common layer based on a graph representation for the merging of different annotated corpora. Teufel (1995) and Chan & Wu (1999) were concerned with the merging of several source lexica for part-of-speech tagging. The merging of more complex lexica has been addressed by Crouch and King (2005) who produced a Unified Lexicon with lexical entries for verbs based on their syntactic subcategorization in combination with their meaning as described by WordNet, Cyc (Lenat, 1995) and VerbNet (Kipper et al., 2000).

In this context, proposals such as the Lexical Markup Framework, LMF (Francopoulo et al. 2008) become an attempt to standardize the format of computational lexica as a way to avoid the complexities of merging lexica with different structures.

In what follows, we will first introduce some background information about SCF lexica, and describe each resource involved in the experiment. We will also demonstrate an issue of encoding: how the same phenomena can be represented differently in each lexica, and introduce the structure of features that will be merged. In section 3, we will present the merging process, analyze the results obtained and introduce the need of adjusting results. Finally, in section 4, we will draw conclusions from our experiment, and advance future lines of research to further pursue the goal of reducing human intervention to only at the verification step.

2. Information encoded in SCF lexica

Subcategorization frames (SCF) are meant to make explicit the number and role of the complements that a predicate, most typically a verb, needs for forming a correct sentence and, more importantly, being correctly interpreted. Thus, the interpretation of sentence “John eats every morning” crucially depends on the knowledge that the verb “to eat” can be intransitive, that is, there is no

need to take a noun phrase as a complement. Note that the most usual case is that one lemma has more than one SCF, as is shown in Table 2. For every instance of one lemma in a text, the corresponding SCF should be chosen regarding its complements. As we have seen in the last example, the meaning of a sentence is strongly related to the complements of the verb. The decision on whether or not an element is a complement of a particular verb is made by a syntactic analysis which implies a parser. Parsers must be supplied with information to describe the syntactic behavior of each verb such as the number and characteristics of the complements that every verb takes, whether the occurrence of these complements is obligatory or not, and on how every particular complement contributes to the meaning of the whole sentence. Currently, both rule-based and statistical parsers benefit from this lexical information, first in the analysis step and the latter in the learning process (Jurafsky and Martin 2009 and Manning and Schütze 1999, for a discussion of the benefits of lexicalized statistical parsing). It is important to note that SCF phenomena differ substantially among language families. For instance, for Romance languages to encode how verbs behave with respect to cliticization phenomena, including “se” pronominalization is mandatory.

In the experiment we report here, we merged two subcategorization lexica developed for rule-based grammars; the Spanish working lexicon of the Incyta Machine Translation system (Alonso, 2005) and the Spanish working lexicon of the Spanish Resource Grammar, SRG, (Marimon, 2010) developed for LKB framework (Copestake, 2002). Note that different senses under the same lemma are not distinguished in these lexica, and thus, are not addressed in the research reported here. In the case of one lexicon enriched with different senses for one lemma, the merging mechanism would be the same. The difference would stay in the lexicon indexation. Instead of grouping the SCFs with respect to a lemma, they will be grouped under each pair’s lemma-sense. Following is a brief description on how these two lexica encode SCF information.

2.1. The encoding of SCF in the Incyta lexicon

In the Incyta lexicon, each verb entry is represented as a list of tags. The subcategorization information for each verb is encoded in the 'ARGS' feature as a parenthesized list of all the possible subcategorization patterns that a given verb can have, even if the different patterns imply a change in the meaning of the verb.

