Empirical study of the relevance of semantic information for anaphora resolution: the case of adverbial anaphora

Klara Ceberio [1], Itziar Aduriz [2], Arantza Díaz de Ilarraza [1] and Inés García Azkoaga [3]

[1] IXA Group,
University of the Basque Country, 649 p.k.
20018 Donostia
{klara.ceberio} {jipdisaa}@ehu.es
http://ixa.si.ehu.es

[2]Department of Linguistics, University of Barcelona,
08007 Barcelona
itziar.aduriz@ub.edu

[3]Department of Basque Philology, University of the Basque Country,
01007 Vitoria-Gasteiz
ines.garciaazkoaga@ehu.es

Abstract. In this paper we present the referential tagging of part of the EPEC Corpus of Basque (26,000 words). We extended our annotation from pronominal anaphora to include other types of referential structures, such as proper names, and nominal and adverbial anaphora. We describe the criteria defined for the annotation and the problematic cases we found in the tagging process. We particularly focus here on adverbial anaphora, with the aim of drawing some conclusions related to the features of antecedents in order to test whether semantic information can be helpful in future computational treatment of this phenomenon.

Keywords: reference, anaphora, adverbial anaphora, reference resolution.

1 Introduction

This paper presents some results of the referential tagging of part of the EPEC Corpus of Basque. This corpus consists of 26,000 words, corresponding to newspaper texts, classified in different domains. We particularly focused on adverbial anaphora and our aim has been to draw some conclusions related to the features of antecedents which must be correctly linked to their referents, testing whether semantic information can be helpful in anaphora resolution.

In previous works, we dealt with the tagging of pronominal anaphora (Aduriz et al., 2007) and we defined the guidelines for referential tagging (Aduriz et al., 2008) taking into account the problems we found in the annotation process. The typology of

the anaphors we treat is based on the work done by Garcia-Azkoaga (2004), which includes pronominal, nominal and adverbial anaphors. The interpretation of the anaphoric relation in adverbial cases is not always easy to resolve automatically, especially when the interpretation cannot be grammatical and must be done using semantic information. This interpretation could be based on different relations, such as the part-whole relation (Winston et al., 1987), or discursive, in other words, indirect.

In the prototypic example below, world knowledge is needed to correctly interpret the anaphors.

(1) ***Herri batera*** *iritsi ginen.* ***Eliza*** *mendixka batean zegoen.*

'We arrived **in a village**. **The church** was located on a hill'.

The *church* is a building mentioned previously, *village* semantic information is also needed to find the antecedent in some locative adverbial anaphors we found in the target corpus.

(2) *Profesional mailako klub guztiek* ***euren prestakuntza zentroa*** *dute 14 urtetik gorako gazteentzat.* ***Han*** *ikasketak eta futbola dira betebehar nagusiak.*

'All professional football clubs have **their own training centre** for young people. Their main obligations **there** are football and studies'.

As the antecedent of the local adverb *han* /there/ is a noun phrase in the nominative case (without morphological information about place), world knowledge is needed to interpret that *prestakuntza zentroa* /training centre/ is a place where young footballers go to prepare and train hard.

There are other types of anaphors, like the associative ones, where the correct antecedent can be chosen according to semantic criteria. The information we obtain from the semantic field of the anaphoric word can be helpful; for example, we can connect 'flood' with 'water'.

(3) ***Uholde batek*** *Blayaisko (Gironde departamendua) zentral nuklearreko aktibitateen gelditzera behartu du zenbait astetarako.* ***Urak*** *Bordeleren ondoko zentralaren zati oso bat estali zuen abenduaren 27an gauean eta goizaldean izaniko ekaitzaren ondorioz". (EGUNKARIA, 2000-01-07, 11. or.)*

'A flood forced the activity of the Blayais (Department of Gironde) nuclear power station to be stopped for some weeks. **The water** covered a part of the nearby power station on the night of 27th December and the following morning'. (EGUNKARIA, 7-01-2000, p. 11)

This paper is organized as follows: in section 2 we describe the main features of Basque and the basis of the corpus we have annotated. In section 3, the reference tagging process is explained, in which we have established a typology for our tagging process, including: proper nouns, pronominal anaphors, nominal anaphors and local adverbs with anaphoric value. The analysis and results obtained from our study are presented in section 4. After that, section 5 deals with the anaphoric adverb and the

relevance of semantic information for its resolution. Finally, in section 6, some conclusions are drawn and future work is outlined.

