Reusability of a corpus and a treebank to enrich verb subcategorisation in a dictionary

Arantza Díaz de Ilarraza, Koldo Gojenola and Maite Oronoz
Department of Computer Languages and Systems
University of the Basque Country, P.O. box 649, E-20080 Donostia
jipdisaa, jipgogak, jiporanm@si.ehu.es

Abstract

This paper deals with the reusability of a corpus and a treebank to enrich verb subcategorisation in a static resource, a dictionary. Two experiments have been performed to propose: a) new subcategorisation information for verb entries included in the dictionary, and b) new verb entries. For the verb subcategorisation enrichment, inconsistencies between the information obtained from the corpus and the dictionary were found by means of a tool called *Saroi*. The same tool is used to propose new entries. A verb is proposed for its inclusion in the dictionary if it is found in the corpus but not in the dictionary, and it also appears in the treebank.

1 Introduction

This paper deals with the reusability of a corpus and a treebank to enrich verb subcategorisation in a dictionary. Dictionaries are a basic and very rich source of lexical information. However, their creation is very time consuming and sometimes dictionaries do not reflect changes in language usage. Several works have been carried out with the aim of automatically enriching dictionaries. They tackle a great variety of aspects going from the sources from which data was extracted to the output resources to be created. For example, in [8], a dictionary of word combinations was automatically enriched using information extracted by means of a dependency parser. In another work, the Prague Dependency Treebank was used to learn verb subcategorisation frames for Czech by means of machine learning techniques. In [10] frequencies about words were extracted from a corpus and added to the Longman Dictionary.

In our case, the dictionary we want to enrich is a general purpose monolingual dictionary called *Euskal Hiztegia* (EH)[11]. Since its creation the Basque Academy has made new decisions about the standard forms of some words. Moreover, we assume that corpora better reflect the changes in the language.

In order to reduce manual work to the checking of the results, we reuse already developed resources: a) a corpus to extract verbs and their realisation schemas, b) the EH dictionary to obtain verbs and their subcategorisation patterns, and c) the Basque Dependency Treebank. To manage all these resources, we have used a dependency-tree inspection tool called *Saroi*.

Our aim is to enrich the dictionary in two ways; a) adding verb subcategorisation information after looking for inconsistencies between the verbs that appear in a corpus and those that appear in the dictionary, and b) enriching EH with verb entries found in the corpus but that are missing in the dictionary after checking its existence in the treebank. The enrichment proposal lists will be presented to linguists. A *feedback* process has been performed as we use the dictionary to enrich itself.

The remainder of this paper is organised as follows: section 2 describes the used resources; in section 3 we will analyse *Saroi*, a dependency-tree inspection tool; section 4 explains the preprocessing work, and sections 5 and 6 show the performed experiments. Finally, some conclusions are outlined in section 7.

2 Resources

Basque is an agglutinative language with relative free order among sentence elements. In finite verbs, the verb agrees in tense and mood with the subject, object or indirect object of the sentence. As [9] says, "The simplest forms of intransitive verbs are monovalent and mark agreement with the subject (NOR). Intransitive verbs can also have bivalent forms marking agreement with an absolutive argument (subject) and a dative argument (NOR-NORI). Finite transitive verb forms are minimally bivalent, marking agreement with an ergative argument (subject) and an absolutive (direct object) argument (NOR-NORK). In addition, there are trivalent forms that add agreement with a dative argument (NOR-NORI-NORK)". The type of auxiliary verb used by each of these four types of verbs has been pointed out between parentheses. Three different resources are used:

- **Corpus**. It consists of verb realisation schemas obtained as a result of the automatic analysis of a corpus composed of 10,032,133 word-forms taken from a Basque newspaper [4]. A group of 2,541 verbs (including 367 multiword verbs) was extracted from this corpus with the aim of identifying their verbal syntactic pattern or realisation schema. In this list each verb is accompanied by the syntactic components found in its context, together with information about the type of auxiliary verb, and the proportion in which each type of auxiliary verb appears. Table 1 shows the data we extracted for the verb *etorri*.

Etorri "to come"(5649 occurrences)		
Aux. type	#	%
NOR-NORK	2	0.03 %
NOR	5331	94.37 %
NOR-NORI-NORK	0	0 %
NOR-NORI	316	5.59 %

Table 1: *Auxiliary verb types with the verb* etorri.

