A Method Towards the Fully Automatic Merging of Lexical Resources

Núria Bel
Universitat Pompeu Fabra
Barcelona, Spain
nuria.bel@upf.edu

Muntsa Padró
Universitat Pompeu Fabra
Barcelona, Spain
muntsa.padro@upf.edu

Silvia Necsulescu
Universitat Pompeu Fabra
Barcelona, Spain
silvia.necsulescu@upf.edu

Abstract

Lexical Resources are a critical component for Natural Language Processing applications. However, the high cost of comparing and merging different resources has been a bottleneck to obtain richer resources and a broader range of potential uses for a significant number of languages. With the objective of reducing cost by eliminating human intervention, we present a new method towards the automatic merging of resources. This method includes both, the automatic mapping of resources involved to a common format and merging them, once in this format. This paper presents how we have addressed the merging of two verb subcategorization frame lexica for Spanish, but our method will be extended to cover other types of Lexical Resources. The achieved results, that almost replicate human work, demonstrate the feasibility of the approach.

1 Introduction

The automatic production, updating, tuning and maintenance of Language Resources for Natural Language Processing is currently being considered as one of the most promising areas of advancement for the full deployment of Language Technologies. The reason is that these resources that describe, in one way or another, the characteristics of a particular language are necessary for Language Technologies to work.

Although the re-use of existing resources such as WordNet (Fellbaum, 1998) in different applications has been a well known and 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, which normally suffer from the lack of broad coverage resources. In the research reported in this paper, we wanted to merge two hand-written, large scale Spanish subcategorization lexica to obtain a new one that is larger and validated. Because subcategorization frames contain highly structured information, difficult to compare, it was considered a good scenario for testing new lexical resource merging methods. Other experiments merging resources containing different levels of information are also envisaged.

1.1 Related Work

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, a proposal such as the Lexical Markup Framework, LMF (Francopoulo et al. 2008) is understood as an attempt to standardize the format of computational lexica as a way to avoid the complexities of merging lexica with different structures. But it only considers manual comparison of resources and manual mapping from non-standard into the standard.

Despite the undeniable achievements of the research just mentioned, most of it reports the need for a significant amount of human intervention to extract information of existing resources and to map it into a format in which it can be compared with another lexicon, or towards proposed standards, such as the mentioned LMF. Thus, there is still room for improvement in reducing human intervention. This constituted the main challenge of the research reported in this paper: finding a method that can perform, without human intervention, semantic preserving information extraction and format mapping operations to allow for automatically merging two lexical resources, in this

Proceedings of Workshop on Language Resources, Technology and Services in the Sharing Paradigm, pages 8–15,
Chiang Mai, Thailand, November 12, 2011.

particular case two subcategorization frame (SCF) lexica for Spanish. The best results achieve up to 92% in precision and 93% in recall when comparing automatically and manually extracted entries, show the potential of our approach.

1.2 Merging Lexica

Basically, the merging of lexica has two well defined steps (Crouch and King, 2005). In the first, because information about the same phenomenon can be expressed differently, the information in the existing resources has to be extracted and mapped into a common format, making merging possible in a second step, where the extracted information from both lexica is mechanically compared and combined to form the new resource.

While automation of the second step has already proved to be possible, human intervention is still critically needed for the first. In addition to the cost of manual work, note that the exercise is completely ad-hoc for particular resources to be merged. The cost is what explains the lack of interest in merging existing resources, even though it is critically needed, especially for under-resourced languages. Any cost reduction will have a high impact in the actual re-use of resources.

Thus, our objectives were: first, to carry out a more traditional merging exercise achieving some improvements for step two by using graph unification as the only, basic mechanism. Second, to investigate to what extent we could reduce human intervention in the first step, we devised a semantic preserving mapping algorithm that covers the extraction of the information of any particular lexicon and its mapping onto another format that allows, later, the merging with another resource.