2 Main features of Basque

Basque is not an Indo-European language and differs considerably in grammar from the languages spoken in the surrounding regions. It is an agglutinative language, in which grammatical relations between components within a clause are represented by suffixes.

The Reference Corpus for the Processing of Basque (EPEC) is a 300,000 word collection of written standard Basque that has been automatically tagged at different levels (morphology, surface syntax, phrases). Part of this collection was obtained from *Euskaldunon Egunkaria* (not accessible at this moment), the only daily newspaper written entirely in standard Basque, in the second half of 1999 and in 2000. The articles were chosen so that they covered an assorted range of topics (economics, culture, international, local, opinion, politics, sports, entertainment). This corpus is being used for Natural Language Processing and, despite its small size, it is a strategic resource for a minority language like Basque.

We based this study on the part of EPEC related to newspapers because, from a socio-discursive perspective (Maingueneau, 2005), the discourse of these types of texts is related to a concrete activity, depending on the potential audience and covering general topics. We considered the following domains: economics, international, Europe, politics, sports. Knowing these domains allowed us to better specify the semantic fields related to the adverbial anaphoric expressions.

3 Tagging process

Our annotation begins with an annotated corpus that provides us with an easier environment in which to work, focusing on the specific structures that could be part of a reference chain.

The corpus has been morphosyntactically analysed by means of MORFEUS (Alegria et al., 1996). After that, two automatic taggers (rule-based and stochastic taggers) disambiguate at lemmatization level. Finally, entities, chunks and complex postpositions are identified by means of the following tools: i) EIHERA, which identifies entities (Institution, Person and Location) (Alegria et al., 2006); ii) IXATI Chunker (Aduriz et al., 2006), which identifies verb chains, noun phrase units, and complex postpositions.

As previously mentioned, we extended the annotation from pronominal structures to include nominal anaphoric and other referential structures, such as proper nouns, in order to cover a wider range of this phenomenon. The tagged corpus contains 26,000 words and the MMAX2 application (Müller & Strube, 2006) was used for this tagging (adapted to the established requirements). The tagging process was carried out manually.

In order to give consistency to our guidelines, the tagging process was split up into two steps: 1) we asked two linguists to tag the same part of the corpus and the problematic cases were checked with the help of a third linguist, acting in the role of final arbiter and 2) when the guidelines had been revisited, taking into account based on the conclusions reached in the first step, one linguist manually annotated the corpus, and the problematic cases were checked with another linguist.

After obtaining the results of this annotation, we obtained the semantic information for the adverbial anaphors by looking them up manually in EusWordNet (Basque version of WordNet) (Agirre et al., 2006).

3.1 Reference typology

We have already commented that a study of the pronominal anaphora in Basque was carried out before we extended our annotation. The typology we work with is based on the work of Garcia-Azkoaga (2004) and includes proper names and pronominal, nominal and adverbial anaphora. We defined and tagged the following types of references:

a) Proper Names

Repetition of proper names (entities) referring to people, places or organizations.

(4) *Bigarren itzulian hautatuko dute errektorea,* ***Montero*** *eta* ***Perezen*** *artean.* ***Monterok*** *lortu zituen boto gehien atzoko bozketan eta* ***Perez*** *izan zen bigarrena, hamar boto gutxiagorekin.*

'The rector will be elected in the second round, between **Montero** and **Perez**. **Montero** achieved the most votes in the first election and **Perez** was second, with 10 votes less'.

b) Pronominal anaphors

This type of anaphora is represented by a pronoun that needs an antecedent to be interpreted correctly. We must mention that Basque lacks real pronouns and that demonstratives are used for this purpose (*hau* /this/, *hori* /that/ (nearer than '*hura'*) and *hura* /that/), with all declension cases.

(5) *Adituek uste dute* ***lau DF5*** *baino ez daudela zabalduta, baina* ***haietako*** *bakoitzak lau megatoi ditu eta EEBBak, Errusia edo Europa jotzeko gaitasuna.*

'The experts think that only **four DF5s** are deployed, but each of **them** contains four megatons and the capacity to destroy the USA, Russia and Europe'.

c) Nominal anaphors

We distinguished two types of nominal anaphors; loyal anaphors and conceptual ones:

c.1) Loyal anaphors

These repeat a part of the antecedent or the whole antecedent. They can take the same lexeme and repeat it, adding or changing grammatical information using declension cases (in Basque), postpositions or attributive elements (adjectives).