Corpora offer a vast and complete description of verb structures, nevertheless, as the information is automatically collected, errors can be produced.

- **Dictionary**. All the verb patterns were extracted from the *Euskal Hiztegia* (EH) dictionary.

  We have used a TEI-conformant (*Text Encoding Initiative*) XML version of the dictionary as a source of information about 4,016 verbs. Apart from the headword, we extracted a tag that identifies the kind of auxiliary verb. Possible types are DA, DU, DIO, ZAIO, DA-DU .... The dictionary specifies the senses of each entry word. For most of the verb senses the type of the auxiliary is marked. For example, the verb *eratu* has two senses with an auxiliary mark, and one without it. The sense similar to *konpondu* ("to adapt") carries out a DA type auxiliary, while the second sense, similar to *moldatu, antolatu* ("to repair"), goes with a DU type auxiliary. In this case a combined DA-DU tag is automatically assigned to the verb *eratu* to collect both sense uses. The auxiliary type tags in the dictionary differ from those used in the corpus (see table 2).

Dictionary	Corpus
DA	NOR and NOR-NORI
DU	NOR-NORK and NOR-NORI-NORK
ZAIO	NOR-NORI
DIO	NOR-NORI-NORK

Table 2: *Equivalences between auxiliary verb tags.*

- **Treebank**. The 3LB project annotated corpora for Catalan, Spanish and Basque. The syntactically annotated Basque corpus contains 25,000 word-forms from a reference corpus [1] and 25,000 from newspapers. The corpus used for the treebank and the one for the realisation schemas are disjoint. The treebank was annotated using a dependency framework similar to [5] and the *Conference on Computational Natural Language Learning 2007* format.

We consider these three resources complementary to each other as the corpus reflects the real use of the nowadays language, the dictionary was compiled after a vast manual linguistic analysis, and the treebank combines both viewpoints.

3 A treebank inspection tool

The main goal of the *Saroi* system is to look for linguistic information in dependency-trees by means of rules that express the characteristics of the information we search. This system was also used to look for agreement errors in dependency-trees [6].

The system is composed of three main modules: a) a robust syntactic analyser, b) a rule compiler, and c) a module that coordinates the results of the analyser, applying different combinations of the already compiled rules. The specification language for the description of the rules is abstract, general, declarative, and based on linguistic information. Figure 1 shows the architecture of the system.

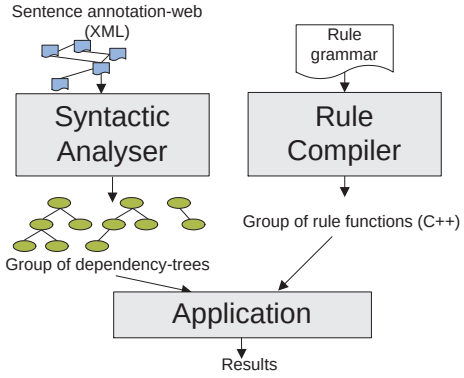

Fig. 1: *Architecture of* Saroi.

3.1 Syntactic analyser

The input of the syntactic analyser module is an annotation-web that follows an XML stand-off markup approach and that represents the linguistic information obtained by the analysis chain. The analysis chain [2] is composed of a morphosyntactic analyser, a tagger/lemmatiser [7], a chunker, and finally, a parser that obtains dependency-trees.

The information gathered in the XML documents that represent the dependency trees is ambiguous. That is, a document can store multiple dependency parses. *Saroi* deals with this ambiguity and creates independent dependency-trees.

