Discourse Deixis and Coreference: Evidence from AnCora

Marta Recasens
CLiC - Centre de Llenguatge i Computació
Department of Linguistics
University of Barcelona
08007 Barcelona, Spain
mrecasens@ub.edu

Abstract

Few empirical studies have been conducted on discourse deixis, and no such study exists for Catalan or Spanish. This paper presents an empirical analysis of 200 000 words from the AnCora corpora annotated with discourse deixis. It returns to and tests assumptions previously made, laying out the linguistic problems we still need to account for. To this end, proposals are put forward with regard to (i) the detection of abstract anaphors, and (ii) the way their antecedents should be understood, drawing on the theory of underspecification. The quantitative and qualitative corpus analysis casts light on ways of improving the performance of coreference resolution systems by shifting the focus from the delimitation of antecedents to the detection of abstract anaphors.

1 Introduction

Natural Language Processing (NLP) work dealing with discourse deixis (e.g. (1)) has been little to date in comparison with the considerable amount of effort devoted to the automatic resolution of pronominal individual anaphora. It is probably the relative ease of identifying term-denoting NPs as well as their relatedness to named entities that accounts for this higher attraction. Following Webber's (1988) terminology, by *discourse deixis* it is meant NPs that refer to a previous discourse segment.[1] The discourse segment is referred to as an *abstract antecedent*, and the NP as an *abstract anaphor*. The abstract antecedent can be either the referent situation(s) or circumstances expressed by the stretch of text (1), or the proposition itself ("wording") as a linguistic object (2).

(1) Al voltant d'aquesta passarel·la està previst *crear un espai lliure amb vistes cap al riu Cardener.* Això implica una modificació del traçat del carrer Sant Antoni. (C)[2]
'Around this walkway it is planned *to create a free area facing the River Cardener.* This implies a modification of the route of Sant Antoni Street.'

(2) *"L'euro té potencial per a una apreciació, basada en el creixement i l'estabilitat de preus interna,"* van declarar ahir els ministres d'Economia i Finances. (C) La declaració institucional va ser emesa pel Consell de l'Euro.
'*"The euro has potential for appreciation, based on the internal growth and stability of prices,"* declared yesterday the ministers of Economy and Finance. The institutional declaration was expressed by the Euro Council.'

Only small datasets (Byron and Allen, 1998; Eckert and Strube, 2000; Navarretta and Olsen, 2008) have been annotated with discourse deixis and limited to a few anaphoric expressions. Byron (2002) emphasized that demonstrative pronouns referring to clauses or larger stretches of text abound in natural discourse, and the corpus-based study of the use of demonstrative NPs in Portuguese and French conducted by Vieira et al. (2002) also pointed out the limitation of systems restricted to anaphors with a nominal antecedent. Later annotation efforts such as OntoNotes (Pradhan et al., 2007) and ARRAU (Poesio and Artstein, 2008) have tried to overcome these limitations.

This paper presents a corpus study of discourse deixis in Catalan and Spanish. With the future goal of building a coreference resolution system for these two languages, the AnCora corpora (already annotated from the morphological to the semantic levels) are being enriched with

[1] Although from a semantico-logical point of view, discourse deixis overlaps but is not the same as *reference to abstract objects*, I will use both indistinctively, thus treating events as abstract objects. Abstract reference is also called *situation reference* (Fraurud, 1992).

[2] The language appears indicated at the end of each example: (S) for Spanish, (C) for Catalan.

Johansson, C. (Ed.)
Proceedings of the Second Workshop
on Anaphora Resolution (2008)

coreference relations. The annotation includes all NPs, which implies that those whose linguistic antecedent is a discourse segment are also encoded. I focus on written texts (newspaper articles) unlike former work done on dialogs. In addition, not only pronouns but also full NPs are annotated. Being Catalan and Spanish pro-drop languages, zero pronouns are also considered. This empirical analysis of the 200 000 words that are available at present (100 000 for each language) returns to and tests assumptions previously made. It thus lays out the linguistic problems we still need to account for.

In order to make sense of the real data, proposals are put forward with regard to (i) the detection of abstract anaphors (distinguishing between nominalizations and labels in Francis' (1994) terms), and (ii) the way their antecedents should be understood, drawing on the theory of underspecification (Poesio et al., 2006). These ideas have a twofold effect. On the one hand, they suggest that it is feasible for a coreference resolution system to automatically detect references to abstract objects, thus improving the overall performance of the system. On the other hand, they argue for the likely failure of delimiting abstract antecedents on the basis of exact boundaries.

