LEARNING ARGUMENT/ADJUNCT DISTINCTION FOR BASQUE

Izaskun Aldezabal, M.ª Jesús Aranzabe, Koldo Gojenola,
Kepa Sarasola, Aitziber Atutxa

Abstract

This paper presents experiments performed on lexical knowledge acquisition in the form of verbal argumental information. The system obtains the data from raw corpora after the application of a partial parser and statistical filters. We used two different statistical filters to acquire the argumental information: Mutual Information, and Fisher's Exact test. Due to the characteristics of agglutinative languages like Basque, the usual classification of arguments in terms of their syntactic category (such as NP or PP) is not suitable. For that reason, the arguments will be classified in 48 different kinds of case markers, which makes the system fine grained if compared to equivalent systems developed for other languages.

This work addresses the problem of learning subcategorization frames by distinguishing arguments from adjuncts, being the last ones the most significant source of noise in subcategorization frame acquisition.

Introduction

In recent years a considerable effort has been done on the automatic acquisition of lexical information. As several authors point out, this information, mostly subcategorization information, is useful for a wide range of applications. For example, Carroll et al. (1998) show how adding subcategorization information improves the performance of a parser (automatic syntactic analyzer). With this in mind, our aim is to build a system that automatically obtains subcategorization frames. The following figure shows the general schema of a subcategorization acquisition system.

The basic idea behind any system like the one presented in this paper is the following. Starting from a corpus, syntactic information is attained as a result of a parsing phase. As a consequence, each verb will get a set of frames assigned to it. These frames represent the different syntactic environments in which the verb appeared in the corpus. Once these frames are available, statistical filters apply to distinguish subcategorized elements from non-subcategorized ones. As we can see in figure 1. there are two ways to perform this filtering. (A) consists in applying the filters to verb-case pairs to distinguish subcategorized elements (arguments) and non-subcategorized ones (adjuncts). (B) consists in applying statistics directly to the frames to distinguish

[ASJU Geh 46, 2003, 75-93]
http://www.ehu.es/ojs/index.php/asju

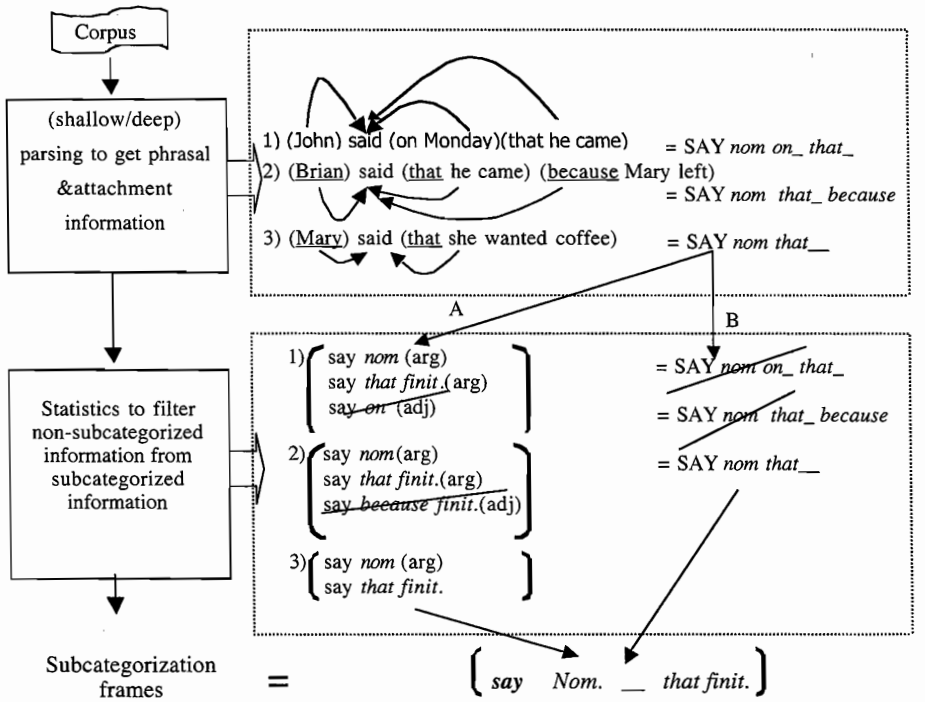

Figure 1. A general schema for a subcategorization acquisition system

subcategorization frames from appearing frames. Following the first filtering, an additional step is required to achieve subcategorization frames; to go back to the original frames and eliminate the elements considered to be adjuncts, because original frames without adjuncts are supposed to be subcategorization frames. The second filtering yields to subcategorization frames directly. The system presented here employs the first filtering approach. We will explain the reasons for this choice in section 2.2.

As we just said, the statistical filters included in the system will perform the argument/adjunct discrimination. But it is well known that this is not a trivial task since there is no clear cut between arguments and adjuncts. However, we decided to pursue it, but under certain limitations, both theoretic and pragmatic.