The information contained in the SCF includes a list of the possible complements, indicating for each of them the grammatical function (\$SUBJ, \$DOBJ, \$IOBJ, \$POBJ, \$SCOMP, \$OCOMP, \$ADV), the phrase type that can fulfill each grammatical function ('N1' for noun phrase, 'N0' for clausal phrase, 'ADJ' for adjective phrase) and the preposition required in the case of prepositional objects (\$POBJ). In the case of clausal complements, the information is further specified, indicating the type of clause (finite, 'FCP', or non-finite, 'ICP') in the interrogative ('INT') or non-interrogative ('0') forms, and the mode ('SUB' or 'IND' in the case of a finite clause) or

the control structure ('PIV \$SUBJ', 'PIV \$DOBJ', etc.), in the case of non-finite clauses. Incyta further specifies if one of the complements can be fulfilled by a reflexive and/or reciprocal pronoun ('\$DOBJ APT RFX'). Apart from the number and type of the complements, the subcategorization pattern includes further characteristics, represented by the GFT tag (General Frame Test). For example, whether the verb is impersonal for weather-like verbs (LEX-IMPS T) or if it can take the “se” clitic (RFX), that is, pronominal verbs as explained in 2.3, or if it can occur in the form of an absolute past participle construction.

2.2. The encoding of SCF in the SRG lexicon

The SRG is grounded in the theoretical framework of Head-driven Phrase Structure Grammar, HPSG, (Pollard and Sag, 1994), a constraint-based, lexica-list approach to grammatical theory where all linguistic objects (i.e. words and phrases) are represented as typed feature structures. In the SRG lexicon, each lexical entry consists of a unique identifier and lexical type (one among about 500 types, defined by a multiple inheritance type hierarchy).

Verbs are encoded by assigning a type and adding specific information to the lexical entries. Verbal types are first distinguished by the value for the SUBJ-list. Thus, we have subtypes for impersonal verbs taking an empty SUBJ-list, verbs taking a verbal subject and verbs taking a nominal subject.

The feature COMPS is a list of the complements which specifies the phrase structure type of each complement; i.e. noun phrase (NP), clause phrase (CP), prepositional phrase (PP), adjectival phrase (AP), adverbial phrase (ADV), and subject complement (SCOMP). Verbal complements are specified for their form (finite or infinitive), mode (indicative or subjunctive), and control or raising relation of verbal complements. Marking prepositions for some verbs are given in the lexicon itself, while for the others just the preposition’s type is specified. Alternations of complements, as well as other valence changing processes that verb frames may undergo, are dealt with by the grammar rules, which are triggered by lexical feature-value attributes that encode whether a verb is, for instance, reflexive or pronominal.

2.3. Issues in information merging

It is evident from previous section, that the SRG and the Incyta lexica encode the same phenomena but in a slightly different way. For a task like automatic merging, information about the same facts must be represented exactly in the same way as to compare and decide whether (Crouch and King, 2005):

- it is the same information
- it is different information that should be kept in the resulting lexicon
- it is different information that points at some gap or inconsistency in one of the lexica, if not directly to an error

In addition to mere formal differences, i.e. different tags, there can be differences in the semantics of a given tag, i.e. one tag covers what in another dictionary covers two tags. One of the most complex cases we found was the

encoding of reflexive and pronominal verbs in both lexica. Now, we will briefly review the implications of this phenomenon, the complexity of representing it and how these two lexica encode it differently, which was one of the more interesting issues to study in the results of the merging experiments.

In Spanish (but also in other languages like French, Italian, Dutch, German, etc.) the presence of the reflexive pronoun triggers, in combination with different verbs, different interpretations related to diathesis and the number and interpretation of the arguments. The so called pronominal verbs are those that are lexically marked, which in some constructions occur with a pronominal clitic particle 'se', without referential value. The lack of referential value distinguish these constructions from other clitic occurrences like the expletive use of clitics to refer to an obligatory, but not mentioned, object such as (1), or a true reflexive occurrence like in (2).

1. *La vi*
'I saw her'
2. *Me lavo las manos*
'I wash my hands'

Pronominal verbs are normally classified into two groups. *Inherent/absolute pronominal verbs*: Their frame obligatorily requires the occurrence of the clitic and because it depends only on the lexical item, it must be encoded in the lexica.