In the next section we introduce the two SCF lexica that we used to validate our proposal. Section 3 reports on the work done in manually extracting the information of our lexica and their mapping onto a common format in order to merge them and thus getting a gold-standard to evaluate the results of the automation exercise. Section 4 presents our proposal to use unification for the merging phase of both the manual and the automatically extracted resources. Section 5 explains how we addressed the problem of automatically mapping the contents of two lexica onto a common format in order to avoid manual extraction. Finally, section 6 states conclusions drawn and further research directions.

2 Information encoded in SCF lexica

Subcategorization frames (SCF) are meant to explicitly demonstrate 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. Note that the most usual case is that one lemma has more than one SCF.

In the experiment we report here, we merged two subcategorization lexica, developed for rule-based grammars, with the goal of creating a SCF gold-standard for Spanish. The two lexica are the Spanish working lexicon of the Incyta Machine Translation system (Alonso, 2005) and the lexicon of the Spanish Resource Grammar, SRG, (Marimon, 2010) developed for LKB framework (Copestake, 2002). Note that different senses under the same lemma were 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.

SRG and Incyta lexica encode phenomena related to verbal complements, their role and categorical characteristics expressed as restrictions. SCFs in the SRG lexicon are formulated in terms of feature-attribute value pairs, so they have a graph structure. In the Incyta lexicon, SCFs are represented as a list of parenthesis-marked components, each with a list-based, non structured information[1] declaration. In next sections we briefly introduce the format of both lexica.

2.1 The encoding of SCF in the Incyta lexicon

In the Incyta lexicon, the subcategorization information for each verb is encoded 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 each SCF includes a list of the possible complements, indicating for each of them the grammatical function ($SUBJ, $DOBJ, $IOBJ, $POBJ, $SCOMP, $OCOMP, $ADV), the phrase type that can fulfill each grammatical function ('N1' for noun phrase, 'N0' for clausal complement, 'ADJ' for adjective phrase) and the preposition required in case of prepositional objects ($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

[1] Decorated lists, parenthetical or otherwise marked, have been a quite common way of representing SCF information, i.e. COMLEX, VERBNET among others.

the complements can be fulfilled by a reflexive and/or reflexive pronoun ('$DOBJ APT RFX'). Apart from the number and type of the complements, the subcategorization pattern includes other subcategorization requirements, represented by the GFT tag (General Frame Test), such as whether the verb is impersonal for weather like verbs (LEX-IMPS T), can take the "se" clitic (RFX), that is, pronominal verbs, or can occur in the form of an absolute past participle construction.

2.2 The encoding of SCF in SRG lexicon

The SRG is grounded in the theoretical framework of Head-driven Phrase Structure Grammar, HPSG, (Pollard and Sag, 1994), a constraint-based, lexicalist approach to grammatical theory where all linguistic objects (i.e. words and phrases) are represented as typed feature structures. In the SRG, each lexical entry consists of a unique identifier and a 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 of 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 has as value a list of the complements which specifies the phrase structure type of each complement; i.e. NP, PP, AP, ADV, and 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 are given as specific information in the lexicon and included as variables in the types. Alternations of complements, as well as other valence changing processes that verb frames may undergo are dealt with lexical rules, which are triggered by lexical feature-value attributes that encode whether a verb can enter, for instance, a passive or a pronominal construction.

2.3 The encoding of SCF in the common lexicon

As we have said, in order to execute the merging of these two lexica, we first needed to convert them into a common format. In order to approach current proposals for standard formats (Francopoulo et al. 2008; Ide & Bunt, 2010) that recommend graph-based and attribute-value formalisms, we chose to map Incyta information towards the SRG format. Since this format already had a graph structure, it was compliant to the standard recommendations. Furthermore, the use of feature structures has several strong points for our work:

- It allowed us to easily combine the information contained in two lexica by graph unification, as we will see in section 4.
- Since graphs are structured representations, they can easily be transformed, after merging, to other standard formats for further reuse, so we consider them a good representation for our final SCF gold-standard.