As for the evaluation, we first evaluated performance of the statistical filters in the argument/adjunct distinction. Second, we evaluated the quality of whole acquired subcategorization frames. We approached the first evaluation (the filter evaluation for the argument/adjunct distinction) in two different fashions; one way consisted in evaluating the resulting list of verb-case marker pairs (tagged either as argument or adjuncts), with the values a human would assign to each verb-case marker pair in the list automatically obtained. Note that the annotator did not have more context than the list of verb and the cases. The second way consisted in selecting some sentences and evaluating over these sentences (that is to say, within a sentential context); again, the statistical filter marked each case phrase the parser attached to the verb in the sentence as argument or adjunct. We compared this marking with the values (argument/adjunct)

assigned by the human annotator to those same verb and case-phrases, but note that the annotator was provided with the sentence and therefore could make use of the sentential context to establish the meaning of the verb. Both methods of evaluation yield significantly different results. Evaluating this way, we wanted to reach some conclusions on the importance of the context for the argument adjunct distinction task. Finally, we also evaluated subcategorization frames obtained using the results of the statistical filter by manually annotating each subcategorization frame obtained by the machine as correct or incorrect. In this case we did not make use of the sentential context.

The paper is divided into five sections. The first section is devoted to explain the theoretical motivations underlying the process. The second section is a description of the different stages of the system. The third and fourth sections present the results obtained by the application of statistical filters to discriminate arguments from adjuncts, and the results of the whole subcategorization frame acquisition, respectively. The fifth section reviews previous work on automatic subcategorization acquisition. And last but not least, we present the main conclusions.

1. The argument/adjunct distinction

As said before, Carroll et al. (1998) showed how adding subcategorization information improves the performance of a parser. Their experiment was developed for English, which is considered to be a fix word order language and head initial.

1. Josuk alde egin zuen etxetik bere amarekin jateko.
   Josu-erg left aux home-from his mother-with eat-to

   "Josu left home to eat with his mother"

2. Josuk alde egin zuen seietan bere amarekin jateko
   Josu-erg left aux. six-at his mother-with eat-to

   "Josu left at six to eat with his mother" or "Josu left to eat with his mother at six"

Both *etxetik* (from home) and *seietan* (at six) are postpositional phrases superficially appearing in between *alde egin* (to leave) and *jateko* (to eat), so in principle, and without the help of any subcategorization information, the parser would not be able to decide where to attach in each case. It would treat both the same way. Either it would consider that in both cases these intermediate postpositional phrases are attached to both verbs, or either it would have to make a heuristic decision. For example attach them to the first verb.

Subcategorization information would allow performing the right attachment of the ablative case (*from*) since the ablative is subcategorized by *alde_egin* (to leave) and not by *jan* (to eat). It would also make possible to attach correctly the inessive case to both verbs because the inessive case (at) is not subcategorized by either *alde egin* (to leave) or *jan* (to eat). At this point, we hope we have shown the importance of learning and applying subcategorization information. But such an enterprise is as difficult as important. The argument/adjunct distinction is probably one of the most unclear issues in linguistics. The distinction was presented in the early days in the following

way: subcategorized elements (arguments) are those elements appearing obligatorily while non subcategorized elements (adjuncts) are not. Nowadays we know that this definition is too naive. Several problematic cases are not considered under this definition, for example under-specified elements, elements showing dative case, object shift constructions and so on.

a. Under specified elements

3. I arrived safely.

In principle, *arrive* is taken to be an unaccusative verb, with a single argument.

4. I arrived safely at the station

But in this sentence, *at the station* seems to be an argument too.

b. Object shift constructions

5. I loaded the truck
6. I loaded bricks on the truck

Would we say that the subcategorization is different for these last two cases?

Another definition considers as subcategorized elements those ones participating in the event and as non subcategorized those ones contextualizing or locating the event. This is a semantic definition of what an argument is.[1] It is still not clear, in example 4, whether *at the station* would be an argument or an adjunct. One could say that it participates in the event since it marks the end of the event. Under some aspectual thesis (Tenny 1987) both *the truck* and *on the truck* could also be considered as participants of the event, again because they mark the end of the event. But leaving aside aspectual issues, take a look to the following examples:

7. Yesterday I talked with Mary.
8. Yesterday I played soccer with Mary.

Here, *Mary* is a participant of the event in both cases, therefore under the given definition in both cases *Mary* would be a subcategorized element. But this is contradictory to what traditional views consider in practice. *To play* does not require two participants (though it can have them), while *to talk* (under the sense of communicating) seems to require two participants.

Finer argument/adjunct distinctions have also been proposed differentiating between basic arguments, pseudo-arguments and adjuncts. Basic arguments are those required by the verb. Pseudo-arguments are those that even if they are not required by the verb, when appearing they extend the verbal semantics, for example, adding new

[1] It would be also syntactic because depending wether it is a participant or not the elements will get projected in different positions (external or internal predication).

participants. And finally adjuncts, which would be contextualizers of the event. (For further reference on the argument/adjunct distinction see Gawron 1986, Grimshaw 1990, Schutze 1995, Verspoor 1997).

Though there is an extensive literature on subcategorization, up to day, we did not find a way to establish a clear cut between subcategorized and non subcategorized elements. Nevertheless, from the different diagnostics proposed in the literature some are quite consistent among various authors (Pollard and Sag 1987, Grishman et al. 1994, Verspoor 1997):

1. Obligatoriness condition. When a verb demands obligatorily the appearance of an element, this element is an argument.

   a. John put the book *on the table*
   b. *John put the book

2. Frequency. Arguments of a verb occur more frequently with that verb than with the other verbs.

   a. I came *from home* (argument).
   b. I heard it *from you* (adjunct).