(6) *Galarza Mexiko hiriburuko **Reclusorio Norte espetxean** sartu dute behin-behinekoz, estradizio tramitea egin arte. **Espetxe honetako** giza baldintzak oso txarrak dira.*

'Galarza was arrested and taken to **the Reclusorio Norte prison** during the extradition procedure. The living conditions **in this prison** are really bad'.

c.2) Conceptual anaphors

These anaphors conceptualize, sum up or evaluate an antecedent by means of a noun phrase. These conceptual anaphors always end with a demonstrative. The antecedent can be either a noun phrase or an utterance. We can usually identify this type of anaphor replacing the entire noun phrase using only the demonstrative. These anaphors provide us with an intratextual reference.

(7) *Gaur egun eskola inoiz baino ahalegin handiagoa egiten ari da ikasleen irakurzaletasuna bultzatzen. **Asko irakurtzen duen ikasleak arrakasta handiz gauzatzen dituela ikasketak** diote. Eta **baieztapen borobil honen aurrean** eskola asko irakurtzen duten haurrak ateratzen saiatzen da, bide pedagogiko guztiak urratuz.*

'Nowadays, the school is making a big effort to motivate students to read. It is said that **students who read a lot complete their studies successfully**. **This categorical assertion** leads many schools to take children who read a lot, breaking all pedagogical rules'.

d) Adverbial anaphors (locative):

These are represented by local adverbs: *hemen* /here/, *hor* /there (near here)/, *han* /over there/. Their interpretation is anaphoric, because they usually refer to a place or space previously mentioned in the text. There is another adverb, *bertan*, that can be used in all three cases (*hemen/bertan* /here/, *hor/bertan* /there (near here)/, *han/bertan* /over there/).

(8) *UNAMET Nazio Batuetako misioaren egoitzaren **bi eraikini** eraso zieten, timortarrak **han** babestuta zeudela.*

'**Two buildings** of the UNAMET United Nations mission were attacked, while the people of Timor were **there**'.

After defining this reference typology, we tagged the 26,000 word corpus. In the next section some of the results are outlined.

4 Analysis and some results

As previously mentioned, the input of the tagging process is obtained automatically from the output of our chunker. Most of the chunks have been marked as markable by means of a simple preprocessing. Temporal, predicative connectors and locutions have been removed from the set of markables.

Of the resulting 4950 markables, 1026 were included in an anaphoric set or a referential chain. Table 1 shows the frequency of occurrences of each type of

anaphoric expression for the domains selected in the corpus. The following table gives a brief summary of the results:

Table 1. Frequency of anaphoric structures in percentages.

Domains	a) P. Nouns	b) Pronominals	c) Nominals		d) Adverbials
			Loyal	Conceptual	
Sports	28.20	8.97	**47.43**	11.53	1.28
Europe	31.45	20.16	**38.70**	8.06	0.80
Politics	30.61	16.32	**43.53**	5.44	1.70
Economics	28.57	16.48	**40.65**	13.18	0
World	**40.31**	17.99	32.11	5.01	0.91

In this first approach, we appreciate that the domain of the news texts determines the number of each type of reference that appear. Nevertheless, there is a general tendency for the most common type of anaphor to be the loyal one, and the least common to be the adverbial anaphor. Proper nouns are more frequent in news about the world.

In the previous section, the relation of Basque demonstratives (mostly used for pronominals) and adverbs was mentioned and this is the main reason for choosing adverbial anaphoric structures for this work. The declension case (morphological form) of the antecedent does not always provide us with the grammatical information needed to link this type of anaphor to the correct referent. As the grammatical information is insufficient, we attempted to determine whether semantics can be helpful.

5 Adverbial anaphora

In this study we wanted to emphasize the importance of using semantic information for the automatic processing of references. The first steps of our research focused on pronominal anaphora (Aduriz et al., 2007). We extended the study of pronominal anaphora to other types of references, such as repetitions and proper nouns (Aduriz et al., 2008) and, in this paper we focused mainly on adverbial anaphora, specifically on spatial adverbial anaphora.