The paper is organised as follows. Section 2 outlines previous work on discourse deixis in the field of NLP: from early theoretical accounts to more recent corpus-based approaches. The AnCora corpora are described in Section 3, where the coreference coding scheme and the reliability study are also presented. The annotated data prompts a revision of previous assumptions in Section 4 on the basis of both a quantitative and qualitative corpus analysis. Section 5 tries to make sense of problematic issues discussed in Section 4 by borrowing from other linguistic accounts. Finally, Section 6 summarizes the conclusions and outlines for future work.

2 Previous work

Reference to abstract objects first came to the scene in NLP when systems began to be developed to resolve pronominal anaphora. Soon it was realized that neuter pronouns such as *it*, and especially demonstratives like *this* and *that* often referred to linguistic units other than NPs.[3] The need to account for these pronouns required going beyond the NP level, and discourse model theories entered the scene. Webber (1988) introduced the term "discourse segments" to refer to the clausal mention in (1). Karttunen (1976) had previously talked of entities introduced by NPs and referred back to in the discourse as "discourse entities." Webber's term was meant as a complement to Karttunen's, claiming that discourse segments have their own mental reality apart from the discourse entities they contain.

These first approaches were rather theoretical. Although they used some real examples, these were selected according to what they wanted to prove and no systematic empirical study was conducted. Four ideas underlie Webber's (1988) seminal work, and these recur in subsequent works. I limit myself to mentioning them here (in her own words). Section 4 returns to this point to collate the assumptions with the empirical data from AnCora.

1. Preference for demonstratives: "Subsequent reference to a sequence of clauses is most often done via deictic pronouns."
2. Referent coercion: "Once the speaker has referred to it [discourse segment] via *this/that*, it must now have the status of a discourse entity since it can be referenced via the anaphoric pronoun *it*.
3. Required presence: "The demonstratum being something [explicitly] present in the shared context."
4. Ambiguity:[4] "All pointing is ambiguous ... The listener's choice depends on what is compatible with the meaning of the rest of the sentence."

Around the turn of the century, collections of real data became a reality and they have made it possible to collate early theoretical claims with real occurring data. Table 1 presents some of the corpora where discourse-deictic NPs (in some cases only pronouns or only demonstratives) have been annotated. These corpora were developed either with a view to developing and testing algorithms or to extracting quantitative figures about linguistic phenomena. Work in progress includes the OntoNotes coreference

[3] Notice that not all instances of these pronouns are referential in English. There are no counterparts in Catalan and Spanish to the English dummy-*it* constructions, and expletive uses are restricted to a very few constructions.

[4] Although I will argue for "non-specification" in Section 4.4, the term "ambiguity" is kept here in fidelity to Webber's original words.

ISSN 1736-6305 Vol. 2
http://dspace.utlib.ee/dspace
/handle/10062/7129

System/Study	Corpus	Anaphor	Antecedent	
			NP	Clause
Byron and Allen (1998)	English dialogs	pers.pr.	75%	7%
(PHORA)	383 pronouns	dem.pr.	25%	35%
Eckert and Strube (2000)	English dialogs	pers.&dem.pr.	45%	23%
Navarretta and Olsen (2008)	Danish texts (60K)	pers.&dem.pr.	26%	29%
(DAD)	Italian texts (55K)	(zero) pers.&dem.pr.	85%	10%
Vieira et al. (2002)	Portuguese 50 dem.NPs	dem.full NP	62%	38%
	French 50 dem.NPs	dem.full NP	68%	32%
Botley (2006)	English (300K)	*this*		56%
	spoken discourse	*that*		32%
	news	*these*		10%
	literature	*those*		2%

Table 1: Corpora annotated with discourse deixis

corpus (Pradhan et al., 2007) (although only the heads of VPs are considered as antecedents), and the ARRAU corpus (Poesio and Artstein, 2008), where all clauses are presented as potential antecedents for the coders to decide.

Annotating this information, however, is still an open problem, since "it is not completely clear the extent to which humans agree on the interpretation of such expressions" (Poesio and Artstein, 2008). The largest existing corpora annotated with coreference information (for the MUC and ACE campaigns) all restrict to encoding NPs whose antecedent is also an NP. No corpus-based study exists for Catalan or Spanish.

3 Coreference annotation in AnCora

The AnCora corpora – Annotated Corpora for Catalan and Spanish[5] (Taulé et al., 2008) consist of two 500 000-word corpora for Catalan (AnCora-Ca) and Spanish (AnCora-Es), mainly newspaper and newswire articles. Both corpora are annotated at different levels of linguistic description: morphological (PoS and lemmas), syntactic (constituents and functions), and semantic (argument structures, thematic roles, semantic verb classes, named entities, and WordNet nominal senses). They are being enriched with coreference annotation: 100 000 words for each language are available at present.