Although SRG lexicon had already a graph structure, we still needed to perform some preprocessing, related to how we wanted to encode different subcategorization phenomena in our final SCF lexicon[2].

In both lexica, there were some phenomena to be treated by lexical rules which we decided to encode according to the following rules:

- The SCFs that contain an optional complement are split into two SCFs, one with the optional complement and one without it.
- SRG handles some phenomena, such as systematic complement alternations, by lexical rules. These rules are applied in order to create one SCF for each possible complement type. For example, a verb that has a complement that may be fulfilled by both a finite and an infinitive clause is represented with just a type that triggers a lexical rule that will produce the alternation in processing time. Thus, in this example one SRG frame would be converted into two: one with finite and one with an infinite clause complement.

We applied these preprocessing rules to SRG lexica, and converted Incyta lexicon into the graph-based format of SRG, ensuring that SCF patterns and the above mentioned phenomena are encoded in the same way.

3 Manual Extraction Phase

As previously said, the first step of the unification process was to convert Incyta lexicon into the chosen standard graph format, in this case, the feature-value structures of SRG lexicon.

This exercise of converting information contained in a lexicon is referred to by Crouch and King (2005) as the extraction phase. As a first exercise, we performed this conversion with several rules that were manually written according to the intended interpretation of the encoding found in the lexica. These extraction rules mapped the information of Incyta lexicon into a graph represented as an attribute-value matrix. This is what we called the manual extraction phase.

[2] The ultimate goal of the merging was to produce a complete lexicon that could be used as gold-standard in a SCF automatic acquisition experiment.

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

- Different information granularity. For example, this was the case of the Incyta tag "N0" for referring to the verbal category of the phrase that can fulfill a particular complement. The SRG encoding had a different tag for the finite clause case than for the infinitive case.
- Different grammatical coverage. For instance, the Incyta lexicon lists bound prepositions, while the SRG lexicon sometimes refers to the type of the bound prepositions (i.e. locative or manner).

This exercise was very time consuming, since it was necessary to study the codification of Incyta lexicon and to develop several rules to map them into SRG feature structures.

4 Unification Step

After the manual effort of conversion into a ready to unify format, the second step was the unification of the two lexica represented with the same structure and features. The objective of merging two SCF lexica is to have a new, richer lexicon with information coming from both. The resulting lexicon was richer in SCFs for each lemma, on average, as shown in Table 1.

Once the SCFs were converted into comparable graphs (in the sense that they have the same structure and possible feature-value pairs), we used the basic unification mechanism for merging the list of entries, i.e. lemmas, and the SCFs under the same lemma, from the two lexica. We used the implementation of feature structure representation and unification available in NLTK (Bird et al., 2009). 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.

Thus, the resulting lexicon contains 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:

- 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.
- SCFs that, though not identical in both lexica, unify into a third SCF, so they are compatible. This is due to SCF components that were present in one of the lexica but not in the other. For example, assume one SCF in the Incyta lexicon is equal to one SCF in SRG lexicon except that in the Incyta lexicon it contains information about the bound preposition (e.g. has the component "prep=in") while in SRG lexicon it contains only information about the preposition type (e.g. "prep_type=location"). The result of unifying these two SCFs is a richer SCF that contains both, the information of preposition and of preposition type.
- SCFs that were present in one of the lexicon but not in the other: the Incyta lexicon contains SCF_1, while the SRG lexicon contains SCF_2 under the same lemma. SCF_1 and SCF_2 cannot unify, thus the resulting lexicon contains for this lemma both frames, SCF_1 and SCF_2.

Group (3) can signal the presence of inconsistent information in one or the two lexica, like a lack of information in one lexicon (e.g. SCF_1 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, we can detect conflicting information searching the lemmas with SCFs that do not unify at all, or SCFs in one or the other lexicon that never unify with any other SCF. In a further step, with a human specialist, this information can be manually analyzed and eventually eliminated from the final lexicon. Nevertheless, in our work we do not approach this analysis step, so our final lexicon, contained all SCF obtained by unification and also those that did not unify with another SCF.