3. Iterability. Several instances of the same adjunct can appear together with a verb, while several instances of an argument cannot appear with a verb.

   a. I saw you in Washington, in the Kenedy Center.
   b. *I saw you in Washington, in N.Y.

4. Relative order. Arguments tend to appear closer to the verb than adjuncts.

   a. I put the book on the table *at three*
   b. *I put *at three* the book on the table

5. Implicational test. Arguments are semantically implied, even when they are optional.

   a. I came to your house (from x)
   b. I heard that (from x)

The third and fourth tests were not very useful to us. Iterability test is quite weak since it seems to rely more on some other semantic notions such as part/whole relation than on the argument/adjunct distinction. For example, sentence 3.a would be grammatical due to semantic plausibility. *The Kennedy Center* is part of *Washington*, therefore to see somebody in *the Kennedy Center* and see him in *Washington* are not semantically incompatible, so it is plausible to say it. In the case of 3.b *N.Y* is not a part of *Washington* and therefore it is not plausible to see (in the same event) somebody in two different places.

The relative order test is difficult to apply on a free word order language like Basque. The first and fifth tests are robust enough to be useful in practice. But only the two first diagnostics can be captured statistically by the application of association measures like Mutual Information. We did not come out with any straightforward way to apply the fifth test computationally.

Before introducing the different statistical measures applied, we will present step by step the whole process we pursued for achieving the argument/adjunct distinction. Talking about Subcategorization Frames (SCF) means talking about arguments. Many existing systems acquire directly a set of possible SCFs without any previous filtering of adjuncts. However, adjuncts are a substantial source of noise and sparseness.[2] If we wanted to acquire directly the right subcategorization frames without making any previous filtering we would need more than the million and a half words that we have. The reason is that on the basis of verb-case markers (of course obtained from the frames appearing in the corpus) we can apply some statistics because the arguments appear more frequently than adjuncts, because they appear in more frames, and the frequency distinction is usually relevant enough as to be able to apply statistics on it. But it is not so frequent to see a bare real subcategorization frame (in other words, a frame where all the cases are only arguments). In most of the cases there is an adjunct, and moreover the range of different adjuncts is huge. This means that the argument and adjunct combination number into frames is very high besides, the frequency distinction between the combinations is not relevant enough. Therefore we decided to pursue the argument/adjunct distinction as a way to obtain real subcategorization frames (option A in Figure 1).

2. The acquisition process

Our starting point was a raw newspaper corpus from of 1.337.445 words, where there were instances of 1.412 verbs. From them, we selected 640 verbs as statistically relevant because they appear in more than 10 sentences.

As we said earlier, our goal was to distinguish arguments from adjuncts. When starting from raw corpus, like in this case, it is necessary to get instances of verbs together with their dependents (arguments and adjuncts). We obtained this information applying a partial parser (section 2.1) to the corpus. Once we had the dependents, statistical measures helped us deciding which were arguments and which were adjuncts (section 2.2).

2.1. The parsing phase

Aiming to obtain the data against which statistical filters will be applied, we analyzed the corpus using several available linguistic resources (for more information see Aldezabal et al., in this volume):

— First, we performed morphological analysis of the corpus, based on two-level morphology (Koskenniemi 1983, Alegria et al. 1996) and disambiguation using the Constraint Grammar formalism (Karlsson et al. 1995, Aduriz et al. 1997).

[2] When the frequency of an event is too distributed into different occurrences, and the frequency of each occurrence is very similar. So statistically there is no occurrence that is more significant than the others.

— Second, a shallow parser was applied (Aldezabal et al. 2000), which recognizes basic syntactic units including noun phrases, prepositional phrases and several types of subordinate sentences.

1. ... *(a) [EEBBetako lehendakariak] (b) [UEko 15 herrialdeetako merkataritza ministroekin] (c) [bazkaldu zehar zuen] (d) [negoziazioen bilgunean]...*
2. ... the president of the USA had to eat with the ministers of Commerce of 15 countries of the UE in the negotiation center...

a) *[EEBB-etako lehendakari-a-k]*
[USA-of president-the-erg.]

NP-ergative (president, singular)
The president of the USA

b) *[UE-ko 15 herrialde-etako merkataritza ministro-ekin]*
[UE-of 15 countries-of Commerce ministers-with]

PP (with)-commitative (minister, plural)
with the ministers of Commerce of 15 countries of the UE

c) *[bazkaldu behar zuen]*

[to eat had]
verb (eat)
had to eat

d) *[negoziazio-en bilgune-an]*
[negotiation-of center-in]

PP (in)-inessive (center, singular)
in the negotiation center

Figure 2. Example of the output of the shallow parsing phase: 1) Input (in Basque), 2) English translation, Below (c) Verb phrase and (a,b,c) verbal dependents (phrases), and also case+head information

— The third step consisted in linking each verb and its dependents. Basque lacks a robust parser as in (Briscoe and Carroll 1997, Kawahara et al. 2001) and, therefore, we used a finite state grammar to link the dependents (both arguments and adjuncts) with the verb (Aldezabal et al. 2001). This grammar was developed using the Xerox Finite State Tool (Karttunen et al. 1997). Figure 2 shows the result of the parsing phase. In this case, both commitative and inessive cases (PPs) are adjuncts, while the ergative NP is an argument.

