Contribution of complex lexical information to solve syntactic ambiguity in Basque

Aitziber ATUTXA Eneko AGIRRE Kepa SARASOLA
Ixa Group, University of the Basque Country (UPV/EHU),
Manuel Lardizabal pasealekua, 1 · 20018 Donostia-San Sebastián
{aitziber.atucha,e.agirre,kepa.sarasola}@ehu.es

ABSTRACT
In this study, we explore the impact of complex lexical information to solve syntactic ambiguity, including verbal subcategorization in the form of verbal transitivity and verb-noun-case or verb-noun-case-auxiliary relations. The information was obtained from different sources, including a subcategorization dictionary extracted from a Basque corpus, the web as a corpus, an English corpus and a Basque dictionary. Functional ambiguity between subject and object is a widespread problem in Basque, where 22% of subjects and objects are ambiguous, and this ambiguity surfaces in 33% of the sentences. This problem is comparable to PP attachment ambiguities in other languages. Our results show that, using complex lexical information, our results are better than a state-of-the-art statistical parser, obtaining a statistically significant error reduction of 20%. The disambiguation system is independent on the actual parsing algorithm used. The analysis revealed that the most relevant information are the case carried by the noun and the transitivity of the verb.

TITLE AND ABSTRACT IN BASQUE

Informazio lexikal konplexuaren ekarpena euskarazko anbiguotasun sintaktikoen ebazpenean

Lan honetan informazio lexikal konplexua erabiltzearen garrantzia aztertzen dugu euskarazko anbiguotasun sintaktikoen ebazpenean. Aditzen iragankortasuna erakusten duen azpikategorizazioaren ekarpena aztertu dugu, baita aditz-izen-kasu eta aditz-izen-kasu-laguntzaile erlazioena ere. Informazio horiek hainbat iturritatik jaso ditugu: euskarazko corpus batetik, webetik berau corpus gisa hartuta, ingelesezko corpus batetik eta euskarazko hiztegi batetik. Subjektu eta objektuaren arteko anbiguotasun funtzionala maiz aurkitzen dugu euskarazko testuetan; subjektua edo objektua bereiztea kasuen %22an ambiguoa da, eta hori gertatzen da perpausen %33an. Horrela, arazo horren garrantzi handia konparagarria da beste hizkuntza batzuek duten PP attachment arazoarenarekin. Gure sistemaren emaitzak hobeak dira artearen egoerako analizatzaile sintaktiko estatistiko batenak baino, estatistikoki esanguratsua den %20ko errore-murrizketa lortzen baitu. Analisi sintaktikoa egiteko edozein algoritmorekin erabil daiteke desanbiguazio-sistema hau.

KEYWORDS: Syntactic ambiguity resolution, subcategorization, web as a corpus.

KEYWORDS IN BASQUE: Anbiguotasun sintaktikoaren ebazpena, azpikategorizazioa, amarauna corpus gisa.

Proceedings of COLING 2012: Technical Papers, pages 97–114,
COLING 2012, Mumbai, December 2012.

1 Introduction

Due to typological differences, ambiguities in some languages differ from ambiguities in other. For example, while prepositional phrase (PP) attachment ambiguity occurs in about 50% of the English sentences in the Penn treebank (Volk, 2006), it occurs in less than 0.1% of sentences in Basque Dependency Treebank (Aduriz et al., 2003). By contrast, the subject-object ambiguity does not pose any problem in English, but it does in Basque, where 22% of the subjects or objects are ambiguous and 33% of the sentences show this ambiguity. The pervasive nature of this ambiguity makes it worth specific analysis. This problem is relevant to processing similar ambiguities in other morphologically rich languages. For instance, Urdu and Hindi also display subject-object ambiguity due to erg-abs markings, as well as null-case markings (Dixon, 1994; Husain and Agrawal, 2012).

In the literature, we find several approaches to improve syntactic disambiguation. One of them involves focusing on solving a relevant ambiguity by using a problem specific classifier (Kübler et al., 2007; Anguiano and Candito, 2011). This allows to deeply understand the features involved in the ambiguity. The results obtained over the localized ambiguous relations are either used in a post-process by replacing those of a parser (a process also known as *parse correction*) or, after analyzing the most informative features, those are incorporated in the treebank and used to improve a statistical parser (Husain and Agrawal, 2012).

In this paper we will follow the first approach, targeting the resolution of the subject-object ambiguity in Basque, comparing our results to those obtained by the Malt parser (Nivre et al., 2007). In any case, our method is independent of the parser used, and could be incorporated to statistic or rule-based parsers (Aranzabe et al., 2004) as well.

The paper is structured as follows. We will first review the subject-object ambiguity in Basque. In Section 3 we review Basque Dependency Treebank and the methods to acquire verbal subcategorization information. Section 4 presents the features that are informative when resolving the ambiguity. In Section 5 the method to create the gold standard is presented, alongside our disambiguation method and the results. The related work is discussed in Section 6. Finally, the conclusions and future work are drawn.

2 Subject-Object ambiguity in Basque

Typologically, Basque is a highly inflectional head-final language (Ortiz de Urbina, 1989; Laka, 1996). It belongs to what has been called MoR-FWO languages, that is, morphologically rich, free word order languages (such as Czech, Turkish, Hindi etc). In most of the cases the relation between a head and its dependent gets realized through a morphological marker, neither bore by the head nor the dependent. This implies examination of elements occurring in non-local environments.

(1) [***Pertsona*** *nagusi gehien**ek***], *euren etxeetan bizitzen jarraitu nahi dute.*
[**People** aged most-**erg**], their homes live keep-on want auxiliary.
[Most of aged people] want to keep living in their own home.