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 1: Results of merging exercise of manually extracted lexica

Table 1 shows the results of the manual merging exercise in terms of number of SCFs and lemmas in each lexicon. 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, may seem surprising. Nevertheless, a closer study showed that for 50% of the lemmas we have a complete unification; thus, the high number of SCF's in the merged lexicon comes from the many-to-many unification, that is, from the fact that one SCF in one lexicon unified with several SCFs in the other lexicon, so all SCFs resulting from these unifications will be added to the final

lexicon. This is the case for cases of different granularity, as explained before.

The final lexicon contains a total of 4,324 lemmas. From those, 94% appeared in both lexica, which means the resulting lexicon contained 274 lemmas that appear just in one lexicon. Those lemmas are added directly to the final lexicon. They are good proof that the new lexicon is richer in information.

Regarding lemmas that are in both lexica, 50% of them unified all their SCFs, signifying a total accord between both lexica. This is not surprising given that both are describing the same phenomena. On the other hand, 37% of lemmas contained some SCFs that unified and some that did not, which revealed differences between both lexica, as explained in section 3.

Only 274 lemmas (6,3%) did not unify any SCFs because of conflicting information, which we consider a very good result. These verbs may require further manual analysis in order to detect inconsistencies. An example of complete unification failure comes from the inconsistent encoding of pronominal and reflexive verbs in the lexica.

To summarize, the resulting lexicon is richer than the two it is composed of since it has gained information in the number of SCFs per lemma, as well as in the information contained in each SCF. Furthermore, note that the unification method allowed us to automatically detect inconsistent cases to be studied if necessary. For more information about these results and a more accurate discussion, see (*autocite*, 2011).

5 Automatic Mapping

Thus far, we have introduced our proposal to perform automatic merging of two lexica once they are represented as graph-based feature structures. Nevertheless, the most consuming part of the previous task was the extraction and mapping from the original format of a lexicon to a common graph structure. In this section, we present our proposal to automatically perform this mapping, which is the main contribution of this paper. In section 5.2 we will compare the results of the manual and the automatic extraction and mapping phase to assess the usability of our approach.

Our experiment to avoid manual intervention when converting the two lexica into a common format with a blind, semantic preserving method departs from the idea of Chan and Wu (1999) to compare information contained in the same entries of different lexica, looking for consistent, significant equivalences validated by a significant number of cases in the whole lexica. However, they were only mapping part-of-speech tags, while we needed to handle complex, structured information. Thus, our main goal was to reduce human intervention especially including the need to know the internal structure and semantics of the lexica to be merged. The basic idea behind the devised method is to let the system find semantically equivalent pieces of information coming from different resources and to substitute one with the other, in our case to substitute the parenthetical list of Incyta lexicon with the attribute-value equivalent matrix in the SRG lexicon.

5.1 Methodology

The only requirement of the following proposal for automatic mapping is to have a number of lemmas encoded in both lexica. With the same lemmas in both lexica, it is possible to assess that a piece of code in lexicon A corresponds to a piece of code in lexicon B, and to validate this hypothesis if a significant number of other lemmas hold the same correspondence. Thus, when a correspondence is found, the relevant piece in A can be substituted by the piece in B, performing the conversion into a common format to allow for the real merging. This is the basis of our method for carrying out the extraction phase automatically.

In order to maximize comparisons, each SCF was split into pieces in both lexica. Thus, the system had to search for parts of Incyta SCFs that correspond to parts of SRG graphs, i.e. single attribute-values or groups of them. Nevertheless, this search for relevant pieces had to be done automatically and only formal characteristics would be used. Since we did not want our method to be informed by human knowledge of the particular lexica to be merged, and in order to make it applicable to more than one lexicon, the first point to solve was how to compare two different SCFs code with no available previous information about their internal semantics. The only information used was that SCFs in the SRG lexicon were formulated in terms of feature-attribute value pairs and in the Incyta lexicon in terms of a list of parenthesis with less structured internal information.