3. *Juan se ha atrevido a pedir un aumento*
'John CLI dared to ask for a raise'
*Juan ha atrevido a pedir un aumento

Argument reducing pronominals: When appearing with the clitic, its otherwise transitive structure is reduced in one element and the internal argument becomes external. It is normally related to anticausativization phenomena (Bosque, 1999):

4. *El capitán ha hundido su barco*
'The captain sank his ship'
5. *El barco se ha hundido*
'The ship has sank'
6. *Juan ha roto un vaso*
'John broke the glass'
7. *El vaso se ha roto*
'The glass CLI broke'

In Spanish a further problem arises because of the surface similarity between these 'pronominal verb' constructions and the impersonal and reflexive passive sentences also expressed with the clitic 'se'. Most verbs can enter in these constructions where the 'impersonal' value comes from the fact that when appearing with 'se' they inflect in the 3rd. person, here is no lexical subject and they have not or they do not imply reference to any definite subject as they would do if the particle 'se' was eliminated.

8. *Se vive bien en Barcelona*
'People live well in Barcelona'
9. *Se han suspendido las negociaciones*
'The negotiations have been suspended'

In a reflexive passive construction the verb agrees in number with the nominal element which is considered grammatically to be the subject, producing thus a reduction in the number of complements too, like the pronominal case just mentioned.

Due to this variety of possible uses of "se" and the subtle nuances of their interpretations, there is a significant degree of hesitation, if not confusion, when encoding reflexive and pronominal verbs in the lexicon. Our two lexica were not an exception and, most probably because of the difficulties of consistently encoding pronominal verbs, each lexicon has opted for a different strategy and, critically, they do not always agree in the classification of a verb as pronominal or reflexive, the two cases where specific information in the SCF lexicon is required. The Incyta lexicon encodes the possibility of bearing a "se" clitic and taking part in an argument reduction phenomenon with the tag "GFT RFX" annotating the whole SCF. Besides, it marks the possibility of an argument taking a reflexive pronoun adding the feature-value "(APT RFX)" as an annotation in \$DOBJ and \$IOBJ complements.

The SRG lexicon distinguishes with different types among the reflexive or pronominal interpretation of a verb when occurring with "se".

	Reflexive			Pronominal		
	#verbs	#both	#singles	#verbs	# both	#singles
SRG	835	190	645	712	597	115
Incyta	204		14	1204		607

Table 1: Differences of reflexive and pronominal encoding in the two lexica

In table 1 we can see the number of verbs encoded as reflexive (as 1) and pronominal (as 4 and 5) and the overlapping of the two lexica expressed as the number of verbs equally encoded in both lexica. Singles refer to those that are only encoded as reflexive or pronominal in one of the lexica. Despite the difference in quantities, one can observe that the overlap is far from being in the majority, and that there is a significant amount of systematic differences within the encoding.

2.4. The encoding of SCF in the common lexicon

As we said before, our objective was to merge two SCF lexica by graph unification which allows us to combine the information contained in two lexica. This method fulfills our objective to create a complete and correct SCF lexica using information from two manually created resources. By unification, we validate the common information, exclude the inconsistent and add the unique information that each lexicon contains.

The first step of the process is to convert each lexicon into a format which supports graph unification. We decided to use feature-value structures, which form directed acyclic graphs, i.e. the features are arrows and the values, nodes. A graph being a structured representation, intuitively presents the lexical information and it can be easily

transformed, after merging, to other standard formats for further reuse.

The exercise of converting the information contained in a lexicon is referred to as the extraction phase and several rules were manually written according to the intended interpretation of the encoding found in the lexica in order to make it match only within the cases wanted, respecting different information that must occur in the new resource, and indicating when contradictory information occurs for the same verb.

The extraction phase revealed major differences between the two lexica in the following cases:

(i) Different information granularity. This was the case of the Incyta tag “N0” for referring to the category of the phrase that can fulfill a complement. It had to be split according to their form, into a ‘finite’ or ‘infinitive’ clause in order to compare with the SRG encoding.

(ii) Different grammatical coverage. For instance, the Incyta lexicon lists bound prepositions, while the SRG lexicon can refer to the type of the bound prepositions (i.e. locative or manner).