Example 1 shows a noun phrase headed by *pertsona*, where the ergative marker (**ek**) is carried by the last element of the noun phrase, in this case *gehienek*.

Basque is a morphological ergative language (Dixon, 1994) as well in both case-marking and verbal auxiliary morphology. Thus, case-marking for subjects of intransitive verbs and objects of

$$absolutive = \begin{cases} subject\ \ of\ \ intransitive\ \ verbs \\ object\ \ of\ \ transtive\ \ verbs \end{cases}$$
$$ergative = subject\ \ of\ \ transitive\ \ verbs$$

Figure 1: Two cases in Basque, and their respective syntactic functions depending on the transitivity of verb.

transitives and their morphological cross-reference within the verbal agreement auxiliary are identical and different from subjects of transitive verbs. The case-marking of transitive subjects is the ergative case (-ak when singular,-ek when plural). The case-marking of intransitive subjects and objects of transitives is the absolutive case (-a when singular, -ak when plural). Figure 1 summarizes the functions for each of these case-markers. And the following examples illustrate those ambiguities.

(2) *Etiketatzaile**a*** *agertu* *da.*
Tagger-**abs-singular** showed up intransitive-auxiliary.
The tagger-**subj** showed up.

(3) *Etiketatzaileak* *erlazio**a*** *desanbiguatu* *du.*
Tagger-erg-singular relation-**abs-singular** disambiguate transitive-auxiliary.
The tagger-subj has disambiguated the relation-**obj**.

When an element is in the absolutive case, its function (subject or object) is ambiguous, that is, case-marking by itself does not tell whether the element is a subject or an object; it is the transitivity of the verb, which in finite sentences appear to be lexicalized in the auxiliary, along with the case marking which makes disambiguation possible. Examples 2 and 3 show that elements bearing the absolutive case can be either objects or subjects depending on the transitivity present in the auxiliary. In these examples, the ambiguity is resolved with the information made explicit by the auxiliary, but there are exceptions.

(4) *Erlazio**a*** *erortzean* *gertatu* *zen.*
Relation-**abs-singular** when-dropped happened transitive-auxiliary.
It happened when the relation-**subj** dropped.

(5) *Erlazio**a*** *ikustean* *gertatu* *zen.*
Relation-**abs-singular** when-seen happened transitive-auxiliary.
It happened when the relation-**obj** was seen.

In the case of infinitive verbs, which do not have auxiliaries and thus do not explicitly mark for transitivity, the syntactic function of absolutive noun phrases is ambiguous. Examples 4 and 5 show two sentences where *erlazioa* is either subject or object (respectively) depending on the subcategorization information of the verb (drop or see, respectively).

(6) *Sukaldari**ak*** *egin* *ditu.*
Cook-**erg-singular** make transitive-auxiliary.
The cook-**subj** made it.

(7) *Opil**ak*** *egin* *ditu.*
Cake-**abs-plural** make transitive-auxiliary.
(S)he made the cakes-**obj**.

(8) *Erlazio**ak*** *ikusita,* *ezin* *dut* *asmatu.*
Relation-**abs-plural** seeing, not transitive-auxiliary figure-out.
Seeing the relation-**obj**, I can't figure it out.

Another source of ambiguity arises from the ambiguous morphological marker **-ak**, which can mean absolutive plural or ergative singular. Basque is a 3-way pro-drop language (Ortiz de Urbina, 1989), and thus subjects and objects can be elided (note the difference with English, where there must be always a subject). This means that in Basque we can have sentences like 6 and 7 above, where the object (subject in example 7) does not surface. Note also that position does not help disambiguating subjects and objects, as Basque is a relatively free-word order language. These two examples are syntactically ambiguous, and can only be disambiguated using semantic information, that is, cooks tend to be the subjects of transitive verbs and cakes tend to be objects of transitive verbs.

Both sources of ambiguity can occur together. Example 8 shows an example where the verb is infinitive and there is an ambiguous marker -ak, making the dependent (*Erlazio-ak*) ambiguous between absolutive/ergative and subject/object interpretations.

These ambiguous instances are very common in Basque, making up to 22% of all objects and subjects in Basque Dependency Treebank (cf. Section 5.1).

3 Resources

This section starts by describing Basque Dependency Treebank (Aduriz et al., 2003) which we used to extract potentially ambiguous occurrences and evaluate our methods. Next, it presents the methods to learn subcategorization information for Basque verbs. We finally review Malt parser.

3.1 Basque Dependency Treebank

Basque Dependency Treebank has more than 150,000 tokens, distributed in 11,125 sentences coded in CONLL-X format (Aduriz et al., 2003). Each token (dependent) is represented by the following information: index, word form, lemma, syntactic category (part of speech) and subcategory, morphological features corresponding to the different markers attached to the lemma, index of the head and syntactic relation between the dependent and the head. Example 9 shows a sentence from the treebank. Note the first, second and fourth tokens (nouns (ize) and determiner (det) respectively), which show inesive (ine), ergative (erg) and absolutive (abs) markers in singular (sg) and plural (pl). The verb (adi) shows perfective aspect (buru) and participle (part) form, and the auxiliary (adl) appears in the past form (b1), agreeing with a 3rd person singular ergative (hark) and a 3rd person plural absolutive (haiek). The corresponding treebank file is shown in Figure 3.

index	wordform	lemma	category	subcategory	morp.	head	relation
1	Martxoan	martxo	ize	arr	ine sg	6	ncmod
2	Millarrek	Millar	izb	-	erg sg	5	ncsubj
3	gurpil	gurpil	ize	arr	-	5	ncobj
4	guztiak	guzti	det	oro	abs pl	3	detmod
5	puskatu	puskatu	adi	sin	part buru	0	root
6	zituen	*edun	adl	-	b1 hark haiek	5	auxmod
7	.	.	punt	-	_	6	punc

Figure 2: Treebank file in CONLL-X format corresponding to example 9. See text for explanations of tags.