The linking of dependents to a verb is not trivial considering that Basque is a language with free order of constituents, and any element appearing between two verbs could be, in principle, dependent on any of them. Many problems must be taken into account, such as ambiguity and determination of clause boundaries, among others. We evaluated the accuracy up to this point, obtaining a precision over dependents of 87% and a recall of 66%. So the input data to the next phase was relatively noisy.

2.2. The argument selection phase

In the data resulting from the shallow parsing phase we counted up to 65 different cases (types of arguments, including postpositions and different types of suffixes). These are divided in two main groups:

— 43 correspond to postpositions. Some of them can be directly mapped to English prepositions, but in many cases several Basque postpositions correspond to just one English preposition. This set also contains postpositions that map to categories other than English prepositions, such as adverbs.
— 22 types of sentential complements (For instance, English *that* complementizer corresponds to several subordination suffixes: *-la, -n, -na, -nik).*

This shows to which extent the range of arguments is fine grained, in contrast to other works where the range is at the categorial level, such as NP or PP (Brent 1993, Manning 1993, Merlo and Leybold 2001).

Due to the complexity carried by having such a high number of cases, we decided to gather postpositions that are semantically equivalent or almost equivalent (for example, English *between* and *among*). Even if there are some semantic differences between them they do not seem to be relevant at the syntactic level. Some linguists were in charge of completing this grouping task. Even considering the risk of making mistakes when grouping the cases, we concluded that the loss of accuracy due to having too sparse data (consequence of having many cases) would be worse than the noise introduced by any mistake in the grouping. The resulting set contained 48 cases. The complexity is reduced but it is still considerable.

Most of the work on automatic acquisition of subcategorization information (Carroll and Briscoe 1997, Sarkar and Zeman 2000, Korhonen 2001) apply statistical methods (hypothesis testing). Basically the idea is the following: they get "possible subcategorization frames" from automatically parsed data (either completely or partially parsed) or from a manually annotated corpus. Afterwards a statistical filter is employed to decide whether those "possible frames" are or not real subcategorization frames (option B in Figure 1). These statistical methods can be problematic mostly because they perform badly on sparse data. In most of the cases the systems pursuing this approach (option B) are able to decrease the noise because they already have some subcategorization information coming from dictionaries (Carroll and Briscoe 1997). In our case, there is no dictionary carrying such information, therefore and in order to avoid as much as possible data sparseness, we decided to design a system that starts learning the arguments/adjuncts of a given verb instead of learning whole frames. Frames are combinations of arguments, and considering that our system deals with 48 cases, the number of combinations was high, resulting in sparse data. So we decided to work at the level of the argument/adjunct distinction. Working on this distinction is also very useful to avoid noise in the subcategorization frame, since in this task adjuncts are synonyms of noise. A system that tries to get subcategorization frames without previously making the argument/adjunct distinction suffers of having sparse and noisy data.

To accomplish the argument/adjunct distinction we applied two measures: Mutual Information (MI), and Fisher's Exact Test (for more information on these measures, see

Manning and Schütze 1999). MI is a measure coming from Information Theory, defined as the logarithm of the ratio between the probability of the co-occurrence of the verb and the case, and the probability of the verb and the case appearing together calculated from their independent probability.

$$MI = \log \frac{P(\text{verb, case})}{P(\text{verb})\, P(\text{case})}$$

So higher Mutual Information values correspond to higher associated verb and cases (see table 1).

Table 1. Examples from MI values for verb-case pairs

Verb	case	MI
atera (to take/go out)	Ablative (from)	1,830
atera (to take/go out)	instrumental (with)	–0,955
erabili (to use)	*gisa* (as)	2,255
erabili (to use)	instrumental (with)	–0,783

Mutual Information shows higher values for *atera-ablative* (to go/take out), *erabili-gisa* (to use-as). These pairs were manually tagged as arguments, therefore Mutual information makes the right prediction. On the contrary, *atera-instrumental* (to go/take out-with), *erabili-instrumental* (to use-with) were manually tagged as adjuncts. Mutual Information values in table 1 go along with the manual tagging for these last pairs as well, because these Mutual Information values are low as should correspond to adjuncts.

Fisher's Exact Test is a hypothesis testing statistical measure.[3] We used the left-side version of the test (see Pederssen 1996). Under this version the test tells us how likely it would be to perform the same experiment again and be less accurate. That is to say, if you were repeating the experiment and there were no relation between the verb and the case, you would have a big probability of finding a lower co-occurrence frequency than the one you observed in your experiment. So higher left-side Fisher values tell us that there is a correlation between the verb and the case (see table 2.)

Fisher's Exact values show higher values for *atera-ablative* (to go/take out), *erabili-gisa* (to use-as). These values predict correctly the association between the verbs and cases for these examples. The low values for the *atera-instrumental* (to go/take out-with), and *erabili-instrumental* (to use-with) pairs, should be interpreted as the non-association between the verbs and the cases in these examples, that is to say, they are adjuncts. And again, the prediction would be right according to the annotators.

[3] There are two ways of interpreting Fisher's test, as one or two sided test. In the one sided fashion there is still another interpretation, as a right or left sided test.

Table 2. Examples of Fisher's Exact Test values for verb-case pairs

Verb	case	Fisher
atera (to take/go out)	Ablative (from)	1,0000
atera (to take/go out)	instrumental (with)	0,0003
erabili (to use)	*gisa* (as)	1,0000
erabili (to use)	instrumental (with)	0,0002

These tests are broadly used to discover associations between words, but they show different behaviour depending on the nature of the data. We did not want to make any a priori decision on the measure employed. On the contrary, we aimed to check which test behaved better on our data.