In Basque, there are three basic place adverbs: *hemen* /here/, *hor* /there (near here)/ and *han* /over there/, and their equivalent *bertan* /here/, /there/, or /over there/. These adverbial forms, deriving from the demonstratives *hau* /this/, *hori* /that (near here)/ and *hura* /that/ (Hualde & Ortiz de Urbina, 2003), show the three degrees of proximity and they establish an intratextual reference. The adverbial phrases can be marked by a declension case that indicates space (*etxera* /(to) home/) as we will see in the next example:

(9) *Amaia **etxera** joan zen. **Han** zegoen bere ama bazkaria prestatzen.*

'Amaia went **home**. Her mother was **there,** preparing the meal'.

In this case, the antecedent of the adverb *han* /there/ is *etxera* /(to) home/, and this antecedent is easy to detect because it is marked by a spatial morpheme. However, it is not always like this:

(10) *Federazioak aurreprestakuntzari arreta berezia jarri dio azken urtetotan, eta horretarako **Clairefontaineko Institutua** sortu zuen, Paris inguruan dauzkan instalazioetan. **Hara** 12 urterekin iristen dira mutikoak, kalitate kontrola pasatu ostean'.*

'In the last few years, the federation has paid particular attention to development; **the Claire Fontaine Institute** was created for this purpose in the installations near Paris. The 12-year-old boys arrive **there** after having a test.

The adverb *hara* /there/ refers to *Clairefontaineko Institutua* /the Claire Fontaine Institute/. This referent does not contain any declension case or grammatical information that would help to identify the referential link to the adverbial anaphor.

Thinking of a future automatic resolution algorithm, this could be problematic, since there is no possibility of establishing a direct grammatical relation between the two elements. In these cases the interpretations have to be discursive, which means using our world or encyclopaedic knowledge.

There are other types of adverbial anaphors where more information will be needed, as can be seen in example 11:

(11) *Joan den urtean **Chelsearekin** fitxatu zuen, eta denboraldi bat egin du **bertan**.*

'Last year he signed up with **Chelsea**, and he played a season **there**'.

In (11), /there/ refers to the football club called /Chelsea/ and not a city or a person.

From our point of view, this is a good reason for continuing with the study of adverbial anaphora (locative) and seeing how semantic information could be relevant for future automatic applications (Hendrickx et al., 2008)

5.1. Relevance of semantic information

Several researchers have used WordNet as a lexical and semantic resource for certain types of bridging anaphora; WordNet has also been used as an important feature in machine learning of coreference resolution using supervised training data (Meyer and Dale, 2002).

Apart from using syntactic information (POS and identification of noun phrases for pronominal anaphors; named entities for proper names), we are considering using semantic information to tackle the problem of detecting the correct antecedent of the anaphoric element in adverbial types.

We have carried out some experiments in extracting this semantic information from EusWordNet (Basque version of WordNet) (Agirre et al., 2006). The extraction made with the antecedents (only the core of the noun phrase is considered) obtained in the manual annotation explained in this paper gave us the first clues about how to deal with semantic information. Let us take one of the examples we studied (example 12): The hyperonyms of *areto* /hall/ in WordNet are: "hall> room> area> structure> construction> artefact> object, physical object> entity, something".

The same process was carried out for all instances of the adverbial anaphors tagged and similar chains were found.

In brief, most of the hyperonyms (77%) of noun phrases led to the same semantic classes, such as 'artefact> object, physical object'.

Other types of hyperonyms ended in 'event' or 'group'; another semantic feature to be taken into account.

(12) *Bi lagunak kazetari gisa sartu ziren **aretora** eta **bertan** zeudela Mujika sartu zen.*

'Two friends went **to the theatre** and Mujika went in while they were **there'.**

After a deeper study of hyperonym, hyponym and synonym relations in WordNet, we can obtain lists of semantically related concepts. The lists will contain referent candidates in those cases where adverbial anaphors are detected. For example, the list in Table 2 includes some of the concepts linked to the conceptual classes 'hall', 'room', 'area' and 'structure', hyperonyms of *areto* 'hall' in example 12, by means of using information about hyperonym, hyponym and synonym relations.

Table 2. List of hyperonym, hyponyms and synonyms linked to *areto* 'hall'.