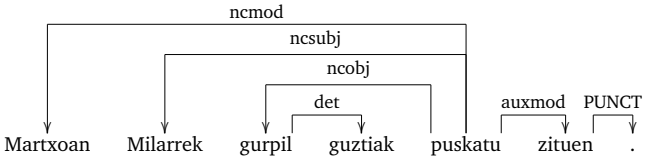

Figure 3: Dependency graph corresponding to example 9.

(9) *Martxoan* *Millarrek* *gurpil* *guztiak* *puskatu* *zituen.*
March-ine-sg Millar-erg-sg wheel all-abs-pl break auxiliary.
Millar broke all the wheels in March.

3.2 Acquisition of verbal subcategorization information

Verbal subcategorization information was extracted from 4 different sources: a subcategorization dictionary built from monolingual Basque corpus, web queries, monolingual English corpus, and a traditional monolingual Basque dictionary.

The subcategorization dictionary was obtained from Basque monolingual corpus, initially built with the purpose of making attachment decisions for a shallow parser on its way to full parsing (Aldezabal et al., 2002). For each of the 2,571 verbs this dictionary lists information about transitivity of the verb, noun[1]-case-verb triples or noun-case-verb-auxiliary quadruples and estimated frequency of each.

This dictionary was automatically built from raw corpora, comprising a compilation of 18 months of news from *Euskaldunon Egunkaria* (a journal written in Basque). The size of the corpus is around 780,000 sentences, approximately 10M words. From the 5,572 different verb lemmas in the corpus, the subcategorization dictionary was compiled for the 2,751 verbs occurring at least 10 times. The corpus was parsed by a chunker (Aduriz et al., 2004) which includes both named-entity and multiword recognition. The chunker uses a small grammar to identify heads, postpositions and verb attachments of NPs and PPs. The grammar was developed

[1]To simplify we just mention nouns, but there are also adjectives, determiner etc, anything that could be the head of a DP or CP.

based on the fact that Basque is a head-final language and it includes a distance feature as well. Phrases were correctly attached to the verb with a precision of 78%. Note that the auxiliary verb in Basque allows to unambiguously determine the transitivity of the main verb. The information captured for each verb corresponds to noun-case-verb triples and the noun-case-verb-auxiliary quadruples.

The second source is an English monolingual corpus. The assumption here is that the subject-object relation is stable when translating across languages, that is, if an element-verb relation is labeled as a subject relation in English it will also be that way in its Basque translation. This is a strong assumption, and we expect it to work better with certain verbs (e.g. activity or achievement verbs) than others (e.g. static verbs), but we did not make any distinction so far. In this approach, for each Basque ambiguous element-verb pair we collected frequencies over unambiguous English examples acquired from automatically parsed English data. We used the BNC corpus parsed with the RASP parser (Briscoe et al., 2006) containing 47,145,584 syntactic relations, where 10,447,129 are verb-noun dependency relations. The method has the following steps:

1. Translate the dependent lemma and the verb lemma using a bilingual dictionary.
2. Build all possible translation pairs.
3. Collect frequencies of each pair in the English corpus, depending on the label (subject or object) assigned by the English parser.

The third source is the result of directly querying the web. The web can be seen as a vast corpus (Bansal and Klein, 2011; Nakov, 2007; Lapata and Keller, 2005) where, in principle, we have bigger chances of finding low frequency combinations not found with the monolingual or crosslingual approaches presented so far. The following steps were pursued:

1. Obtain the lemma of the ambiguous element, create all possible subject and object unambiguous inflected forms using a language generation tool, that is, lemma+ergative+plural and lemma+absolutive+singular pairs.
2. Obtain the three different inflected forms of the verb (verb+future aspect marking, verb+perfect aspect marking and verb+ imperfect aspect marking) using the same generation tool.
3. Generate the corresponding different transitive and intransitive auxiliaries as well. It is not feasible to use all possible transitive/intransitive auxiliaries because the query number would explode so we took into account the 20 most frequent forms.
4. Construct all possible element+case+verb+auxiliary quadruples. For each element-verb candidate we get approximately 60 quadruples,
5. Search Google and collect hits.

Unfortunately Google does not recognize documents in Basque as a separate language. Some authors (Leturia et al., 2008) add certain common Basque words to the query, in order to reduces the number of texts in other languages returned by Google. We solved the problem using another heuristic, restricting to documents not in Spanish nor in English. The first one because Basque borrows vocabulary from Spanish and several times Basque texts are wrongly tagged as Spanish texts, and the other because of the same reason plus the fact that is the most common language on the web. Variability on the web does not cause a problem in this case because all the searches concerning each element-verb candidate are performed at the same time.

The last source is a traditional dictionary (EH dictionary, (Sarasola, 1996)), where each verbal entry carries information on the transitivity of the verb. As each verb usually shows more than one sense, we just considered the first sense. This dictionary uses 7 different markers to capture transitivity: du/da, du, du/dio, dio, da, zaio, da/zaio. For instance, du/da represent verbs that can appear in transitive or intransitive contexts, such as inchoative verbs like *break* that show a transitive/inchoative alternation: Leioa-subj apurtu da (The window-subj broke) or Mikelek-subj lehioa-obj apurtu zuen (Mikel-subj broke the window-obj).