In the syntactic analysis module there is an *enrichment module* that carries out two processes: makes explicit the agreement information in auxiliary verbs and enriches main verbs with the information described in section 2. Figure 2 shows part of the morphosyntactic analysis of the verb *etorri* ("to come") after the addition of the information extracted from the corpus (see table 1), and the dictionary.

```
<fs id="V-etorri-1" type="VerbInfo">
 <f name="frequency-features">
  <fs type="verb-frequency">
   <f name="occurrences"><nbr value="5649"/></f>
   <f name="NOR-%"><nbr value="94.37"/></f>
   <f name="NOR-NORK-%"><nbr value="0.03"/></f>
   <f name="NOR-NORI-NORK-%"><nbr value="0"/></f>
   <f name="NOR-NORI-%"><nbr value="5.59"/></f>
  </fs>
 <f name="NOT_in_EH" org="list">
  <sym value="Not_NOR-NORK"/>
  <sym value="Not_NOR-NORI-NORK"/>
 </f>
</fs>
```

Fig. 2: *Part of the analysis of the verb* etorri *after the enrichment process.*

3.2 Rule compiler

The rule grammar that constitutes the input of the rule compiler has been defined by means of a general specification language. The aim of this language is to search for any linguistic structure in a dependency tree. The use of an abstract specification language has several advantages: a) declarativeness, b) maintainability and, c) efficiency, as the abstract rules will be compiled to an object language (C++). The rules allow the traversing of the dependency tree while at the same time checking syntactic constraints.

In the rules we use linguistic information such as tags that define dependency relations between the elements of the sentence (e.g. ncsubj, ncobj,...), as well as tags defining features of the syntactic elements (number, case, ...). Apart from this, some operators have been defined to navigate vertically the dependency-tree and to inspect linguistic features.

The rules, written in an abstract language, cannot be directly applied to a dependency tree because they must first be translated into executable statements. We defined and implemented a syntax-directed translation scheme [3] for that purpose.

4 Preprocessing

Therefore, we have three linguistic data resources with very different origins: a) a group of verbs together with information about the types of auxiliary verbs they appear with, extracted from a corpus, b) another group of verbs with the same information but extracted from a dictionary, and lastly, c) a treebank of correct and standard Basque. In addition, we have a system, *Saroi*, that looks for linguistic information in treebanks. So, we can reuse all these elements to enrich the dictionary. As the enrichment module manages verbal information from different origins, we can use this to obtain different lists of verbs. The verbs obtained from the corpus are 2,541 and those extracted from the dictionary, 4,016, with a total of 5,264 different verbs, showing that not all the verbs appear in both sources:

- 1,248 verbs only appear in the corpus ("Corpus Only, CO"): i) Verbs appearing in journalistic style but not in the dictionary, e.g. *klonatu* ("to clone"), ii) Mistyped verbs, and iii) Multiword verbs that do not appear neither as entries nor as subentries in the dictionary.

- 2,723 verbs are exclusively gathered in the dictionary "Dictionary Only, DO"). Examples of these verbs are those marked in the dictionary as: i) Infrequent verbs, e.g. *urgoitu* ("to get tired"), ii) Dialectal variants, e.g. *haurridetu* ("to make sister cities") used in the French speaking area, and iii) Verb entries marked as highbrow. An example is, *hatsanditu* ("to get out of breath").

- 1,293 verbs appear in both sources, corpus and dictionary ("Both, B").

5 Finding inconsistencies

The resources used for this experiment are the *"Both, B"* list of verbs and *Saroi*. The main objective in this first experiment is to look for inconsistencies between the subcategorisation information that appears in the corpus and in the dictionary. For us an inconsistency occurs when the types of auxiliary verb in the corpus and in the dictionary are different. For example, the verb *zauritu* ("to wound") appears with auxiliaries of type NOR in the corpus and with a DU tag in the dictionary (DU is equivalent to NOR-NORK and NOR-NORI-NORK, see table 2).

5.1 The experiment

Let us see step by step the process followed:

1. Analysis of the *"B"* verb list by means of the analysis chain mentioned in section 3.1.

2. Enrichment of these verbs with the information extracted from the corpus and from the dictionary. After the enrichment process has concluded, each of the verbs will have information similar to the one showed in figure 2.

3. Application of a set of four rules, one for each auxiliary verb type, to the resulting verb list using *Saroi*. Figure 3 shows the rule for detecting inconsistencies in auxiliary verbs of type NOR.

```
RULE INCONSISTENCY_IN_NOR_TYPE
   Detect ( @.pos == 'ADI' & @.occurrences >4 &
            @.NOR-% >50 & @.Not_NOR )
```

Fig. 3: *Detecting inconsistencies in* NOR *auxiliaries.*

The rule in figure 3 can be paraphrased as: mark that a tree fulfils this rule if the current node ('@') has as part of speech (@.pos) ADI (verb), the verb appears in the corpus more than four times and goes with an auxiliary of type NOR with a proportion of more than 50%. But besides this, the entry in the dictionary indicates that the same verb does not usually carry a NOR auxiliary. So, we notice a clear inconsistency between the data extracted from the corpus (the verb appears more than half of the times with the NOR auxiliary) and those extracted from the dictionary (it does not appear with auxiliaries of type NOR).