(iii) Different treatment of systematic complement alternations. SRG handled them by lexical rules while Incyta explicitly declared them as possible SCF or disjunctions included in one of them. For example, a verb that has a complement that may be fulfilled by both a finite and an infinitive clause is represented with a type that includes a lexical rule that will produce the alternation when needed. In the Incyta lexicon this phenomenon is encoded as two different realizations in the SCFs, one for the finite clause (FCP) and one for the infinite (ICP). Thus, in this example, one extraction rule would convert one SRG frame into two: one with finite and one with an infinite clause complement.

These differences in encoding resulted in a different number of SCF, which we will comment upon later.

In general terms, the extraction rules mapped the information of each lexicon into a graph that can be represented as an attribute-value matrix. The attribute and values used are the following (the names are used for internal purposes, but a translation into recommended LMF labels is planned):

- ‘subj’ specifies the category of the subject, i.e. Noun Phrase (NP), Complementizer Clause Phrase (CP);
- ‘comp_1’ and ‘comp_2’ specify the category of the first, respective to the second, verb complement, i.e. none (no complement), adverbial (adv), indirect object (ppa), adjectival (adj), NP, CP or PP.
- ‘passive’ specifies if the verb accepts to undergo passive.
- ‘apc’ specifies if the verb occurs as absolute past participle construction
- ‘rpc’ specifies if the verb is reciprocal verb as in the example: “Juan y María se escriben cartas” (*Juan and María write letters to each other*).
- ‘rfx_prn’ is a complex valued attribute that

specifies if the verb is reflexive ([clitic=‘yes’; rfx=‘yes’; prn=‘no’]), pronominal ([clitic=‘yes’; rfx=‘no’; prn=‘yes’]) or none of them ([clitic=‘no’; rfx=‘no’; prn=‘no’]).

The attributes ‘subj’, ‘comp1’ and ‘comp2’ can be simple structures, i.e. NP, or complex structures. In the latter case, they include a list of specific features, indicating the category, their form (finite or infinite, affirmative or interrogative), the verbal mode (indicative or subjunctive) or the preposition required.

Another complex structure is the ‘rfx_prn’ attribute containing the ‘clitic’ feature discussed in section 2.3, which takes ‘yes’ if the verb is a pronominal or reflexive and ‘no’ if it does not accept this type of pronoun. ‘prn’ and ‘rfx’ information is triggered by the SRG lexicon which encodes, whether a verb is pronominal or reflexive. For verbs which accept both types of pronouns, they will have two different SCF, one for each behavior.

As we see in table 1, the agreement among the two lexica regarding the encoding of reflexive and pronominal phenomena was far from being complete. This would be a handicap for unification algorithms which unify only if the values of all common features are compatible (in this case if the verb is reflexive or pronominal). Thus, we had to standardize the division in reflexive and pronominal classes from one of the lexica. After a manual inspection we decided to preserve the SRG information because it was richer, we collapsed the Incyta information in only one feature “clitic=yes”, i.e. the verb is pronominal or reflexive, and let the SRG make the final decision during the moment of unification. For each verb in the SRG, a further “clitic=yes” was added, at the same level as ‘prn’ and ‘rfx’ features, to prevent unification with those entries that had no information, and thus can unify without restrictions.

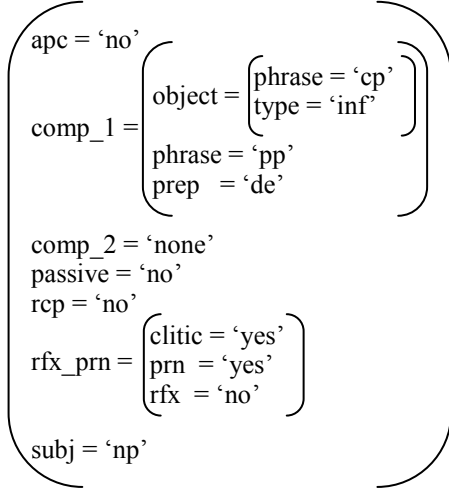
When we organize information from both dictionaries in a common format, we looked for a structure that keeps the information valid during the process of graph unification. For instance, a PP with a CP object cannot unify with a PP with a NP object.