3. Evaluation of the argument/adjunct distinction

We found in the literature two main approaches to evaluate a system like the one proposed in this paper (Briscoe and Carroll 1997, Sarkar and Zeman 2000, Korhonen 2001):

— Comparing the obtained information with a gold standard.
— Calculating the coverage of the obtained information on a corpus. This can give an estimate of how well the information obtained could help a parser on that corpus.

Under the former approach a further distinction emerges: using a dictionary as a gold standard, or performing manual evaluation, where some linguists extract the arguments in a corpus (this would be the gold standard) and compare them with the arguments obtained automatically.

We decided to evaluate the system both ways, that is to say, using a gold standard and calculating the coverage over a corpus. The intention was to determine, all things being equal, the impact of doing it one way or the other.

3.1. Evaluation 1: comparison of the results with a gold standard

From the 640 analyzed verbs, we selected 10 for evaluation. For each of these verbs we extracted from the corpus the list of all their dependents. The list was a set of bare verb-case pairs, that is, no context was involved and, therefore, as the sense of the given verb could not be derived, different senses of the verb were taken into account. We provided 4 human annotators/taggers with this list and they marked each dependent as either argument or adjunct. The taggers accomplished the task three times. Once, with the simple guideline of the implicational test and obligatoriness test, but with no further consensus. The inter-tagger agreement was low (57%). The taggers gathered and realized that the problem came mostly from semantics. While some taggers tagged the verb-case pairs assuming a concrete semantic domain the others took into account a wider rage of senses (moreover, in some cases the senses did not even match). So the tagging was repeated when all of them considered the same semantics to the different

verbs. The inter-tagger agreement raised up to a 80%. The taggers gathered again to discuss, deciding over the non clear pairs.

The list obtained from merging[4] the 4 lists in one is taken to be our gold standard. Notice that when the annotators decided whether a possible argument was really an argument or not, no context was involved. In other words, they were deciding over bare pairs of verbs and cases. Therefore different senses of the verb were considered because there was no way to disambiguate the specific meaning of the verb. So the evaluation is an approximation of how well would the system perform over any corpus. Table 3 shows the results in terms of Precision and Recall.

Table 3. Results of Evaluation 1 (context independent)

	Precision	Recall	F-score
MI	62%	50%	55%
Fisher	64%	44%	52%

Precision measures from the elements marked by the machine as arguments, how many where really arguments, in other words, how many where also tagged as arguments by the human annotators. In this case it tells us that from the elements marked as arguments using MI, 62% were real arguments, the rest either were adjuncts or attachment errors made by the parser that have been considered by the machine as arguments (or elements which were not well attached). As for the elements marked as arguments using Fisher, 64% were real argument, the rest adjuncts or errors. Recall measures, how many of the elements marked as arguments by the human annotators were not marked as such by the machine. That is, how many of the real arguments were left out. F-score is just a way to normalize both precision and recall, so for example MI gets better recall results than Fisher, and Fisher gets better precision results than MI. F-score provides a way to select which one is relatively better considering both precision and recall.

3.2. Evaluation 2: calculation of the coverage on a corpus

The initial corpus was divided in two parts, one for training the system and another one for evaluating it. From the fraction reserved for evaluation we extracted 200 sentences corresponding to the same 10 verbs used in the "gold standard" based evaluation. In this case, the task carried out by the annotators consisted in extracting, for each of the 200 sentences, the elements (arguments/adjuncts) linked to the corresponding verb. Each element was marked as argument or adjunct. Note that in this case the annotation takes place inside the context of the sentence. In other words, the verb shows precise semantics.

We performed a simple evaluation on the sentences (see table 4), calculating precision and recall over each argument marked by the annotators.[5] For example, if a

[4] Merging was possible once the annotators agreed on the marking of each element.

[5] The inter-tagger agreement in this case was of 97%.

verb appeared in a sentence with two arguments and the statistical filters were recognizing them as arguments, both precision and recall would be 100%. If, on the contrary, only one was found, then precision would be 100%, and recall 50%.

Table 4. Results of Evaluation 2 (inside context)

	Precision	Recall	F-score
MI	93%	97%	95%
Fisher	93%	93%	93%

3.3. Discussion

It is obvious that the results attained in the first evaluation are different than those in the second one. The origin of this difference comes mostly, on one hand, from semantics and, on the other hand, from the nature of statistics:

— Semantic source. The former evaluation was not contextualized, while the latter used the sentence context. Our experience showed us that broader semantics (non-contextualized evaluation) leads to a situation where the number of arguments increases with respect to narrower (contextualized evaluation) semantics. This happens because in many cases different senses of the same verb require different arguments. So when the meaning of the verb is not specified, different meanings have to be taken into account and, therefore, the task becomes more difficult.

— Statistical reason. The disagreement in the results comes from the nature of the statistics themselves. Any statistical measure performs better on the most frequent cases than on the less frequent ones. In the first experiment all possible arguments are evaluated, including the less frequent ones, whereas in the second experiment only the possible arguments found in the piece of corpus used were evaluated. In most of the cases, the possible arguments found were the most frequent ones.

