Pronominal anaphora in Basque: annotation of a real corpus

Aduriz, Itziar
Universitat de Barcelona
Gran Vía de les Corts
Catalanes, 585
08007 Barcelona
jiradagi@si.ehu.es

Ceberio, Klara
Euskal Herriko Unibertsitatea/
Universidad del País Vasco
Manuel Lardizabal 1
20018 Donostia
jibcebec@sc.ehu.es

Díaz de Ilarraza, Arantza
Euskal Herriko Unibertsitatea/
Universidad del País Vasco
Manuel Lardizabal 1
20018 Donostia
jipdisaa@si.ehu.es

Abstract: This paper describes the process followed in the annotation of pronominal anaphora in the Eus3LB corpus[1] of Basque. Our aim is to use this annotation as the basis for later computational treatment of our language. We present the linguistic analysis carried out, the criteria defined for the tagging and some relevant linguistic conclusions about the features of the antecedents needed to link them correctly to their anaphoric elements.

Keywords: Pronominal anaphora, anaphoric tagging of the corpus.

Resumen: En este artículo se describe el proceso de etiquetado manual de la anáfora pronominal en el corpus Eus3LB, corpus de 54.000 palabras de texto escrito en euskera etiquetado a nivel sintáctico y que servirá de base para posteriores tratamientos computacionales. Presentamos aquí el estudio lingüístico previo, los criterios de etiquetado establecidos y algunas conclusiones lingüísticas relevantes sobre las características de las relaciones entre la anáfora pronominal y su correspondiente antecedente.

Palabras clave: anáfora pronominal, etiquetado anafórico del corpus.

1 Introduction

Anaphora resolution is a wide-open research field in the area of Natural Language Processing (NLP) and it is crucial for the task of understanding the language at discourse level. Anaphora resolution, like other types of language's automatic treatment, needs corpus annotation. Mitkov (2002) highlights the importance of an annotated corpus for research purposes:

"The annotation of corpora is an indispensable, albeit time-consuming, preliminary to anaphora resolution (and to most NLP tasks or applications), since the data they provide are critical to the development, optimisation and evaluation of new approaches".

Taking this statement into account, we began annotating Eus3LB corpus with anaphoric information.

This paper describes the way followed in this tagging. First, we began determining the main subject of our study, the pronominal anaphora in Basque. The fact that in this language there is not a specific theoretical work in this area made the previous study larger than we expected.

[1] The Eus3LB corpus is a part of the corpus of the 3LB project (Palomar et al., 2004).

In section 2 we make a brief presentation of the anaphora from a linguistic point of view. Section 3 is dedicated to explain the criteria we defined for the annotation of the Basque pronominal anaphora, the demonstratives *hau* (this), *hori* (that) and *hura* (that/ he/ she/ it) in the above-mentioned corpus. Finally, in section 4 some conclusions and future work are presented.

2 The anaphora in a linguistic context. Pronominal anaphora in Basque

According to Hirst (1981): "*anaphora, in discourse, is a device for making an abbreviated reference (containing fewer bits of disambiguating information, rather than being lexically or phonetically shorter) to some entity (or entities)*".

This reference can appear before the anaphoric element (*Lisa* could see the stars in the sky. *She* was very lucky) or after it (The elevator opened for *him* on the 14th floor, and *Alec* stepped out quickly[2]). In this case we call it cataphora. Actually, there are classifications, which have been defined depending on the grammatical category of the anaphora, its type, the position of its referent in the text, etc. (Mitkov, 2002).

In this work we studied the classification based on the anaphora's grammatical category and we have specifically focused on the pronominal anaphora[3] (Mitkov, 2002), (Ferrández, 1998).

Additionally, we did a contrastive analysis to compare the use of the pronominal anaphora in different languages. At this stage, we noticed some characteristics particular to the Basque language. On the one hand, we do not have different forms for the third person pronouns, and demonstrative determiners are used as third person pronominals (Laka, 2000). On the other hand, there is no gender distinction in the Basque pronominal system.

We also studied the description of the demonstratives with anaphoric function in some Basque grammars (Euskaltzaindia, 1985; 1993), (Zubiri & Zubiri, 1995), (Patrick & Zubiri, 2001), (Laka, 2000) and (Hualde & Ortiz de Urbina, 2003) as well as in some previous studies of this phenomenon (García Azkoaga, 1998) and (Garzia Garmendia, 1996). However, these works study the subject from another point of view.