3. Unification

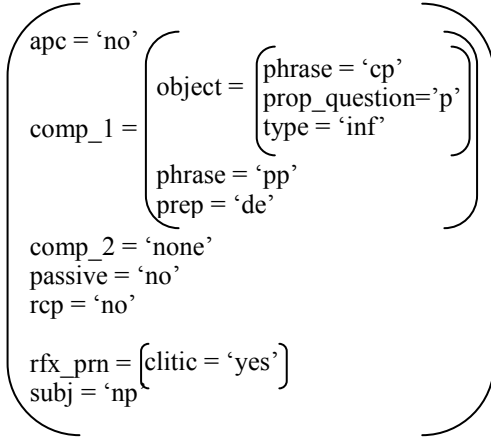
After the manual effort of conversion into a ready to unify format, the second step was the unification of the two lexica that contain the same structure and features. The objective of merging the two SCF lexica was to have a new, richer lexicon with information coming from both. After each lexicon was mapped into a common format, the results were mechanically compared and combined to form the new resource.

Once the SCF was converted into graphs, we used the basic unification mechanism implemented in NLTK (Bird et al., 2009) for each verb to merge its SCF from the Incyta lexicon with those from the SRG lexicon. For a better understanding of the unification process, in Figure 1 we present the results of the unification for the verb ‘reprimir’ (to repress), where it is interesting to note the resulting values of the ‘rfx_prn’ feature. This verb is considered in the Incyta lexicon as a ‘clitic’ verb, without

SRG:



Incyta:



Result:

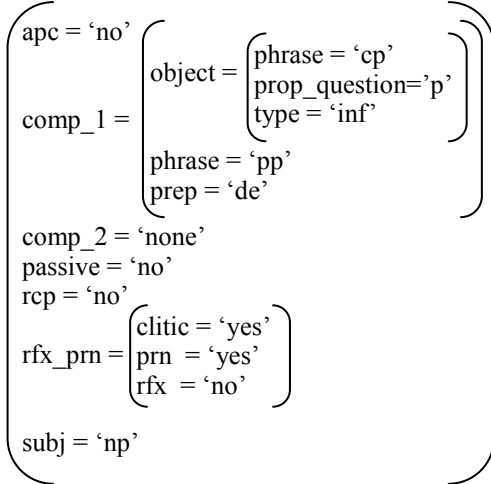


Figure 1: The results of the merging for the verb 'reprimir' (to repress).

expressing the values for 'rfx' and 'prn' features. However, in the SRG lexicon, it is encoded as a pronominal verb; therefore the final SCF lexicon considers it also a pronominal verb. In addition, the same example presents a case of lack of information in SRG because it does not specify the differences of the causal phrase from the PP in the case that it is a statement or a question. The resulting structure fulfills our objective to maintain information from both lexica.

The unification process tries to match many-to-many SCFs under the same lemma. This means that for every verb, each SCF from one lexicon tries to unify with each SCF from the other lexicon. The resulting lexicon is richer in SCFs for each lemma, on average, as shown in Table 2, where we present the results of merging the two lexica in terms of SCFs, lemmas and the average of SCF per lemma. Note that we present both the number of unique SCFs in the three lexica and the number of total SCFs that can be found in them.

The resulting lexicon will contain lemmas from both dictionaries and for each lemma, the unification of the SCFs from the Incyta lexicon with those from the SRG lexicon. The unified SCFs can be split in three classes:

- (1) SCFs of verbs that were present in both dictionaries, i.e. A_{SCF} is contained under one lemma in both lexica, thus the resulting lexicon, contains A_{SCF} under this lemma;
- (2) Information on SCF's components that were present in one of the lexicon but not in the other, i.e. the Incyta lexicon contains A_{SCF} , while the SRG lexicon contains B_{SCF} under the same lemma. A_{SCF} and B_{SCF} unify in C_{SCF} , where C_{SCF} contains the common information and also the information just in A_{SCF} or just in B_{SCF} ;
- (3) SCFs that were present in one of the lexicon but not in the other: the Incyta lexicon contains A_{SCF} , while the SRG lexicon contains B_{SCF} under the same lemma. A_{SCF} and B_{SCF} cannot unify, thus the resulting lexicon contains for the same lemma both frames, A_{SCF} and B_{SCF} .