At this point it is important to notice that the system deals with non-structural cases. In Basque there are three structural cases (*ergative, absolutive* and *dative*) which are special because, when they appear, they are always arguments. They correspond to the subject, direct object and indirect object functions. These cases are not very conflictive when deciding on their argumenthood,[6] mainly because in Basque the auxiliary bears information about their appearance in the sentence. So they are easily recognized and linked to the corresponding verb. That is the reason for not including them in this work. Precision and recall would improve considerably if they were included because they are the most frequent cases (as statistics perform well over frequent data), and also because the shallow parser links them correctly using the information carried by the auxiliary. Notice that we did not incorporate them because

[6] As we said in section 1, the nature of the dative case is not very clear.

our aim is to use the subcategorization information obtained to help our parser, and the non-structural cases are the problematic ones.

4. Eliminating the adjuncts from then original frames

Until now we presented a part of the system which is able to decide whether a case phrase corresponds to an argument or an adjunct[7] by means of the occurrence frequency of verb-case pairs in the data. Next step consisted in going back to the original case frames obtained by the partial parser, and eliminating the cases tagged by the machine as adjuncts. Remember that the partial parser tries to attach the case phrases surrounding the different verbs to the corresponding verb. This way, for each verb in a sentence, the parser will provide a frame, or in other words, the combination of case phrases attached to it. This is what we would call an original frame. We used the list resulting from the application of MI. Thus, for example, take *bazkaldu* (to have lunch). The frames obtained by the parser are the following ones:

1. occurrences ### 8,3 DU-erg[8]
2. occurrences ### 3,8 DU-erg-ine
3. occurrences ### 2,9 DU-erg-soc
4. occurrences ### 1 DA-abs-ala
5. occurrences ### 1 DU-abl-erg-ine
6. occurrences ### 1 DU-abs-erg
7. occurrences ### 1 DU-abs-erg-ine-soc
8. occurrences ### 0,7 DA-abs
9. occurrences ### 0,2 DA-abs-ine
10. occurrences ### 0,1 DA-abs-soc

Figure 3. Frames obtained by the parser for the verb *bazkaldu* (to eat)

As we said before, by applying the statistical filters the system got for each verb the list of arguments and adjuncts.

bazkaldu ine: 0,504482
bazkaldu soc: 2,065221
bazkaldu ala: 0,210678
bazkaldu abl: 0,430152

Figure 4. List of arguments/adjuncts obtained by the parser for the verb *bazkaldu* (to eat)

[7] Or an error coming from the heuristics applied by the parser to attach the different phrases to the verbs.

[8] Remember that we did not recover the absolutive case when the auxiliaries are DU or DIO since it is quite usual to find incorporation of the internal argument into the verb with some transitive verbs.

These mutual information values tell us that *sociative case* (with) is an argument *bazkaldu(to have lunch)*.[9] Now, as said before, all cases but the *sociative* (with) will be eliminated from the initial frames, and the result is:

1. occurrences ### 13,1 DU-erg[10]
2. occurrences ### 2,9 DU-erg-soz
3. occurrences ### 1 DU-abs-erg
4. occurrences ### 1 DU-abs-erg-soz
5. occurrences ### 1,9 DA-abs
6. occurrences ### 0,1 DA-abs-soz

Figure 5. Frames obtained for the verb *bazkaldu* (to eat) after eliminating adjuncts

4.1. Evaluation

Once, we got these new frames, our goal was to see if these new frames could be considered as the real subcategorization frames. We know that certain cases are always adjuncts for a given verb, but there are also some cases acting either as arguments or adjuncts depending on the frames they appear in. More over, sometimes the frame in which that case is an argument, and the frame in which that same case acts as an adjunct belong to two different meanings of the verb. For example, consider the following case and frames:

atera 6,95061728395062 ### DA-abs-ala-ine
atera 41,3703703703704 ### DA-abs-ine

Figure 6. Examples of frames obtained for the verb *atera* (to go out/to publish)

If the machine was marking the *inessive case* as adjunct, we would go back to these frames and erase the *inessive case* from them, without making a further distinction. The problem comes from the meaning associated to each of these frames. When looking at the examples we noted that for the first frame the *inessive case* is really an adjunct

[9] In this case, it seems that the machine makes a mistake, but when we take a look to the examples one realizes that eat appears meaning *to gather* or *to meet*. So we will go back to the original frames, and the other cases will be eliminated.

[10] Remember that we did not recover the absolutive case when the auxiliaries are DU or DIO since it is quite usual to find incorporation of the internal argument into the verb with some transitive verbs.

because it is associated to *atera* (to go out) and as movement verb the inessive acts as an adjunct. Contrastively, for the second frame, the meaning changes and *atera* is not a movement verb, it would be equivalent to the English *to publish*. In this case, *inessive* would not be an adjunct but an argument.

Going back to the evaluation, the results were obtained as follows: the manual annotators were provided with both the list of these new frames obtained by the machine by deleting the adjuncts and the list of the original frames obtained initially by the parser. The annotators were marking in both lists each frame as correct or incorrect for the given verb. This time they did not have any sentential context to make the decision, again the decisions were made over raw lists of verbs and frames, therefore they could not know the meaning of the verb associated to each frame.