We have only inspected the verbs with more than 4 occurrences in the corpus to avoid the appearance of mistyped words erroneously marked as verbs. In addition, we think that a clear inconsistency occurs if the proportion of an auxiliary verb in the corpus is more than 50%.

5.2 The results

In a list of 1,293 verbs, 53 (4%) present inconsistencies referring the auxiliary verb. In 45 of the cases (84.9%) there is an inconsistency of type NOR. 6 cases (11.3%)

showed a NOR-NORK inconsistency. 2 times (3.77%) a NOR-NORI difference appears, while no NOR-NORI-NORK inconsistencies are marked.

A priori, we expected a high proportion of NOR type inconsistencies before seeing the results. In Basque, when the verbs are used as impersonal, the ergative argument of the sentence (the subject of the clause) is ellided and verbs of type NOR-NORK turn into NOR. This fact is not reflected in the dictionary.

A linguist made manually a deeper analysis of the inconsistencies and found the following casuistry:

- In 36 of the cases (67.9%) there was a lack of some verb alternation (impersonal, inchoative, ...) in the dictionary. In this case, the alternating syntactic structures in the corpus together with their examples can be added to the dictionary.
- In 11 of the cases (20.7%) the verb usage in the corpus and in the dictionary differs. These are interesting for examining the real verb usage and the reasons for changes in language use.
- 5 errors (9.4%) were identified in the dictionary. We manually verified that when the subcategorisation tag in the dictionary indicated an auxiliary type, examples in the dictionary showed others.
- In one of the cases (1.9%), although the word-form was the same, the senses of the verb in the corpus and in the dictionary were different.

We have observed that from a list of 1,293 verbs 53 (4%) are marked by *Saroi* as inconsistencies. A linguist has confirmed that all the proposals present real inconsistencies, so we have obtained reliable results. The inconsistencies have been used to propose the inclusion of new verb alternations and new verb usage in the dictionary, and confirm the usefulness of the corpus as a source of language use information.

6 Adding new entries

The objective of this second experiment is to enrich EH with new verb entries found in the corpus and the treebank but that are missing in the dictionary. In this case we have used the *"Corpus Only, CO"* list of verbs, and the treebank. We consider that a verb could be proposed to be part of the dictionary if in addition to being in the corpus, it also appears in the treebank. As the treebank was manually tagged and contains correct linguistic information, we think that it offers enough guarantee for the purpose we follow. Treebanks have the advantage of having less noisy data compared to that obtained by automatic parsers.

6.1 The experiment

The process followed to look for verbs that appear in the corpus and in the treebank, but not in the dictionary, is the following one:

1. As we are looking in the treebank for specific verbs lemmas, first, we have automatically created a rule similar to the one in figure 4 for each of the verbs appearing in the corpus (1,248 rules). In the rule in the figure 4, only the nodes in the dependency-trees with the ADI (verb) POS tag are inspected and if we find one with the lemma *"ados etorri"*("to agree") occurring in the corpus more than 10 times, the dependency-tree that fulfils the conditions is marked.
2. The rules are applied to the treebank using *Saroi*.

```
RULE VERB_ADOS_ETORRI
      Detect ( @.pos == 'ADI' &
               @.lemma == 'ados etorri' &
               @.occurrences >10 )
```

Fig. 4: *Detection of a verb in the treebank.*

6.2 The results

Table 3 presents in detail the results of this experiment. The first column shows the candidate verbs. Column 2 indicates the number of occurrences in the corpus while column 3 (Treeb.) shows the number of times in which rules have been activated in the treebank. This column has been divided into two, a) the part of the treebank that is composed of literary texts and, b) the part composed of journalistic texts. Finally, the last column (Propo?) indicates whether an expert proposes or not the verb for its inclusion in the dictionary. The reasons used by the linguist for accepting or rejecting the verbs that appear only in the corpus are diverse:

Verb	**Corp.**	**Treeb.**		**Propo?**	
not in dictionary	*occurs*	*EEBS occurs*	*Journ. occurs*	*Propo?*	*Reason*
baloratu	>50	1	2	Reject	1R
blokeatu	>50	0	5	Accept	1A
diseinatu	>50	0	1	Accept	1A
exijitu	>50	1	2	Doubt	D
inbertitu	>50	0	4	Accept	1A
hitzartu	>50	0	7	Accept	2A
hitzeman	>50	0	2	Accept	2A
kaltetu	>50	0	1	Accept	2A
justifikatu	>50	0	1	Accept	1A
planteatu	>50	2	3	Accept	3A
menperatu	>50	3	0	Reject	2R
afiliatu	>10	0	1	Accept	1A
berdintsu izan	>10	0	1	Doubt	D
deskubritu	>10	4	0	Reject	1R
erlazionatu	>10	1	0	Accept	1A
errebindikatu	>10	0	2	Accept	1A
errekurritu	>10	0	3	Doubt	D
finantzatu	>10	0	6	Accept	1A
kargugabetu	>10	0	1	Accept	1A
kartzelaratu	>10	1	0	Accept	1A
kolaboratu	>10	0	1	Accept	1A
komentatu	>10	1	0	Doubt	D
konplikatu	>10	1	0	Accept	1A
lanpetu	>10	1	0	Reject	3R
ingresatu	>10	0	2	Reject	1R
inkomunikatu	>10	0	6	Accept	1A
inkulpatu	>10	0	1	Reject	4R
inspiratu	>10	1	0	Accept	1A
integratu	>10	1	1	Accept	1A
konprometitu	>10	0	1	Accept	1A
kotizatu	>10	0	1	Accept	1A
kriminalizatu	>10	0	1	Doubt	D
merkaturatu	>10	0	5	Accept	1A
praktikatu	>10	1	0	Accept	1A
profitatu	>10	1	0	Accept	1A

Table 3: *Candidate verbs.*

- A candidate verb is accepted (A) if:
  - 1A. It has been manually looked up in four dictionaries and it is found in at least two.
  - 2A. It does not appear with this word-form in the EH dictionary but appears with a similar form in a subentry. For example, *hitzartu* ("to agree to") does not appear but *hitz hartu* does with the same sense. The linguist proposes the word-form found in the corpus when it appears with the same spelling in most of the dictionaries.
  - 3A. The candidate verb appears in the dictionary but not as the preferred verb. For example, *planteatu* ("to bring up") is marked as *"spanish influenced word"* and *ezarri* is proposed. In the rest of the dictionaries, *planteatu* is a standard entry. So, the linguist proposed the form found in the corpus.
- The reasons for rejecting (R) a candidate verb are:
  - 1R. Another form is preferred in all the rest of the dictionaries.
  - 2R. It does not appear as a dictionary entry but as a variant of the verb.
  - 3R. It does not appear in the dictionary as a verb but as an adjective.
  - 4R. It does not appear in any dictionary.
- The doubtful (D) verbs are those that could be found in only one of the four dictionaries.

Two thresholds were defined for this second experiment. One asking each verb to appear more than 50 times in the corpus, and a second one reducing the number of occurrences to 10. Table 3 shows that a high number of occurrences in the corpus does not necessarily mean a guarantee in the proposal. When the verb occurs in the corpus more than 50 times 72.7% of the verbs is accepted and 18% refused. For the second group of verbs (more than 10 times in the corpus) 66% is accepted and 16% refused. In this second group the number of refused verbs is lower, but the number of those marked as doubtful is higher. We have the impression that the verbs marked as doubtful probably would be accepted but the conditions we have established are quite strict. Besides, the contribution of verbs to the dictionary in the second group is higher.

The verb lists proposed in both experiments can be easily extended. When looking for inconsistencies, we could reduce the number of occurrences in the corpus, obtaining more inconsistencies. In the case of verb entries, asking, for example, 5 occurrences in the corpus, the proposed list will probably be larger.