4 Feature Space

In this section we will try to collect the information that we deemed was relevant to disambiguate subjects and objects. Each piece of information configures a separate feature. Our feature space F encodes heterogeneous information representing each candidate element-verb.

The features are presented grouped into sets depending on the nature and source of each one. A value close to 1 means that there is evidence for disambiguating to subject. A value close to 0 means that the feature leans for object. In some cases the feature does not predict anything.

Features related to subcategorization information

These features are based on the information mined from each source of subcategorization information, as presented in the previous section.

Subcategorization dictionary (SubcatDict)

- *TransCase(SubcatDict)*: The probability of the element to be a subject depends on the probability of the verb to be transitive $P(TransCase)$ and the actual case marking assigned by the morphological analyzer to the ambiguous element. $P(TransCase)$ is estimated from the SubcatDict, counting the occurrences of the verb as transitive and intransitive. If the case of the element is ergative and $P(TransCase)$ is bigger than 0.5, then this feature takes the value of $P(TransCase)$, that is, the feature will lean towards a subject interpretation. If the case is the absolutive, then the feature will lean towards being a subject if the verb is intransitive, or to object otherwise. The value of the feature is encoded according to the following formula.

$$\textit{TransCase(SubcatDict)} \begin{cases} P(TransCase) = \frac{\#trans}{\#trans+\#intrans} & case = erg \,\&\, P(TransCase) > 0.5 \\ 1 - P(TransCase) & case = abs \,\&\, P(TransCase) < 0.5 \\ 0 & case = abs \,\&\, P(TransCase) > 0.5 \\ none & otherwise \end{cases}$$

- *NCaseV(SubcatDict)*: The probability of the element to be a subject is related to the tendency of the element to bear the ergative case with that verb in the corpus (*P(Erg)*), independently of whether the verb shows transitive tendency. Here we do not consider the actual case marker assigned by the morphological analyzer, but the tendency of the element to bear ergative case when occurring with the verb.

$$\textit{NCaseV(SubcatDict)} \begin{cases} 1 & P(TransCase) > 0.5 \,\&\, P(Erg) > 0.5 \\ 0 & P(TransCase) < 0.5 \,\&\, P(Erg) < 0.5 \\ none & otherwise \end{cases}$$

- *NCaseVAux(SubcatDict)*: The probability of the element to be a subject is related to the tendency of the element to appear in a subject configuration, that is to say, to bear the ergative case with the verb appearing with a transitive auxiliary, or to bear absolutive case with the verb appearing with an intransitive auxiliary. This is estimated as described in the following formula.

$$NCaseVAux(SubcatDict) \begin{cases} \frac{\#(n+abs+v+intransAux)+\#(n+erg+v+transAux)}{\#(n+case+v)} & \#(n+case+v)>0 \\ none & otherwise \end{cases}$$

Web as a corpus (Web)

- Similar to the features for the SubcatDict, we can estimate the same features using the web queries explained in the previous section. This yields three new features: *TransCase(Web)*, *NCaseV(Web)* and *NCaseVAux(Web)*.

English corpus (BNC)

- *Subj(BNC)*: The value is 1 if the element shows a tendency of being a subject in the English corpus, 0 if the tendency is to object and none if it could not be translated or if it was not found in the English corpus.

Dictionary (Dict)

- *TransCase(Dict)*: This feature corresponds to the combination of two informations. The value is 1 if according to the dictionary the verb is transitive and the case is ergative, or if the verb is intransitive and the case is absolutive. The value is 0 if the verb is transitive and the case is absolutive, and none if transitivity could not be established in the dictionary.

Features related to morphological and syntactic information

Here we group weaker indications, as follows.

- *AspectCtrl*: The value is 1 if the verb is an aspectual or control verb such as *begin, end, stop, quit*. Aspectual verbs that occur a verb in the infinitive, force the verb in the infinitive to miss the surface subject. For example, *I started knowing you*, there cannot be a surface subject for *know* because *start* is an aspectual verb and thus both verbs share the same subject. Control verbs like *want* or *expect* act in Basque quite similarly to aspectual verbs In short, when the head of the verb is an aspectual or control verb, the element would tend to be the object, and thus feature will be 0, and none otherwise.
- *Preverb*: the value is 1 if the element appears in the pre-verbal position, as this is a sign of being a subject.
- *Inf*: 1 if the verb appears to be an infinitival, bare-infinitival, infinitival with a relative marker, infinitival nominalization, or in a finite form, that is to say, with an auxiliary, and finite with relative marker, respectively.
- *Erg*: The value is 1 if the case born in the noun phrase where the element is located is ergative, 0 otherwise. Remember from 2, that the head of a noun phrase does not necessarily bear the case marking, which is found in the last constituent of the noun phrase . We applied feature propagation in order to recover the case.
- *-ak*: The value is 1 if the element bears the ambiguous -ak morpheme.

- *Sing*: The value is 1 if the number of the element is singular, 0 otherwise.
- *Entity*: The value is 1 if the element starts with a capital letter and is not in the first position in the sentence. Since the morphological analyzer does not implement any entity recognition, this is an heuristic to code possible entities.

5 Experimental setup

In this section we review the method to create the gold standard, followed by our disambiguation method, the method to train malt parser, and the results of the experiments.

5.1 Creating the gold standard

The gold standard comprises ambiguous instances of noun phrases, which can either be subject or object. The syntactic structure and gold label is taken from Basque Dependency Treebank. In order to make the setting realistic the morphological tags are taken from an automatic morphological disambiguation tagger for Basque (Aduriz et al., 2000). Note that this tagger is conservative and does not always return a single tag. In those cases we use the first tag, as customary with most parsers. The accuracy of this tagger when selecting ergative and absolutive is 87%.