Group (3) consists in inconsistent information in lexica, as it can signal a lack of information in one lexicon (e.g. A_{SCF} appears in Incyta but it does not have a corresponding SCF in SRG) or an error in the lexica (at least one of SCF implicated into the unification is an incorrect frame for its lemma). Thus, for detection conflicting information, we will detect lemmas whose SCFs do not unify at all (the unification number under a lemma is 0), or SCFs in one or the other lexicon that never unify with other SCFs (the total unification number for a SCF is 0). In a further step, by using a human specialist, this information can be manually analyzed and eventually eliminated from the final lexicon. Our objective is to automatically merge a lexica, thus we consider human analysis a possible intervention that would be useful to filter the results, but not a necessary step. The resulted lexicon contains all valid information provided by the unification of lexica and some SCFs that can be incorrect or not.

Lexicon	Unique SCF	Total SCF	Lemmas	Avg.
SRG	326	13.864	4303	3.2
Incyta	660	10.422	4070	2.5
Merged	919	17.376	4324	4

Table 2: Results of the merging exercise

It can be seen from the number of unique SCFs that the Incyta lexicon has many more SCFs than the SRG lexicon. This is due to different granularity of information. For example, the Incyta lexicon always gives information about the concrete preposition accompanying a PP while, in some cases, the SRG gives only the type of preposition, as explained before. The number of unique SCFs of the resulting lexicon, which is close to the sum between the numbers of the unique SCFs in the lexica, was very surprising for us. As shown in Table 3, for 50% of the lemmas we have a complete unification; thus this result comes from the many to many unification rather than from the direct addition of SCFs from both lexica.

Regarding the average number of SCFs per lemma in the different lexica, we use the total number of SCFs to calculate it.

Lemmas	Unification classes (lemmas)	Resulted SCF				
		Total	Unify	No unify		
				SRG	Incyta	
4050	2166	3329	3329	0	0	(1)
	888	7424	1966	3119	2339	(2)
	525	2977	1123	1854	0	(3)
	197	991	600	0	391	(4)
	274	1810	0	1123	687	(5)
274		845	0	778	67	(6)

Table 3: Detailed results of merging:
(1) Unify 100%; (2) There are not unified SCFs in both lexica; (3) There are not unified SCFs in SRG; (4) There are not unified SCFs in the Incyta lexicon; (5) Any SCFs do not unify; (6) Appears only in one lexicon.

Table 3 explains with more details the source of this gain of SCFs. Our final lexicon contains a total of 4,324 lemmas. From those, 4,050 appeared in both lexica (94%). 2,166 lemmas (class (1) from Table 3) unified all their SCFs signifying a total accord between both lexica for 50% of lemmas. Note that for 2,160 of them, every SCF from the Incyta lexicon unifies with one and only one SCF in the SRG lexicon, that is a unification type '1 to 1', while 6 verbs accomplish a many-to-many unification.

1,610 (the classes (2), (3) and (4) from Table 3) lemmas do not unify all the SCFs thus they reveal differences between both lexica, as explained in section 2.4. These lemmas present, in total, 8,637 SCFs in the SRG lexicon and 6,342 SCF in the Incyta lexicon. Through the unification process under the same lemma, 3,689 SCFs unify, while a total of 4,973 SCFs from the SRG lexicon and 2,730 SCFs from the Incyta lexicon are added directly into the resulting lexica. Besides, the resulting lexicon contains 274 lemmas (the class (6) from Table 3) that appear just in one lexicon, 21 lemmas appear just in the Incyta lexicon and 253 lemmas appear just in the SRG lexicon, which are considered as lacking of information. They are the best proof of our results that the new lexicon is more consistent in information.

Only 274 lemmas, 6,3%, did not unify any SCFs because

of conflicting information and require further manual analysis. An example of complete unification failure comes from the inconsistent encoding of pronominal and reflexive verbs in a hand-made lexicon like the one we have introduced in section 2.3.