An example of the code of one SCF in Incyta lexicon is (1):

(1) (($SUBJ N1 N0 (FCP 0 INT) (MD-0 IND) (MD-INT SUB)) ($DOBJ N1))

Therefore, the information that had to be discovered was the following:

- The Incyta lexicon marks each SCF as a list of parenthesis, where the first level of parenthesis indicates the list of complements. In example (1) there are two main parentheses, one representing the subject structure ($SUBJ …) and the other with direct object structure ($DOBJ …).
- Each component of the list begins with an identifier ($SUBJ or $DOBJ in (1)) followed, without necessarily any formal marker, by additional information about properties of the component in the form of

tags. For example, in (1) above, direct object ($DOBJ) is fulfilled by a noun phrase (N1).

- Incyta marks disjunction as a simple sequence of tags. In (1), subject ($SUBJ) may be fulfilled by N1 (noun phrase) or N0 (clause phrase). Furthermore, properties of one of the elements in the disjunction are specified in one or more parenthesis following the tag, as it is the case of N0 in (1). The 3 parenthesis after N0 are in fact properties of its realization: it is a sentential complement (FCP) whose verb should appear in indicative (MD-0 IND) unless it is an interrogative clause (MD-INT SUB). Note that this information is not structured so it was necessary to look for a way to detect that these parentheses refer to N0 and not to N1.

We devised an algorithm to discover and extract this internal structure from scratch. Our algorithm first splits every SCF in all possible ways according to only formal characteristics (minimal but complete parenthetical components for Incyta and minimal but complete attribute-value matrices for SRG) and looks, independently in each lexicon, for the most frequently repeated pieces along the whole lexicon, in order to assess that a particular piece is a meaningful unit in a particular lexicon. Note that we wanted to discover minimal units in order to handle different information encoding granularity. If we would have mapped entire SCFs or large pieces of them, the system could substitute information in A with information in B although possibly missing a difference.

Note that when performing the extraction, we aimed to ensure that as much information as possible from the original lexicon is preserved by splitting the lexicon into small pieces. However, in some cases, this created incomplete SCFs. Nevertheless, as our ultimate goal is to merge the two lexica, it is in the merging step that the partial elements will get the missing parts.

To sum up, our algorithm does the following with the Incyta SCF code:

- It splits SCF into each parentheses that conforms the list (this is, to find $SUBJ and $DOBJ in example (1)).
- For each of these pieces, it considers the first element as its key, and recursively splits the following elements.
- It detects the relationship among the different elements found inside the parentheses by assessing those that always occur together. For instance, in (1), it detects that FCP appears only when there is a N0, and that MD-0 appears only when (FCP 0) appears. In this way, the constituents of the parentheses grouped according to their dependency are automatically identified. The elements that always occur together are treated as minimal units.

On the other hand, it is also necessary to look for minimal units of the SRG lexicon. In this case, these minimal units are the values or features structures obtained when taking the values of the attributes at the first level of embedding. In this way, in the target format the minimal units are guaranteed to be semantically justified.

Once the minimal units of each Incyta and SRG SCFs are extracted, our algorithm does the following mapping:

- For each element extracted from the Incyta SCF, it creates a list of verbs that contain it. This list is represented as a binary vector whose element *i* is 1 if the verb in position *i* is in the list.
- For each minimal unit obtained from the SRG lexicon, it also builds a binary vector with the verbs that contain each element.
- For each Incyta SCF minimal unit, it assesses the similarity with each SRG unit comparing the two binary vectors using the Jaccard distance measure, especially suited for calculating distances between binary vectors and also used by Chan and Wu (1999).
- It chooses as mapping elements those that maximize similarity.

Once we had the mapping elements, new feature structures substituting Incyta units with SRG mapping elements are produced. Thus, a new version of the Incyta lexicon represented with feature-value structures is produced. The new feature structure-based entries could then be merged with the ones in SRG using unification, as we did with the manually extracted feature structures in section 4. Eventually, we obtained a new lexicon by merging the two lexica in a completely automatic way.