The procedure to detect ambiguous instances is the following. We first look up the verbs in the corpus. Depending on the finiteness of the verb, we have two cases:

1. If the verb has an auxiliary annotated as finite by the morphological analyzer, then the agreement features in the auxiliary verb resolve all ambiguities, except in some cases with -ak, which can be morphologically ambiguous between ergative singular or absolutive plural, and thus ambiguous between subject and object (respectively). To be more specific, if the auxiliary verb shows agreement with a singular ergative and a plural absolutive (both occurring with -ak) then both dependents could be either subject or object.
2. If the verb is tagged as infinitive, then if the verb has a dependent with the absolutive case, the dependent is ambiguous between subject and object.

The dependents of the verbs are looked up in the treebank, extracting the head and the case marking. The head is given as the root of the dependent, and the case marking of the dependent is given by the last token under the dependent.

Detecting dependents is a difficult task for parsers. So we filter out those dependents which could be difficult for current parsers, as follows. If the sentence contains a single verb, we then check all dependents of the verb. If the verb has multiple verbs, we then need a clause delimiter to identify which phrases are dependents of which verbs. We use a simple heuristic based on the fact that Basque is head-final: all words before the first verb are assigned to that verb, and the rest are assigned to the second verb. This is a strong baseline for any parser, as it attains 86% accuracy in a study of our own.

The above method to extract ambiguous dependents yields 4,525 ambiguous dependents in 3,617 sentences, that is, 22% of all subjects and objects in the treebank, with one ambiguous instance every 33% of sentences.

Note that if we had searched for ambiguous instances using the gold morphological tags in the treebank, we would find 4,400 subject-object ambiguous relations. This smaller amount is

due to errors by the morphological analyzer, as it incorrectly tags some ergative or absolutive cases, or certain auxiliaries as main verbs, considering the sentence as an infinitive sentence and therefore ambiguous with respect to subject-object. Note that the real ambiguity faced by a parser is closer to that of the automatic analysis.

5.2 Methods

We performed a machine learning experiment to examine the impact of the use of those lexical features to solve subject-object ambiguity in Basque. We used Support Vector Machines (Chang and Lin, 2011) with Radial kernels tuning C and G parameters over the training set using cross-validation to find their best values. The 4,525 ambiguous subject-object relations in the treebank were split into training and testing sets in a proportion of 50%. We also evaluated each feature on its own, and also, used an ablation procedure, learning the classifier with all features but one.

5.3 Malt parser

We chose Malt parser (Nivre et al., 2007) to compare our results with. This parser is a history-based deterministic dependency parser, which using the input and a stack and through 4 main actions, *shift* moving a token from the input to the stack, *left-arc* adding an arc from the token on the input to the token on the top of the stack, removing the token from stack, *right-arc* adding an arc from the token on the top of the stack to the token on the input, moving the token of the input onto top of stack and *reduce* removing a token from the top of the stack. The action is chosen according to a machine learning classifier using a variety of features including the stack, the input and past history. This way, dependency tree gets built in one single pass and in linear time. We used version 1.4, and the configuration was selected using the optimization developed by Nivre and Ballesteros Ballesteros and Nivre (2012).

Maltparser has to face a wider range of ambiguity that our classifier, as in some cases a phrase with the absolutive case can play syntactic functions other than subject or object. In order to make comparison to our classifier fair, we substituted all other tags returned by Malt parser with *ncobj*, the most common tag (around 75% compared to 25% for ncsubj).

5.4 Results

Table 1 shows the results of some features evaluated independently over the training set. For the sake of brevity, we only show those features with an accuracy over 60% are displayed. The baseline consists of assigning always the object tag to any ambiguous element, since it is the most frequent tag (75%). We measure accuracy on finding both subjects and objects, but we are also interested in measuring the performance of the system for detecting subjects since this is the most difficult task. Accuracy is the number of correctly tagged elements over all elements. Precision on subjects is the number of correctly recognized subjects divided by all elements tagged as subject by the system. Recall on subjects is the number of correctly recognized subjects divided by the total number of subjects in the gold set. F1 is the harmonic mean between precision and 1.

The table shows that the *Erg* feature performs best. This feature relies in the tag assigned by the automatic morphological analyzer. TransCase(SubcatDict) provides the second best result in terms of accuracy, still over the baseline, beating all the others in F1 and recall for subjects.

Feature	acc (sbj+obj)	prec (sbj)	rec (sbj)	F1 (sbj)
Baseline	75.29	00.00	00.00	00.00
TransCase(SubcatDic)	76.99	82.58	**74.17**	**78.15**
NCaseV(SubcatDic)	72.21	51.50	48.33	49.86
TransCase(Web)	60.10	80.94	57.47	67.21
NCaseV(Web)	69.21	22.71	19.16	20.78
TransCase(Dict)	60.31	**83.63**	50.26	62.79
Preverbal	62.09	17.93	17.93	17.93
Erg	**86.06**	50.26	50.26	50.26

Table 1: Results on ambiguous elements in the training corpus for each feature evaluated independently. Note that we only with accuracies over 0.6.

Feature	acc	prec(sbj)	rec(sbj)	F1(sbj)
Baseline	75.29	00.00	00.00	00.00
All features	**89.62**	**86.34**	**68.89**	**76.63**
¬SubcatDict	88.23	84.98	63.62	72.76
¬Web	88.32	83.94	65.20	73.39
¬BNC	88.23	84.49	64.14	72.93
¬Dict	87.66	86.25	59.57	70.47
¬SubcatInf	86.06	88.27	**50.26**	70.47
¬CaseNum	**85.28**	**77.64**	56.77	**65.58**
¬NCaseV(Aux)*	87.84	83.84	62.91	71.88

Table 2: Results using 10-fold cross-validation on ambiguous elements in the training corpus. The first lines correspond to the baseline and full classifier, and the rest present the results of feature ablation experiments. Best results and worst results in each column in bold.