In order to assess the quality of the new resource, we performed a manual inspection of lemmas whose frames can't be unified. Our objective was to identify what was the inconsistent information and we had a special interest in the results of the merging of the pronominal and reflexive verbs, which we knew are problematic.

In the Incyta lexicon, most of the reflexive or pronominal verbs have two different SCFs: one for the occurrence of the clitic personal pronoun, no NP complement and the tag for a reflexive verb (e.g. cubrir: "yo me cubro", *I cover myself*) and another one for the NP complement, in this case it is no longer encoded with the tag for reflexive verb ("yo cubro el coche", *I cover the car*).

On the contrary, in SRG, both realizations of a reflexive verb are included in the same frame, indicating both that it may have a NP complement and a reflexive tag. Because the clitic pronoun and the NP complement cannot appear in the same SCFs, when extracting all possible SCFs that a SRG verb may have our set of extraction rules creates two SCFs: one reflexive, without NP complement, and another non-reflexive with NP complement. Using this strategy, we obtained over 3600 unifications for these types of verbs, thus we consider our approach correct. However, we found that around 100 Incyta verbs had been encoded following the same interpretation as the original the SRG lexicon. These verbs have a SCF that contains both the NP complement and the reflexive tag and thus do not unify with the SRG SCF's that have been split into two SCFs. These verbs are a third of the ones that do not unify any SCF. Figure 2 demonstrates these particular feature structures.

As it can be seen from the tables above, the resulting lexicon is richer than the two it is composed of as it has gained information in the number of SCFs per lemma, as well as in the information contained in each SCF. Table 2 shows an increase of SCFs per lemma on average.

In general, automatic merging produces errors that can easily be the object of further refinement, because errors are systematic. However, this tends not to be true of manual merging exercises, where human errors are occasional and hence, inconsistent, as we have seen in the encoding of reflexive verbs in the Incyta lexicon.

Thus, an automatic merging process can have a final step, based on what Crouch and King (2005) call "patch files". Using our observations collected during the final verification, we will consider for the future to devise specific patches that correct or add information in particular cases where either wrong or incomplete information is produced. A first candidate case would be to correct all of the verbs in the Incyta lexicon with SCFs that have both the reflexive tag and the NP complement.

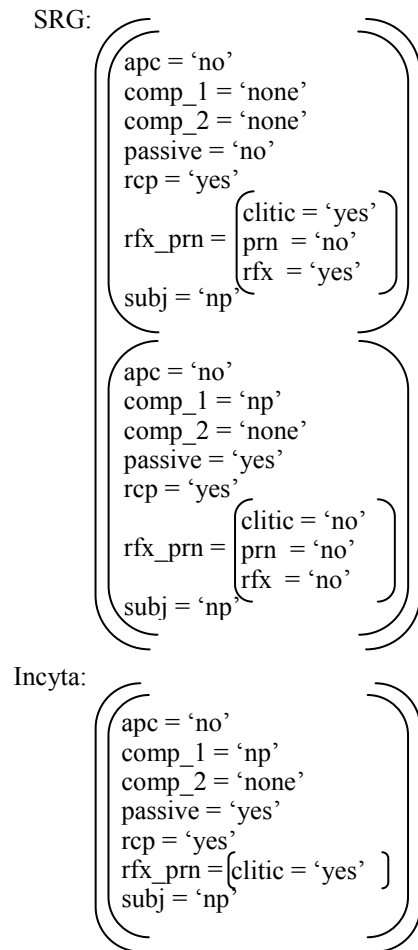


Figure 2: Example of unification problem for a reflexive verbs, such as 'cubrir' ('to cover')

4. Conclusions

We have proposed a method to reduce human intervention in the merging of Language Resources, in particular within the SCF lexica. By using graph unification as the sole operation that controls merging, we support the proposal of Ide and Bunt (2010) for rich annotated corpus merging, demonstrating that it is also possible for lexical merging. Our proposal of extracting information and representing it as a graph in order to only use a unification method for the actual merging is an innovative proposal in the field of dictionary merging. The structure proposed is based on attribute-value feature-based directed acyclic graphs and can be easily transformed into a standard format for further reuse.