Table 5. Results of the frames evaluation

	Precision	Recall	F-score
Eliminate adjuncts from initial comb. (688 → 144)	52%	75%	61%

In this case precision expresses how many frames, from the number of frames the machine marked as subcategorization frames, are really subcategorization frames. Therefore one could say that precision meassures the quality of the data obtained. Recall meassures how many real subcategorization frames were discovered by the machine. For doing that, we take the original list of frames got initially by the parser and we tagged them as real subcategorization frames or errors. And recall was calculated by taking the number of frames marked as real subcategorization frames from the list obtained after eliminating the adjuncts and dividing this number by the number of frames marked as real subcategorization frames in the original list. This way we can get an idea of the lost of information when eliminating the adjuncts.

4.2. Discussion

The approach of eliminating the adjuncts is useful for acquiring subcategorization frames. We were able to reduce sparseness. After eliminating the adjuncts the total number of frames decreased from 688 to 144. This happens because once the adjuncts are eliminated, we found a lot of combinations that were different because of an adjunct and once that adjunct disappeared, the frames could be merged because they were the same frame. So the frequencies linked to them could be added. This way, we are able to get relevant frequency distinctions and a lower number of case combinations (frames) for each verb.

We also have to consider the loss of information. As the recall measure shows we lost 25% of subcategorization frames. That means that when eliminating adjuncts, due to errors, we eliminated arguments also, and therefore we lost correct subcategorization frames that were originally captured before the argument/adjunct filtering occurred.

5. Related work

Concerning the acquisition of verb subcategorization information, there are proposals ranging from manual examination of corpora (Grishman et al. 1994) to fully automatic approaches. Table 6, partially borrowed from Korhonen (2001), summarizes several systems on subcategorization frame acquisition.

Manning (1993) presents the acquisition of subcategorization frames from unlabelled text corpora. He uses a stochastic tagger and a finite state parser to obtain instances of verbs with their adjacent elements (either arguments or adjuncts), and then a statistical filtering phase produces subcategorization frames (from a set of previously defined 19 frames) for each verb.

Briscoe and Carroll (1997) describe a grammar based experiment for the extraction of subcategorization frames with their associated relative frequencies, obtaining 76,,6% precision and 43,4% recall. Regarding evaluation, they use the ANLT and COMLEX Syntax dictionaries as gold standard. They also performed evaluation of coverage over a corpus. For our work, we could not make use of any previous information on subcategorization, because there is nothing like a subcategorization dictionary for Basque.

Sarkar and Zeman (2000) report results on the automatic acquisition of subcategorization frames for verbs in Czech, a free word order language. The input to the system is a set of manually annotated sentences from a treebank, where each verb is linked with its dependents (without distinguishing arguments and adjuncts). The task consists in iteratively eliminating elements from the possible frames with the aim of removing adjuncts. For evaluation, they give an estimate of how many of the obtained frames appear in a set of 500 sentences where dependents were annotated manually, showing an improvement from a baseline of 57% (all elements are adjuncts) to 88%. Comparing this approach to our work, we must point out that Sarkar and Zeman's data does not come from raw corpus, and thus they do not deal with the problem of noise coming from the parsing phase. Their main limitation comes by relying on a treebank, which is an expensive resource.

Kawahara et al. (2000) use a full syntactic parser to obtain a case frame dictionary for Japanese, where arguments are distinguished by their syntactic case, including their headword (selectional restrictions). The resulting case frame components are selected by a frequency threshold.

Maragoudakis et al. (2001) apply a morphological analyzer and phrase chunking module to acquire subcategorization frames for Modern Greek. In contrast to this work, they use different machine learning techniques. They claim that Bayesian Belief Networks are the best learning technique.

Merlo and Leybold (2001) present learning experiments for automatic distinction of arguments and adjuncts, applied to the case of prepositional phrases attached to a verb. She uses decision trees tested on a set of 400 verb instances with a single PP, reaching an accuracy of 86,5% over a baseline of 74%.

Note that both Manning and Merlo and Leybold's systems learn from contexts with just one PP (maximum) per verb (finite state filter). Our system learns from contexts with up to 5 PPs. Furthermore, we distinguish 48 different kinds of cases, hence the number of combinations is considerably bigger.

Table 6. Summary of several systems on subcategorization information

Method	Number of frames	Number of verbs	Linguistic resources	F-Score (evaluation based on a gold standard)	Coverage on a corpus
C. Manning (1993)	19	200	POS tagger + simple finite state parser	58	
T. Briscoe & J. Carroll (1997)	161	14	Full parser	55	
A. Sarkar & D. Zeman (2000)	137	914	Annotated treebank	—	88
D. Kawahara et al. (2000)	—	23,497	Full parser		82 accuracy
M. Maragoudakis et al. (2001)	—	47	Simple phrase chunker	77	
This paper	—	640	Morph. Analyzer + Phrase Chunker + Finite State Parser	55	95

Regarding the parsing phase, the systems presented so far are heterogeneous. While Manning, Merlo and Leybold and Maragoudakis et al. use very simple parsing techniques, Briscoe and Carroll and Kawahara et al. use sophisticated parsers. Our system can be placed between these two approaches. The result of the shallow parsing is not simple in that it relies on a robust morphological analysis and disambiguation. Remember that Basque is an agglutinative language with strong morphology and, therefore, this stage is particularly relevant. Moreover, the finite state filter we used for parsing is very sophisticated (Karttunen et al. 1997, Aldezabal et al. 2001), compared to Manning's.