Note that the performance of some features suffers from the fact that they only apply to a subset of the elements. In fact, *Erg* is the only one which applies to all.

Table 2 shows the results when training an SVM with Radial kernel and 10 fold cross-validation (All features line). Those results beat the baseline and individual features by a large margin. When compared with the best individual feature (*Erg*) the difference in accuracy is smaller, but note that the classifier is better on detecting subjects, showing that its output is better balanced.

The table also shows the results of feature ablation. Features were grouped by their source (SubcatDict, Web, BNC, Dict) or by the linguistic nature of the information they carry, independently of the source. For example, SubcatInf in Table 2 represents all subcategorization information: TransCase(SubcatDic), NCaseV(SubcatDic), NCaseVAux(SubcatDic),TransCase(Web), NcaseV(Web), NCaseVAux(Web), Subj(BNC) and TransCase(Dict). The highest loss in overall accuracy is for CaseNum features, but all features cause a performance drop when removed. This shows that the features are complementary. The loss in F1(sbj), whenever ablation of any kind is carried out, confirms this fact. The highest loss in precision and F1 also occur when information about case and number are left aside. And the highest loss in recall corresponds to

	acc	prec(sbj)	rec(sbj)	F1(sbj)
All features	89.33	82.48	71.74	76.74
MALT	86.72	76.82	65.69	70.82

Table 3: Final results for ambiguous elements on the test set.

	LAS	UAS		prec	rec	F1
MALT	83.17	83.08	sbj	71.57	75.01	73.24
			obj	76.36	73.61	74.95
MALT Post-processed	83.52	83.08	sbj	72.11	75.52	73.77
			obj	81.10	74.39	77.60

Table 4: Final results over all dependencies in test set.

the elimination of subcategorization information.

Finally we run our classifier with all features in the test set, as shown in Table 3. The performance obtained is comparable to that in the train set using cross-validation. The results of Malt are lower both in accuracy and F1 over subjects, with a statistical significant difference (p-value < 0.005). Note that the error reduction is 19.64%.

The above results show that our approach is competitive over Malt for the subset of ambiguous subject-object relations. In order to show that our system can make a relevant difference over the overall performance of a parser, we corrected the output of Malt parser with the result of our classifier and evaluated over all dependencies. Table 4 shows the usual UAS (Unlabeled Attachment Score) and LAS (Labeled Accuracy Score) scores for both MALT and the post-processed MALT. Of course, the unlabeled score is the same for both, as we just changed some labels. The improvement in LAS is of 0.35 absolute points, statistically significant (p-value < 0.00009). In addition we also show the performance over objects and subjects. Note that the post-processed version improves both object and subject recognition.

6 Related Work and Discussion

As mentioned in the introduction, this work is framed following a parse correction strategy. In parsing, not all ambiguities show the same complexity, and not all the languages behave the same way with respect to the distribution of the ambiguities. In English, for example, prepositional phrase attachment (PP-attachment for short) ambiguity traditionally stirred interest since (Hindle and Rooth, 1993; Ratnaparkhi, 1998), among others, for being both common and difficult to solve. These seminal works presented PP-attachment resolution in isolation, with no evaluation over full sentences or integration with a parser or with the results of a parser. With the proliferation of statistical parsers the attention moved from solving specific ambiguities to treat ambiguities as a whole. Statistical parsers learned from treebanks, though, make it difficult to reach any conclusion on what is the relevant information for resolving specific ambiguities, and whether those need to be encoded explicitly in treebanks.

Focusing on a relevant ambiguity is helpful to achieve a better understanding of the intricacies of parsing structures. Along these lines, we find two main approaches, that of parsing correction, or that of transforming and enriching the treebank with additional information. Work in parse correction consists on the creation of corrective models to solve difficult ambiguities, such as PP-attachment ambiguity in English. Thus correction occurs as a post-process to parsing,

replacing the output of the parser (labels or attachments) with alternatives obtained in an independent classification process.

The literature differs on the languages and the criteria used to choose the target information to be corrected. Hall and Novak Hall and Novák (2005) worked on Czech. They highlight the problem of projectivity as particularly problematic when parsing free word-order languages, such as Czech, due to the frequency of sentences with non-projective constructions. They present a corrective model which recovers non-projective dependency structures by training a classifier to select correct dependency pairs from a set of candidates obtained from parses generated by an automatic parser. In this case, the baseline parsers were Collins (2000) and Charniak (Charniak and Johnson, 2005), and the authors showed an improvement on dependency accuracy from 81.6% to 82.8% and from 84.4% to 85.1% (respectively).

Kilian and Menzel Foth and Menzel (2006) developed a classifier to solve PP-attachment ambiguity in German, and integrating the classifier into a rule-based parser. They report an error reduction of 14% and an improvement of overall attachment accuracy from 89.3% to 90.6%. Kluber, Ivanova and Klett Kübler et al. (2007) dealt with English PP-attachment ambiguity. They used the MaltParser (Nivre et al., 2007), showing an overall labeled parser accuracy improvement from 86.2% to 86.5%, and more precisely an improvement in unlabeled attachment from 71.8% to 77.4% in the case of prepositions. Henestroza and Candito (Anguiano and Candito, 2011) focus on PP-attachment and coordination as being difficult attachment ambiguities for French, presenting an improvement from 89.78% to 90.47% over MaltParser, and an improvement from 91.04% to 91.36% over MSTParser.