Hall, hallway, antechamber, anteroom, entrance, foyer, lobby, vestibule, dorm, dormitory, residence, student_residence, manor_hall, mansion, mansion_house, corridor, passageway, passage, way, elbow_room, area, country, arena, domain, field, orbit, sphere, region, construction, altar, arcade, colonnade, loggia, arch, abutment_arch, broken_arch, camber_arch, trimmer_arch, auditorium, assembly_hall, box, batter's_box, bullpen, bema, choir, chancel, ...

It is easy to produce these lists automatically. In a future study, we plan to measure the suitability of the use of these lists in adverbial anaphora resolution.

6 Conclusions and future work

We have presented a preliminary study we carried out to test whether semantic knowledge can be helpful in the computational treatment of the reference in Basque.

The adverbial anaphora and the detection of the semantic class were the focus of our study. We manually analyzed sets of the hyperonyms of each referent in WordNet and detected a common behaviour: in 77% of cases the final semantic class of the referents is classified as "entity". In addition, we manually obtained lists of semantically related classes that can constrain the search for possible referents of adverbial anaphors.

The results obtained from this study will be helpful in further work on the development of an automatic anaphora resolution tool for Basque.

In addition, the examples annotated in this work have enriched the EPEC Corpus, which is a strategic resource for the processing of Basque.

7 References

1. Aduriz I., Aranzabe M., Arriola J. M., Atutxa A., Díaz de Ilarraza A., Ezeiza N., Gojenola K., Oronoz M., Soroa A. & Urizar R. (2006). "Methodology and steps towards the construction of EPEC, a corpus of written Basque tagged at morphological and syntactic levels for automatic processing". Corpus Linguistics Around the World. Book series: Language and Computers. Vol. 56 (1-15). Ed. Andrew Wilson, Paul Rayson, and Dawn Archer. Rodopi, Netherlands.
2. Aduriz, I., Ceberio, K. & Díaz de Ilarraza, A. (2007). "Pronominal Anaphora in Basque: Annotation issues for later computational treatment". 6th Discourse Anaphora and Anaphor Resolution Colloquium. DAARC2007, Lagos Portugal.
4. Ceberio K., Aduriz I., Díaz de Ilarraza A. & García Azkoaga I. (2008). "La anotación de la referencia sobre un corpus periodístico en euskera". XXVI Congreso internacional de AESLA. Almería.
3. Agirre E., Aldezabal I. & Pociello E. (2006). "Lexicalization and multiword expressions in the Basque WordNet". Proceedings of the first International WordNet Conference in Jeju, Korea, 22-26.
4. Garcia Azkoaga I.M. (2004). *Kohesio anaforikoa hiru testu generotan*. Adinaren araberako azterketa, Bilbao, Euskal Herriko Unibertsitatea.
5. Hendrickx I., Hoste V. & Daelemans W. (2008). "Semantic and syntactic features for anaphora resolution for Dutch". In: Proceedings of the CICLing-2008 Conference, Haifa, Israel, 2008, Berlin, Springer, 2008, p. 351-361.
6. Hualde J.I. & Ortiz de Urbina J. (2003). *A grammar of Basque*. Berlin: Mouton de Gruyter
7. Kleiber G. (1994). Anaphores et pronoms. Louvain-la-Neuve, Duculot.
8. Laka I. (2000). "A Brief Grammar of Euskara, the Basque Language". Euskarako errektoreordetza, Euskal Herriko Unibertsitatea. http://www.ehu.es/grammar
9. Maingeneau D. (2005). "L'analyse du discours et ses frontières". Marges linguistiques, 9. 1-12.
10. Meyer J. & Dale R. (2002). "Learning selectional preferences for use in resolving associative anaphora". In Proceedings of the 2002 Australasian Natural Language Processing Workshop, 2nd December, Canberra, Australia.
11. Mitkov R. (2002). *Anaphora resolution*. London: Longman.
12. Müller C. & Strube M. (2006). "Multi-Level Annotation of Linguistic Data with MMAX2". In: Sabine Braun, Kurt Kohn, Joybrato Mukherjee (Eds.): Corpus Technology and Language Pedagogy. New Resources, New Tools, New Methods. Frankfurt: Peter Lang, pp. 197-214. (English Corpus Linguistics, Vol. 3).
13. Winston M., Chaffin R. & Herrmann D. (1987). A taxonomy of part-whole relations, Cognitive Science, Volume 11, Issue 4, pp. 417-444.