5.2 Evaluation and Results

To evaluate the results, we compared the two resulting lexica: the one resulting from the manual extraction and later unification and the lexicon resulting from the automatic extraction by mapping and again unification. Specifically, we use the manually built lexicon as a gold-standard. The evaluation is done using traditional precision, recall and F1 measures for each verb entry because most of them have more than one SCF and then we compute the mean of these measures over all the verbs.

We first counted only identical SCFs in the entries of every verb entry. However, we also took into account what we call the “compatible” entries. Note that in some cases the results of the automatic mapping are parts of SCFs instead of complete SCFs, because of the piece splitting

process. As said, merging by unification automatically adds the information as to complete them in numerous cases, but the Incyta SCFs that did not find any of the SGR SCFs to unify with can result in an additional but incomplete SCF in the final lexicon. They may be considered correct, although incomplete, when they are compatible with the information in the gold-standard, that is, when the automatically created entry subsumes the SCF in the gold-standard. Thus, in a second measurement, we also count these pieces that are compatible with SCFs in the gold-standard as a positive result. We keep figures separated, though, in table 2.

The results, shown in table 2, are near 88% of F1 in the strict case of identical SCFs. If we compare compatible SCFs, the results are even more satisfactory.

	P	R	F1
A-identical	87,35%	88,02%	87,69%
B-compatible	92,35%	93,08%	92,72%

Table 2: Average results of the mapping exercise

For a more detailed analysis of the results, we plot in Figure 1 the system performance in terms of number of SCFs under a lemma that are either identical or compatible in the gold-standard and in the merged lexicon. We also plot the ratio of verbs that have a particular number of SCFs or less (cumulative). The verbs that have one or two SCFs (about 50% of the verbs) obtain high values both in the exact matching and compatible SCFs, as it may be expected. Nevertheless, 95% of verbs (those with 11 or less SCFs per lemma) obtain at least F1=80% when counting only identical resulting SCFs and F1 over 90% when counting compatible resulting SCFs. Note that these figures are the lower threshold, since verbs with less SCFs have better results, as it can be seen in Figure 1. To summarize, the obtained precision and recall of all verbs, even those with more than two SCFs, are very satisfactory and constitute a proof of the feasibility of the approach.

As for the error analysis, the results revealed that some SCFs in the gold-standard are not in the automatically built lexicon. One case is SCFs with adverbial complements. Our algorithm maps adverbials onto prepositional phrases and the resulting SCF misses part of the original information. Nevertheless, our algorithm correctly adds information when there are gaps in one of the dictionaries. It is able to learn correspondences such as "INT" (Incyta for interrogative clause) to "q" in SRG and to add this information when it is missed in a particular entry of the SRG lexicon but available in the Incyta entry.

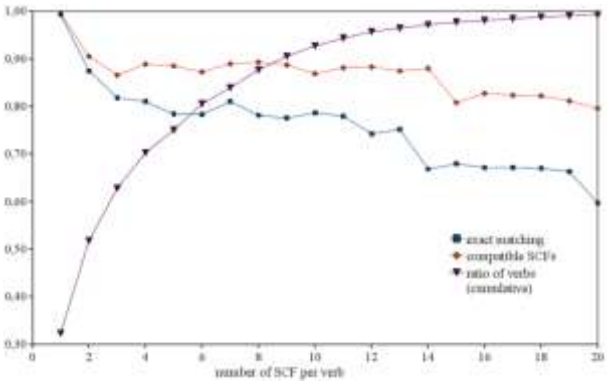

Figure 1: Average F1 and cumulative number of verbs with respect to the number of SCFs

6 Conclusions and Future Work

We have studied a method to reduce human intervention in the merging of lexical resources, and we have proved the concept with two SCF lexica. In order to merge different lexica by means of an automatic operation like unification, the resources need to be mapped into a common format. To reduce the cost of extracting and comparing the lexica contents, we proposed a method to make the mapping automatically. We consider the results obtained, above 80%, very satisfactory. Our method can indicate the possibility of avoiding the manual information extraction phase, which is a big bottleneck for the re-use and merging of language resources.