Attardi and Ciaramita Attardi and Ciaramita (2007) worked on English and Swedish. They show a slightly different methodology in that they do not focus on a particular ambiguity. They collect parsing errors by comparing correct parse trees with incorrect trees produced by the base parser on a training corpus. From those, they define a learning task where the input is a set of features extracted at each node in the parse trees produced by the parser on the training corpus whose output is a set of revised decisions, allowing correction. Their analysis was based on DeSR (Attardi, 2006), a shift-reduce parser.

Hussain and Agrawal Husain and Agrawal (2012) make an exhaustive analysis of the parsing errors committed by MaltParser over the Hindi Dependency Treebank. Hindi, as Basque, is a MoR-FWO language, and ergative as well. They identified that 50% of the errors were related to verb argument structure. The relevant information to avoid the errors was either contained in the treebank in a way that was difficult to manage by the parser, or had to be added to the Treebank, enriching it. They pursued two different experiments on tree-transformation and enrichment of the treebank using linguistic information from a dictionary and VerbNet. They conclude that only tree transformations lead to improvements, while enriching the treebank seems to have no effect whatsoever. In related work, (Husain et al., 2010) analyze the importance of different linguistic features over two dependency parsers on Hindi. They conclude that case marking and several verbal features such as tense, aspect and modality are the most informative. This is in contrast with our findings, where case is important, but also transitivity and subcategorization models.

In a different tone, Atterer and Schütze Atterer and Schütze (2007) are critic on some parse correction experiments. They deem parse correction as being unrealistic because it relies on using the treebank as an oracle to select the ambiguous candidates to be evaluated. They argue that, in contrast to the use of gold morphological analysis from the treebank to select ambiguity cases, parsers do not have access to those gold annotations at parsing time. They also argue that,

due to wrong attachment decisions, the parser might miss some ambiguous head-dependent relations that were mined from gold standard treebanks. To avoid these inconveniences when selecting the ambiguous candidates in our work, we used the morphological tags assigned by a morphological analyzer. Furthermore, we used a positional heuristic for assigning dependents to verbs, leaving aside the elements that could be potentially problematic for a parser as explained in 5.1. This way, we find 2303 ambiguous candidates in the training set, compared to 2338 if we would use the output of Malt parser. The intersection between the two amounts to 2040. In order not to penalize MaltParser on those non-coincident candidates, we assigned the object relation to those cases.

An important argument in favor of parse correction is that it is parser-agnostic, that is, our classifier is an independent module which disambiguates certain difficult ambiguities, and its results can be replaced over the result of any parser. In this paper we report on the correction over MaltParser for Basque(Bengoetxea et al., 2011), but in the future we plan to use our system to correct the output of a rule-based parser for Basque (Aranzabe et al., 2004).

7 Conclusions and Future Work

This works confirms the relevance of complex lexical information when solving syntactic ambiguity. More specifically, we have focused on subject-object ambiguity in Basque where it is one the main ambiguities, as it surfaces in 33% of the sentences. This problem is relevant to processing similar ambiguities in other morphologically rich languages. For instance, Urdu and Hindi also display subject-object ambiguity due to erg-abs markings, as well as null-case markings.

We have explored the impact of complex lexical information, including verbal subcategorization in the form of verbal transitivity and verb-noun-case or verb-noun-case-auxiliary relations. In addition, we have studied several linguistically motivated features, and trained a supervised classifier. Our results show that all sources of lexical information and features employed contribute positively, proving their complementarity. In fact, our classifier is better than a state-of-the-art statistical parser trained for Basque, obtaining a statistically significant error reduction of 20% in the resolution of subject-object ambiguities. When used to correct the output of the parser, the improvement is small, albeit statistically significant, showing that the classifier impacts in the overall parser performance.

The analysis revealed that the most relevant pieces of information are the case carried by the dependent element and the transitivity of the verb. For the future we would like to study the similarities and differences with typologically related languages. In addition, we plan to incorporate some of the features into the treebank and statistical parsers.

Acknowledgements

This research is partially funded by the Ministry of Economy under grants TIN2009-14715-C04-01 (KNOW2 project) and TIN2009-14675-C03-01 (OPENMT-2 project).

References

Aduriz, I., Agirre, E., Aldezabal, I., Alegria, I., Arregi, X., Arriola, J. M., Artola, X., Gojenola, K., Maritxalar, A., Sarasola, K., and Urkia, M. (2000). A word-grammar based morphological analyzer for agglutinative languages. In *Proceedings of the 18th conference on Computational linguistics - Volume 1*, COLING '00, pages 1–7, Stroudsburg, PA, USA. Association for Computational Linguistics.

Aduriz, I., Aranzabe, M., Arriola, J., Atutxa, A., de Ilarraza, A. D., Garmendia, A., and Oronoz, M. (2003). Construction of a basque dependency treebank. In Nivre, J. and Hinrich, E., editors,

Proceedings of the Second Workshop on Treebanks and Linguistic Theories (TLT 2003), pages 201–204.

Aduriz, I., Aranzabe, M. J., Arriola, J. M., de Ilarraza Sánchez, A. D., Galletebeitia, K. G., Oronoz, M., and Uria, L. (2004). A cascaded syntactic analyser for basque. In Gelbukh, A. F., editor, *CICLing*, volume 2945 of *Lecture Notes in Computer Science*, pages 124–134. Springer.

Aldezabal, I., Aranzabe, M., Gojenola, K., Sarasola, K., and Atutxa, A. (2002). Learning argument/adjunct distinction for basque. In *Proceedings of the ACL-02 workshop on Unsupervised lexical acquisition - Volume 9*, ULA '02, pages 42–50, Stroudsburg, PA, USA. Association for Computational Linguistics.