As a conclusion of this study, we decided to focus on the demonstrative determiners *hau* (this), *hori* (that), *hura* (that/he/she/it), only when they behaved as pronouns, because these are the elements which represent the pronominal anaphora in Basque.

We also took into account the specific features of the referents because they are interesting for a later automatic treatment of pronominal anaphora. All these issues will be explained in the next sections.

3 Building an anaphorically annotated corpus

3.1 Previous work

In the literature, we have found bibliographic references to corpora annotated both anaphorically and coreferentially. Some of the sources for this study were: The Lancaster Anaphoric Treebank (UCREL) (Garside et al., 1997), the MUC Coreference Task (MUC-7) (Hirschman, 1997), the corpus of the University of Wolverhampton (Mitkov, 2000), part of the Penn Treebank Corpus (Ge, 1998) and the DRAMA scheme (Passoneau and Litman, 1997). All of the aforementioned corpora annotations have been carried out in and for the English language.

However, we also consulted resources for other languages, such as the TIGER Project (Kunz & Hansen-Schirra, 2003) for German. Similar work has been carried out at the University of Prague (Hajič & Urešová, 2004), where the corpus has been annotated at a pragmatic level, including the annotation of coreferential elements.

Finally, Navarro et al., (2003) carried out the study in Spanish. The Cast3LB[4] Corpus has been tagged pragmatically with the help of an annotation tool. This tool has marked the anaphoric and coreferential relationships (including ellipsis) as well as the corresponding referents.

[2] In Mitkov, 2002; p.19.

[3] When we talk about anaphora, we take into consideration both anaphora and cataphora.

[4] Cast3LB and Eus3LB are part of the 3LB project (Palomar et al., 2004).

3.2 Building an anaphorically annotated corpus in Basque

So far, we have mentioned the different studies carried out in the field of anaphorical and coreferential corpus annotation. In this section, we specify what we have already tagged in the Eus3LB Corpus and we explain the criteria defined for the annotation.

The 50.000 words corpus we worked with consists of news texts and it was first disambiguated at a morphosyntactical level using a Constraint Grammar (CG) (Karlsson et al. 1995). After that, the shallow parser identified the different phrase units (chunks, clauses, etc.). Both the morphosyntactical disambiguation tool (Aduriz et al. 2004) and the syntactic analyser[5] have been developed by the IXA research group[6].

The example in figure 1 shows the morphosyntactic analysis, which is our starting point for anaphora's annotation.

Ben Amor *ere ez da Mundiala amaitu arte etorriko Irunera,* ***honek*** *ere Tunisiarekin parte hartuko baitu Mundialean.*

*(***Ben Amor*** *as well is not coming to Irun until the World Championship is over, because* ***he*** *will play with Tunisia in the World Championship.)*

```
/<Ben>/… ("" /Ben/ IZE IZB … %SIH)
/<Amor>/… ("" /Amor/ IZE IZB … %SIB)
/<ere>/ ("ere" LOT LOK EMEN @LOK)
/<ez>/("ez" PRT EGI @PRT %ADIKATHAS)
/<da>/ ("izan" ADT A1 ... %ADIKATBU)
/<Mundiala>/…("Mundiala" IZE IZB…%SINT)
/<amaitu>/("amaitu" ADI SIN ... %ADIKAT)
/<arte>/ ("arte" IZE ARR DEK ... %SINT)
/<etorriko>/("etorri" ADI SIN ... %SINT)
/<Irunera>/("irun" ADI SIN ... SINT)
/<,>/<PUNT_KOMA>/
/<honek>/ ("hau" DET ... %SINT)
/<ere>/ ("ere" LOT LOK EMEN @LOK)
/<Tunisiarekin>/…("Tunisia" IZE LIB… %SINT)
/<parte>/ ("parte" IZE ARR DEK … %SINT)
/<hartuko>/("hartu" ADI SIN …%ADIKATHAS)
/<baitu>/ ("*edun" ERL MEN …%ADIKATBU)
/<Mundialean>/…("mundial" ADJ IZO… %SINT)
/<.>/<PUNT_PUNT>/
```

Fig. 1 The result of the morphosyntactical[7] parser (automatic analysis)

The morphosyntactical analysis shown in this figure presents the analysis of each word in a line. The tags we have taken into account for the anaphorical annotation are noun phrases, which are marked with the following tags. `%SIH` denotes the beginning of a noun-phrase with two or more elements; label `%SIB` marks the end of a noun-phrase. `%SINT` appears when the noun-phrase has only one element.