Conclusion

This work describes an initial effort to obtain subcategorization information for Basque. To successfully perform this task we had to go deeper than mere syntactic categories (NP, PP...) enriching the set of possible arguments to 48 different classes. This leads to quite sparse data. Together with sparseness, another problem common to every subcategorization acquisition system is that of noise, coming from adjuncts and incorrectly parsed elements. For that reason, we defined subcategorization acquisition in terms of distinguishing between arguments and adjuncts.

The system presented was applied to a newspaper corpus. Subcategorization acquisition is highly associated to semantics in that different senses of a verb will most of the times show different subcategorization information. Thus, the task of learning subcategorization information is influenced by the corpus. As for the evaluation of this

work, we carried out two different kinds of evaluation of the argument/adjunct distinction results. This way, we verified the relevance of semantics in this kind of task.

For the future, we plan to incorporate the information resulting from this work in our parsing system. We hope that this will lead to better results in parsing. Consequently, we would get better subcategorization information, in a bootstrapping cycle. We also plan to improve the results by using semantic information as proposed in A. Korhonen (2001).

Acknowledgements

This research has been supported by the European Commission (MEANING IST-2001-34460), the Spanish Ministry of Science and Technology (Hermes TIC2000-0335-C03-03), the University of the Basque Country (9/UPV00141.226-14601/2002) and the Basque Government (ETORTEK2002/HIZKING21).

References

Aduriz, I., J.M. Arriola, X. Artola, A. Díaz de Ilarraza, K. Gojenola and M. Maritxalar, 1997, «Morphosyntactic disambiguation for Basque based on the Constraint Grammar Formalism». Conference on Recent Advances in Natural Language Processing (RANLP).

Aldezabal, I., K. Gojenola and K. Sarasola, 2000, "A Bootstrapping Approach to Parser Development". *International Workshop on Parsing Technologies (IWPT),* Trento.

—, M. Aranzabe, A. Atutxa, K. Gojenola, M. Oronoz M. and Sarasola K., 2001, "Application of finite-state transducers to the acquisition of verb subcategorization information". *Finite State Methods in Natural Language Processing, ESSLLI Workshop,* Helsinki.

Alegria, A., X. Artola, K. Sarasola and M. Urkia, 1996, "Automatic morphological analysis of Basque", *Literary and Linguistic Computing.* 11.4, Oxford University.

Brent, M.R., 1993, "From Grammar to Lexicon: Unsupervised Learning of Lexical Syntax". *Computational Linguistics* 19, 243-262.

Briscoe, T. and J. Carroll, 1997, "Automatic Extraction of Subcategorization from Corpora", ANLP-97: 356-363.

Carroll, J., G. Minnen and T. Briscoe, 1998, "Can Subcategorization probabilities help a statistical parser?", *Proceedings of the 6th ACL/SIGDAT Workshop on Very Large Corpora,* Montreal.

Gawron, J.M., 1986, "Situations and prepositions", *Linguistics and Philosophy* 9.3, 327-382.

Grimshaw, J., 1990, *Argument Structure.* Cambridge, MA, MIT Press.

Grishman, R., C. Macleod, A. Meyers, 1994, *Complex Syntax: Building a Computational Lexicon.* COLING-94.

Karlsson, F., A. Voutilainen, J. Heikkila, A. Anttila, 1995, *Constraint Grammar: A language-independent system for parsing unrestricted text.* Mouton de Gruyter.

Karttunen, L., J.P. Chanod, G. Grefenstette, A. Schiller, 1997, "Regular expressions for language engineering", *Natural Language Engineering.*

Kawahara, D., N. Kaji and S. Kurohashi, 2000, "Japanese Case Structure analysis by unsupervised construction of a case frame dictionary", COLING-2000, Saarbrucken.

Korhonen, A., 2001, *Subcategorization acquisition.* Unpublished PhD dissertation, University of Cambridge.

Koskenniemi, K., 1983, *Two-level morphology: A general computational model for word-form recognition and production.* PhD dissertation, University of Helsinki.

Manning, C.D., 1993, "Automatic acquisition of a large subcategorization dictionary from corpora". *Proceedings of the 31th ACL.*

— and H. Schütze, 1999, *Foundations of statistical natural language processing.* The MIT Press, Cambridge, Massachusetts.

Maragoudakis, M., K. Kermanidis, N. Fakotakis and G. Kokkinakis, 2001, "Learning automatic acquisition of subcategorization frames using bayesian inference and support vector machines". *The 2001 IEEE International Conference on Data Mining, IMDC'01,* San José.

Merlo, P. and M. Leybold, 2001, "Automatic distinction of arguments and modifiers: the case of prepositional phrases". EACL-2001, Toulouse.

Pederssen, T., 1996, "Fishing for exactness". *Proceeding of the South-Central SAS User Group Conference (SCSUG-96).*

Pollard, C., and I. Sag, 1987, "An information based Syntax and Semantics", volume 13. CSLI lecture Notes, Standford University.

Sarkar, A., and D. Zeman, 2000, "Automatic extraction of subcategorization frames for Czech". COLING-2000, Saarbrucken.

Schutze, C., 1995, "PP Attachment and Argumenthood", *MIT Working Papers in Linguistics.*

Tenny, C., 1994, *Aspectual Roles and the Syntax-Semantic interface,* Kluwer Academic Publishers.

Verspoor, C., 1997, *Contextually-Dependent Lexical Semantics.* PhD dissertation, Brandeis University, MA.