Anguiano, E. H. and Candito, M. (2011). Parse correction with specialized models for difficult attachment types. In *EMNLP*, pages 1222–1233. ACL.

Aranzabe, M., Arriola, M., and de Ilarraza, D. (2004). Towards a dependency parser of basque. In *Proceedings of the Workshop on Recent Advances in Dependency Grammar, COLING'04 Workshop*, pages 49–56.

Attardi, G. (2006). Experiments with a multilanguage non-projective dependency parser. In *Proceedings of the Tenth Conference on Computational Natural Language Learning*, CoNLL-X '06, pages 166–170, Stroudsburg, PA, USA. Association for Computational Linguistics.

Attardi, G. and Ciaramita, M. (2007). Tree Revision Learning for Dependency Parsing. In Sidner, C. L., Schultz, T., Stone, M., and Zhai, C., editors, *HLT-NAACL*, pages 388–395. The Association for Computational Linguistics.

Atterer, M. and Schütze, H. (2007). Prepositional phrase attachment without oracles. *Computational Linguistics*, 33(4):469–476.

Ballesteros, M. and Nivre, J. (2012). Maltoptimizer: A system for maltparser optimization. In Chair), N. C. C., Choukri, K., Declerck, T., Doğan, M. U., Maegaard, B., Mariani, J., Odijk, J., and Piperidis, S., editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

Bansal, M. and Klein, D. (2011). Web-scale features for full-scale parsing. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 693–702, Stroudsburg, PA, USA. Association for Computational Linguistics.

Bengoetxea, K., Casillas, A., and Gojenola, K. (2011). Testing the effect of morphological disambiguation in dependency parsing of basque. Dublin - Irlanda.

Briscoe, T., Carroll, J., and Watson, R. (2006). The second release of the rasp system. In *Proceedings of the COLING/ACL on Interactive presentation sessions*, COLING-ACL '06, pages 77–80, Stroudsburg, PA, USA. Association for Computational Linguistics.

Chang, C. and Lin, C. (2011). LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):27.

Charniak, E. and Johnson, M. (2005). Coarse-to-fine n-best parsing and maxent discriminative reranking. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, pages 173–180, Stroudsburg, PA, USA. Association for Computational Linguistics.

Collins, M. (2000). Discriminative reranking for natural language parsing. In Langley, P., editor, *ICML*, pages 175–182. Morgan Kaufmann.

Dixon, R. (1994). *Ergativity*. Cambridge Studies in Linguistics. Cambridge University Press.

Foth, K. A. and Menzel, W. (2006). The benefit of stochastic pp attachment to a rule-based parser. In *Proceedings of the COLING/ACL on Main conference poster sessions*, COLING-ACL '06, pages 223–230, Stroudsburg, PA, USA. Association for Computational Linguistics.

Hall, K. and Novák, V. (2005). Corrective modeling for non-projective dependency parsing. In *Proceedings of the Ninth International Workshop on Parsing Technology*, Parsing '05, pages 42–52, Stroudsburg, PA, USA. Association for Computational Linguistics.

Hindle, D. and Rooth, M. (1993). Structural ambiguity and lexical relations. *Computational Linguistics*, 19:103–120.

Husain, S. and Agrawal, B. (2012). Analyzing parser errors to improve parsing accuracy and to inform tree banking decisions. *Linguistic Issues in Language Technology*, 7(1).

Husain, S., Mannem, P., Ambati, B., and Gadde, P. (2010). The icon-2010 tools contest on indian language dependency parsing. *Proceedings of ICON-2010 Tools Contest on Indian Language Dependency Parsing, ICON*, 10:1–8.

Kübler, S., Ivanova, S., and Klett, E. (2007). Combining dependency parsing with pp attachment. In *Fourth Midwest Computational Linguistics Colloquium*. Citeseer.

Laka, I. (1996). *A brief grammar of Euskara, the Basque language*. Euskal Herriko Unibertsitatea, Bilbao.

Lapata, M. and Keller, F. (2005). Web-based models for natural language processing. *ACM Trans. Speech Lang. Process.*, 2(1).

Leturia, I., Gurrutxaga, A., Areta, N., and Pociello, E. (2008). Analysis and performance of morphological query expansion and language-filtering words on basque web searching. In Chair), N. C. C., Choukri, K., Maegaard, B., Mariani, J., Odjik, J., Piperidis, S., and Tapias, D., editors, *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2008/.

Nakov, P. I. (2007). *Using the Web as an Implicit Training Set: Application to Noun Compound Syntax and Semantics*. PhD thesis, EECS Department, University of California, Berkeley.

Nivre, J., Hall, J., Nilsson, J., Chanev, A., Eryigit, G., Kubler, S., Marinov, S., and Marsi, E. (2007). Maltparser: A language-independent system for data-driven dependency parsing. *Natural Language Engineering*, 13(2):95.

Ortiz de Urbina, J. (1989). *Parameters in the Grammar of Basque*. Dordrecht, Foris (Studies in Generative Grammar, 33).

Ratnaparkhi, A. (1998). Statistical models for unsupervised prepositional phrase attachement. In *COLING-ACL*, pages 1079–1085.

Sarasola, I. (1996). *Euskal Hiztegia*. Kutxa Fundazioa, Donostia.

Volk, M. (2006). How bad is the problem of pp-attachment?: a comparison of english, german and swedish. In *Proceedings of the Third ACL-SIGSEM Workshop on Prepositions*, pages 81–88, Stroudsburg, PA, USA. Association for Computational Linguistics.