7 Conclusions

This work examines the validity of corpora and treebanks in the enrichment of a more static resource, a dictionary. We have explored two different alternatives to enrich verbal information in a dictionary using both an unannotated corpus and a treebank. The experiments have been designed to obtain on the one hand, a list of verbs that already exist in the dictionary but that present inconsistencies with verbs found in a corpus and, on the other hand, a list of verbs found in the corpus and treebank but that are missing in the dictionary.

By reusing already existing resources, the work carried out to obtain results from the corpus as well as the one to enrich the dictionary, usually a very time consuming task, has been reduced to the minimum.

The experiment for including verb entries in the dictionary shows that, regardless of the threshold used in the corpus, all the verbs appearing more than 4 times in the treebank composed of newspapers are accepted. This part of the treebank combines the actual use of the language with linguistic correctness. The acceptance level of verbs demonstrate the validity of treebanks as information source.

We think that the methodology we use is general for any language although it has a twofold implication: a) the appropriate resources must be available, and b) the linguistic information must be represented following the input specifications for *Saroi* (a general dependency representation in XML).

We are of the opinion that this work is extensible to the rest of words in this dictionary (i.e. *nouns*, *adjectives,* ...). Information concerning POS, examples, or usage domain could be added to the dictionary. Changing the source corpus, domain specific words could also be added.

Acknowledgments. This research is supported by the Univ. of the Basque Country (GIU05/52) and the Basque Government (ANHITZ project, IE06-185). Thanks to I. Aldezabal and A. Atutxa for their invaluable help.

References

[1] I. Aduriz, M. Aranzabe, J. M. Arriola, A. Atutxa, A. Díaz de Ilarraza, N. Ezeiza, K. Gojenola, M. Oronoz, A. Soroa, and R. Urizar. Methodology and steps towards the construction of epec, a corpus of written basque tagged at morphological and syntactic levels for the automatic processing. In *Corpus Linguistics Around the World*. Rodopi, 2006.

[2] I. Aduriz, M. Aranzabe, J. M. Arriola, A. Díaz de Ilarraza, K. Gojenola, M. Oronoz, and L. Uria. A cascaded syntactic analyser for basque. In A. Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing: 5th Int. Conf. CICLing2004, Korea*, volume 2945 of *Lecture Notes in Computer Science*, pages 124–134. Springer-Verlag GmbH, 2004.

[3] A. V. Aho, R. Sethi, and J. D. Ullman. *Compilers: Principles, Techniques, and Tools*. Addison-Wesley, 1985.

[4] I. Aldezabal, M. Aranzabe, A. Atutxa, K. Gojenola, M. Oronoz, and K. Sarasola. Application of finite-state transducers to the acquisition of verb subcategorization information. *Natural Language Engineering. Cambridge University Press.*, 9(1):39–48, 2003.

[5] J. Carroll, G. Minnen, and T. Briscoe. Corpus annotation for parser evaluation. In *EACL 99 workshop on Linguistically Interpreted Corpora (LINC)*, pages 35–41, Norway, 1999.

[6] A. Díaz de Ilarraza, K. Gojenola, and M. Oronoz. Design and development of a system for the detection of agreement errors in basque. In A. Gelbukh, editor, *Computational Linguistics and Intelligent Text Processig: 6th Int. Conf. CICLing2005, Mexico*, volume 3406 of *Lecture Notes in Computer Science*, pages 793–803. Springer-Verlag GmbH, 2005.

[7] N. Ezeiza. *Corpusak ustiatzeko tresna linguistikoak. Euskararen etiketatzaile sintaktiko sendo eta malgua*. PhD thesis, University of the Basque Country, Donostia, 2003.

[8] A. Gelbukh, G. Sidorov, S.-Y. Han, and E. Hernández-Rubio. Automatic enrichment of very large dictionary of word combinations on the basis of dependency formalism. *Lecture Notes in Artificial Intelligence*, (2972):430–437, 2004.

[9] J. I. Hualde and J. O. de Urbina. *A grammar of Basque*. 2001.

[10] A. Kilgarrif. Putting frequencies in the dictionary. *International Journal of Lexicography*, 10(2):135–155, 1997.

[11] I. Sarasola. *Euskal Hiztegia*. Donostia, 1996.