We consider the results obtained in our experiments very satisfactory. Unifying two SCF lexica after converting them to a common operative format by using extraction rules led to a richer resource that will be offered to the community as a gold-standard of verbal SCF for Spanish. During the unification step errors, which are systematic due to the formal merging process, can be detected and then corrected using patch files.

The mapping of information to a common structure remains a very expensive part of resource merging if done manually. It is future work to reduce the cost of information comparison and extraction exercises by

proposing an automatic mapping solution.

5. Acknowledgments

This project has been funded by the PANACEA project (EU-7FP-ITC-248064) and the CLARA project (EU-7FP-ITN-238405).

6. References

- Alonso J.A., Bocsák A. (2005). Machine Translation for Catalan-Spanish. The Real Case for Productive MT; In *Proceedings of the tenth Conference on European Association of Machine Translation (EAMT 2005)*, Budapest, Hungary.
- Bird S., Klein E., Loper E. (2009) *Natural Language Processing with Python*. O'Reilly Media, 1 edition
- Bosque I, Demonte V., Eds. (1999): Gramática descriptiva de la lengua española, R.A.E. - Espasa Calpe, Madrid.
- Chan D. K., Wu.D. (1999). Automatically Merging Lexicons that have Incompatible Part-of-Speech Categories. *Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC-99)*. Maryland.
- Copetake A. (2002). Implementing Typed Feature Structure Grammars. *CSLI Publications*, CSLI lecture notes, number 110, Chicago.
- Crouch D, King T.. (2005). Unifying lexical resources. *Proceedings of Interdisciplinary Workshop on the Identification and Representation of Verb Features and Verb Classes*. Saarbruecken; Germany.
- Farrar S, Langendoen D. T. (2003) A linguistic ontology for the Semantic Web. *GLOT International*. 7 (3), pp.97-100
- Fellbaum C. (1998). WordNet: An Electronic Lexical Database. MIT Press.
- Francopoulo G., Bel N., George M., Calzolari N, Pet M., Soria C. (2008). Multilingual resources for NLP in the lexical markup framework (LMF). *Journal of Language Resources and Evaluation*, 43 (1).
- Hughes J., Souter C., Atwell E. (1995). Automatic Extraction of Tagset Mappings from Parallel-Annotated Corpora. *Computation and Language*.
- Ide N. and Bunt H.. (2010). Anatomy of Annotation Schemes: Mapping to GrAF. *Proceedings of the Fourth Linguistic Annotation Workshop, ACL 2010*
- Jurafsky D., Martin J.H. 2009. Speech and Language Processing: An Introduction to Natural Language Processing, *Speech Recognition, and Computational Linguistics*. 2nd edition. Prentice-Hall.
- Kipper K., Hoa Trang Dang, H.T., Palmer M.. (2000). Class-based construction of a verb lexicon. In *Proceedings of AAAI/IAAI*.
- Korhonen A. (2002). Subcategorization Acquisition. PhD thesis published as *Technical Report UCAM-CL-TR-530*. Computer Laboratory, University of Cambridge
- Lenat D. (1995). Cyc: a large-scale investment in knowledge infrastructure. In *CACM* 38, n.11.
- Manning C.D., Schütze H. (1999). Foundations of Statistical Natural Language Processing. *MIT Press*,

Cambridge, MA, USA.

- Marimon M. (2010). The Spanish Resource Grammar. Proceedings of *the Seventh Conference on International Language Resources and Evaluation (LREC'10)*. Paris, France: European Language Resources Association (ELRA).
- Molinero Miguel A., Sagot Benoît and Nicolas Lionel (2009). Building a morphological and syntactic lexicon by merging various linguistic resources. In Proc. of 17th Nordic Conference on Computational Linguistics (NODALIDA-09), Odense, Denmark.
- Pollard, Sag I.A. (1994). Head-driven PhraseStructure Grammar. The University of Chicago Press, Chicago.
- Teufel S. (1995). A Support Tool for Tagset Mapping. In *EACL-Sigdat 95*