3.2.1 Annotation Format

As mentioned above the anaphoric items we wanted to tag in the corpus are the demonstratives *hau* (this)*, hori* (that) and *hura* (that/ he/ she/ it). Besides, we have to take into account that the texts to mark were syntactically tagged. The noun phrases formed by a demonstrative pronoun have the label `%SINT` and this is in fact, the pronominal anaphora.

First, we marked these demonstrative determiners as anaphoric elements, using the tag `[ANAnum]`[8]. Then we tried to look for their correct antecedents and the anaphora's corresponding referents in the text were tagged with the `[REFnum]`. If referents consisted of more than one element, they were marked with the tag `[REF-Bnum]` at the beginning of the phrase and `[REF-Enum]` at the end of the phrase.

This is exemplified in figure 2: *honek* (he) is the anaphora and *Ben Amor* (proper noun) its antecedent.

```
/<Ben>/…("" /Ben/ IZE IZB …%SIH) [REF-B5]
/<Amor>/…("" /Amor/ IZE IZB … %SIB) [REF-E5]
/<ere>/ ("ere" LOT LOK EMEN @LOK)
/<ez>/ ("ez" PRT EGI @PRT %ADIKATHAS)
/<da>/ ("izan" ADT A1 ... %ADIKATBU)
/<Mundiala>/…("Mundiala" IZE IZB … %SINT)
/<amaitu>/("amaitu" ADI SIN … %ADIKAT)
/<arte>/("arte" IZE ARR DEK … %SINT)
/<etorriko>/ ("etorri" ADI SIN … %SINT)
/<Irunera>/<HAS_MAI>/("irun" ADI SIN…%SINT)
/<,>/<PUNT_KOMA>/
/<honek>/ ("hau" DET … %SINT) [ANA5]
/<ere>/ ("ere" LOT LOK EMEN @LOK)
/<Tunisiarekin>/…("Tunisia" IZE LIB… %SINT)
/<parte>/("parte" IZE ARR DEK … %SINT)
/<hartuko>/("hartu" ADI SIN … %ADIKATHAS)
/<baitu>/ ("*edun" ERL MEN … %ADIKATBU)
/<Mundialean>/…("mundial" ADJ IZO … %SINT)
/<.>/>PUNT_PUNT>/
```

Fig. 2 Example of the anaphoric tagging (manual)

In those cases where we did not find the referent in the three previous sentences, we looked for them in the two subsequent

[5] This chunker has an accuracy of 85%
[6] http://ixa.si.ehu.es
[7] POS and chunking

[8] '*ANA*' for anaphora and *(num)* for its corresponding number, depending on its appearing order.

sentences, due to the fact that they could be cataphora. If this was the case, we marked them with `[CATnum]`. The cataphora's referents, for their part, were marked with the tag `[CAT-REFnum]` if they consisted of a single element. When they had two or more elements, we used `[CAT-REF-Bnum]` for the first element of the phrase and `[CAT-REF-Enum]` for the last one.

For those cases where we found neither a referent in the three previous sentences nor in the two subsequent ones, we defined the label `[?ANAnum]`.

3.2.2 Annotation criteria and features of the referents

Based on our experience in the annotation process, we defined different criteria:

a) Cohesive elements in Basque often include a demonstrative: *hau da* (that is to say), *harekin eta honekin* (with that and this), *honetaz gain* (apart from this), *horren ondorioz* (as a consequence of that). We considered they do not really have any specific referent. Sometimes we considered the referent could be the whole previous paragraph or even the whole text.

b) Sentences with 'to be' as main verb: *Jolasa$_i$ baita hau$_i$* (Because it$_i$ is a game$_i$). In these cases where the demonstrative made reference to the predicative noun phrase were not considered as anaphoric pronouns.

For the moment, we decided to categorize both the mentioned cohesive elements and the sentences with the main verb 'to be' as special cases and not to take into account for future studies.

3.2.3 Annotating the referent: some linguistic conclusions

Once the anaphoric elements and their referents were tagged, we analyzed the next of both:

- The distance between the anaphoric element and its antecedent.
- The grammatical features of the referent; whether it is a single element, a noun phrase or a whole sentence.
- The number of the anaphoric element and its referent.
- The case of the anaphoric element and its referent.