Furthermore, we can see the advantages of representing the lexica as feature structures because it enables the use of graph unification as an automatic mechanism for actual merging.

The strongest point of our method for automatically mapping the lexica into a common format is that it can be applied without the need of knowing the semantics of the lexica to be merged because it finds significant common code in existing lexica as to draw correspondences. This allows us to think our method can be extended to other types of Lexical Resources. The only requirement is that all resources to be mapped contain some common data. Although further work is needed for assessing how much common data guarantees the same results, the current work is indicative of the feasibility of our approach.

It is important to note that the results presented here are obtained without using what Crouch and King (2005) call patch files. Automatic merging produces consistent errors that can be object of further

refinement. Thus, it is possible to devise specific patches that correct or add information in particular cases where either wrong or incomplete information is produced. It is future work to study the use of patch files to improve our method.

Acknowledgments

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

References

Juan Alberto Alonso, András Bocsák. 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.

Steven Bird. 2006. NLTK: the natural language toolkit. In Proceedings of the COLING/ACL on Interactive presentation sessions. Association for Computational Linguistics, Morristown, NJ, USA.

Daniel K. Chan and Dekai Wu. 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.

Ann Copestake. 2002. Implementing Typed Feature Structure Grammars. CSLI Publications, CSLI lecture notes, number 110, Chicago.

Dick Crouch and Tracy H. King. 2005. Unifying lexical resources. Proceedings of Interdisciplinary Workshop on the Identification and Representation of Verb Features and Verb Classes. Saarbruecken; Germany.

Christiane Fellbaum. 1998. WordNet: An Electronic Lexical Database. MIT Press.

Gil Francopoulo, Núria Bel, Monte George, Nicoletta Calzolari, Mandy Pet, and Claudia Soria. 2008. Multilingual resources for NLP in the lexical markup framework (LMF). Journal of Language Resources and Evaluation, 43 (1).

John Hughes, Clive Souter, and E. Atwell. 1995. Automatic Extraction of Tagset Mappings from Parallel-Annotated Corpora. Computation and Language.

Nancy Ide and Harry Bunt. 2010. Anatomy of Annotation Schemes: Mapping to GrAF. Proceedings of the Fourth Linguistic Annotation Workshop, ACL 2010

Karin Kipper, Hoa Trang Dang, and Martha Palmer. 2000. Class-based construction of a verb lexicon. In Proceedings of AAAI/IAAI.

Anna Korhonen. 2002. Subcategorization Acquisition. PhD thesis published as Technical Report UCAM-CL-TR-530. Computer Laboratory, University of Cambridge

Doug Lenat. 1995. Cyc: a large-scale investment in knowledge infrastructure. In CACM 38, n.11.

Montserrat Marimon. 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).

Monica Monachini, Nicoletta Calzolari, Khalid Choukri, Jochen Friedrich, Giulio Maltese, Michele Mammini, Jan Odijk & Marisa Ulivieri. 2006. Unified Lexicon and Unified Morphosyntactic Specifications for Written and Spoken Italian. In Calzolari et al. (eds.), LREC2006: 5th International Conference on Language Resources and Evaluation: Proceedings, pp. 1852-1857, Genoa, Italy.C.J.

Silvia Necsulescu, Núria Bel, Muntsa Padró, Montserrat Marimon and Eva Revilla: Towards the Automatic Merging of Language Resources. In Proceedings of WoLeR 2011. Ljubljana, Slovenia.

Carl Pollard and Ivan A. Sag. 1994. Head-driven Phrase Structure Grammar. The University of Chicago Press, Chicago.

Simone Teufel. 1995. A Support Tool for Tagset Mapping. In EACL-Sigdat 95.