We will explain the conclusions we got from the characteristics noted above.

The demonstrative pronoun *hau* (this) appears 177 times in a 50.000 words corpus. With respect to the position of the anaphora's referents, in 86% of the cases they appear before the anaphoric elements, and in 14% of the cases are cataphora, that is to say, the referent comes after the anaphoric element.

Regard to the distance between the pronoun *hau* (this) and its referent, in 59% of the cases they appear in the same sentence, in some cases they are in the previous sentence and just in few cases they come in the two or three previous sentences.

Finally, we have detected that in 73% of the cases the antecedent, is a noun phrase, while 27% it is a sentence. Here we have a representative example of the first case:

[Ben Amor]$_i$ ere ez da Mundiala amaitu arte etorriko Irunera, honek $_i$ ere Tunisiarekin parte hartuko baitu Mundialean.

([Ben Amor]$_i$ is not coming to Irun before the world championship is finished, since he$_i$ will play with Tunisia in the World Championship).

The demonstrative *hori* (that) appears 251 times in the same corpus. In 99% of the cases, it is anaphora and in the 50% of cases, the referents appear in the same sentence. Regarding the structure of the referent, all of them appear in subordinate clauses and they make reference to an idea or a proposition, rather than to a specific element or person. For example:

(...) [euskaraz egiten den musika oro mespretxatzea]$_i$, hori$_i$ dun kezkagarria iruditzen zaidana.

([despising all music done in Basque]$_i$, that$_i$ is what I think is worrying).

The third demonstrative determiner we have analyzed *hura* (that), is occurs 321 times. This demonstrative is almost always anaphora (98%), rather than cataphora. In 64% of the cases it is in the same sentence of the anaphora and only sometimes in the previous sentence (33%). All the references are noun phrases and more specifically they are usually proper nouns.

Banesto galduta dabil, [Miguel Indurain]$_i$ erretiratu zenetik, hura$_i$ ordezkatuko duen norbaiten bila baitabiltza.

(Since [Miguel Indurain]$_i$ retired, Banesto is lost. They are looking for somebody to replace him$_i$).

Therefore, as in other languages, in Basque most of the demonstrative determiners are also anaphoric while cataphora is not so common (appearing mainly with the demonstrative determiner *hau*). In regard to the distance

between the anaphora and its referents, they mostly appear in the same sentence. In addition the three analyzed determiners involve different grammatical structures: for the determiner *hau* (this), the antecedent is usually a noun phrase, referring to a person or a specific element. Most of the antecedents of the determiner *hura* (that) are also noun phrases and they also refer to a person. The determiner *hori* (that) behaves in a different way; its antecedents are mostly sentences which make reference to a previously stated idea.

Moreover, we also studied the case agreement between the anaphoric element and its antecedents. We determined that it was not an important feature to find antecedents. As concerns the number of both the anaphora and its antecedent, we saw that they agree in 99% of the cases.

The results are compiled in the following table:

	HAU (this)	**HORI** (that)	**HURA** (that/he/ she/it)
Pronouns	**177**	**251**	**321**
Anaphora	**86 %**	**99 %**	**98 %**
Cataphora	14 %	1 %	2 %
Referent in the same sentence	**59 %**	**50 %**	**64 %**
Referent in the previous sentence	32 %	47 %	33 %
Others	9 %	3 %	3 %
Referent Noun Phrase	**73 %**	33 %	**100 %**
Referent sentence	27 %	**67 %**	0 %

Table 1 Results of the study

4 Conclusions and future work

The study carried out has been a useful start in defining some criteria for anaphora's annotation in Basque. Based on these criteria, we plan to tag a larger corpus in the near future. In order to facilitate this task, we are adapting the results of our study to the 3LB-RAT annotation tool (Saiz Noeda et al., 2004).

As of this date we have tagged the demonstrative determiners used as pronouns in Eus3LB corpus. Altogether we have marked 700 anaphoric elements. Actually, the corpus is being tagged by two annotators to obtain annotation agreement and measures of quality.

In regards to the specific features of the annotated determiners, we can say that some of them corroborate the consulted bibliographic statements, while others simultaneously open new perspectives to continue researching other characteristics of the anaphora and its referents. For the moment, we are mainly interested in defining the sources of knowledge needed to identify the referents such as other morphological information, syntactic dependencies or semantic features.

The results obtained from this work will be helpful for allowing the development of an automatic anaphora resolution tool.

5 Bibliography

Aduriz I., Aranzabe J.M., Arriola J.M., Díaz de Ilarraza A., Gojenola K., Oronoz M,. Uria L. (2004). *A Cascaded Syntactic Analyser for Basque.* CICLing 2004. Seoul, Korea.

Euskaltzaindia (1985). *Euskal Gramatika: Lehen Urratsak I*. Nafarroako Foru Gobernua: Euskaltzaindia.

Euskaltzaindia (1993). *Euskal Gramatika Laburra: Perpaus bakuna*. Bilbo: Euskaltzaindia.

Ferrández Rodríguez A. (1998). *Aproximación computacional al tratamiento de la anáfora pronominal y de tipo adjetivo mediante gramáticas de unificación de huecos.* Doktoretza-tesia. Departamento de Lenguajes y Sistemas Informáticos. Universidad de Alicante.

García Azkoaga I. (1998). *Erlazio anaforikoak argudiozko testuetan.* In Koherentzia, kohesioa eta konexioa: testuratze baliabideak. Hizkuntzaren azterketa eta irakaskuntza, Larringan & Idiazabal (ed.). Gasteiz: EHU-UPV & Arabako Foru Aldundia.

Garside, R. Fligelstone, S. and Botley S. (1997). *Discourse annotation: anaphoric relations in corpora*. In Garside, R., Leech, G. and McEnery, A. (Eds.) Corpus annotation: linguistic information from computer text corpora, 66-84. London: Addison Wesley Longman.

Garzia Garmendia J. (1996). *Hura, bera, eta abarren adar gehiago.* Senez, num. 18. EIZIE Euskal Itzultzaile, Zuzentzaile eta Interpreteen Elkartea.

Ge, N. (1998). *Annotating the Penn Treebank with Coreference Information.* Internal

report, Department of Computer Science, Brown University.

Hajič J. & Urešová Z. (2004). *The Prague Dependency Treebank*. Presentation for the IXA Research Group.

Hirst G. (1981). *Anaphora in Natural Language Understanding.* Berlin: Springer-Verlag.

Hirschman, L. (1997) *MUC-7 coreference task definition.* Version 3.0.

Hualde J.I. & Ortiz de Urbina J. (2003). *A grammar of basque*. Berlin: Mouton de Gruyter.

Karlsson F., Voutilainen A., Heikkilä J. Anttila A., editors (1995). *Constraint Grammar: a language-independent system for parsing unrestricted text*. Vol. 4 Natural Language Processing. New York, Berlin Mouton de Gruyter.

Kunz K. & Hansen-Schirra S. (2003). *Coreference Annotation of the TIGER Treebank*. In Proceedings of the Workshop Treebanks and Linguistic Theories, Växjö, Sweden.

Laka I. (2000). *A Brief Grammar of Euskara, the Basque Language*. HTML-ko dokumentua. Euskarako errektoreordetza, Euskal Herriko Unibertsitatea. http://www.ehu.es/grammar

Mitkov, T. and Barbu, C. (2000). *Improving pronoun resolution in two languages by means of bilingual corpora*. Proceedings of the Discourse, Anaphora and Reference Resolution Conference (DAARC, 2000), 133-137. Lancaster, UK.

Mitkov R. (2002). *Anaphora resolution*. London: Longman.

Navarro B., Civit M., Martí M. A., Marcos R., Fernández B. (2003). *Syntactic, semantic and pragmatic annotation in Cast3LB.* In Proceedings of the Shallow Processing of Large Corpora. A Corpus Linguistics Workshop, Lancaster, UK.

Palomar M., Civit M., Díaz A., Moreno L., Bisbal E., Aranzabe M., Ageno A., Martí M.A. y Navarro B. (2004). *3LB: Construcción de una base de datos de árboles sintáctico-semánticos para el catalán, euskera y español.* XX. Congreso SEPLN, Barcelona.

Passoneau, R. and Litman, D. (1997). *Discourse segmentation by human and automated means.* Computational Linguistics 23 (1), 3-139.

Patrick, J. D. & Zubiri, I. (2001). *A student grammar of Euskara.* München: Limcom Europa.

Saiz Noeda M., Izquierdo R. (2004). I*3LB-RAT: una herramienta para la anotación referencial.*beramia 2004. Puebla, México.

Zubiri I., & Zubiri E. (1995). *Euskal Gramatika osoa*. Bilbo: DIDAKTIKER.