The Catalan subset contains 31 079 NPs (10 975 coreferent); the Spanish subset 29 179 NPs (10 499 coreferent). In terms of figures similar to the ones reported in former works (Table 1), it emerges that 42% (AnCora-Ca) and 86% (AnCora-Es) of neuter personal pronouns, and 59% (AnCora-Ca) and 57% (AnCora-Es) of neuter demonstrative pronouns have a clausal antecedent.[6] The Catalan and Spanish neuter pronouns are the equivalent forms of English *it*, *this*, and *that*. The correspondence, however, is not one-to-one, as the range of uses of the Romance forms is much more restricted than those of English. This factor together with differences in the way each corpus has been annotated probably account for the differences with Table 1.

Finally, an interesting ratio not provided by former work is the ratio of discourse-deictic NPs to the total number of coreferent NPs:[7] 3% in Catalan, and 4% in Spanish. Discourse-deictic NPs represent thus a small group in comparison with coreference links between NPs. The fact that reference to abstract objects is more typical of dialogues than newspaper texts contributes to these low figures. However, although discourse deixis accounts for less than 5% of all coreference links, successfully detecting this percentage could result in a statistically significant improvement on the overall performance of a coreference resolution system by reducing the number of false positive links.

[5] Available from `http://clic.ub.edu/ancora`

[6] If relative frequencies are computed including zero pronouns, as done for Italian in (Navarretta and Olsen, 2008), then we obtain that 5% (AnCora-Ca) and 4% (AnCora-Es) of pronouns have a clausal antecedent.

[7] The NP count includes pronouns as well as definite and demonstrative NPs, since these are the forms that can be abstract anaphors.

3.1 Coding scheme

The coreference annotation follows a two-step process: (i) an automatic stage, (ii) a manual one. Only markables corresponding to NPs are automatically encoded with XML tags thanks to the morphosyntactic annotations. Discourse segments are marked at the manual stage when they are needed to mark up a link. The coding guidelines (Recasens et al., 2007) distinguish between identity and discourse deixis relations depending on the type of antecedent: the former have an NP as antecedent, the latter a discourse segment (including at least one clause).

Discourse deixis relations are further split into "segment" (3) and "textual scene" (4) to differentiate those antecedents that fall within the sentence unit from those that go beyond. Segmental discourse deixis takes an attribute specifying the semantic type of the reference: event-token, event-type, or proposition.

(3) *Un pirata informático consiguió robar los datos de 485.000 tarjetas de crédito* ... El robo fue descubierto ... (S)
'*A hacker managed to steal the data of 485,000 credit cards* ... The robbery was discovered ...'

(4) *Latinoamérica concluyó hoy su participación en la "Bolsa de Turismo" de Berlín con un balance preliminar un tanto pesimista porque *0* no tuvo la cantidad de visitantes esperada. La competencia de Asia, los altos precios de los pasajes y la relación dólar-marco alemán, fueron los obstáculos señalados por varios países para impedirles lograr sus objetivos. La escasa presencia de interesados provocó que en algunos puestos el material no se distribuyera por completo. ...*
Fernández se mostró optimista con respecto a que la situación mejore. (S)
'*Latin America ended today its participation in the tourism stock market of Berlin with a preliminary balance rather pessimistic since (it) did not have the expected number of visitors. The competition by Asia, the high prices of the tickets, and the relation dollar–German mark, were the obstacles pointed out by several countries that impeded them to achieve their aims. The scarce presence of interested people caused some stalls not to have all their material distributed. ...*
Fernández showed herself optimistic with respect to the improvement of the situation.'

3.2 Reliability study

The coding scheme of AnCora was tested in a reliability study involving eight participants (six undergraduates and two graduates of linguistics, all of them native Spanish speakers), who annotated the same two texts independently. Given the high cost – both in time and money – of conducting such experiments,[8] this small-scale study was meant as a first approximation to the quality of the scheme. Although high agreement scores (α=.85 and α=.90) were obtained for the coreferent vs. non-coreferent distinction, the four instances likely to be annotated as discourse deixis turned out to be a major source of disagreement. Annotators coincided largely in the NPs chosen as abstract anaphors, but they often disagreed in the extension of abstract antecedents, although the discourse segments usually overlapped. These results are in line with the conclusions reported by Artstein and Poesio (2006) from a similar experiment on dialogues.

4 Evidence from AnCora

A quantitative and qualitative analysis of the 200 000 words coreferentially annotated from the AnCora corpora offer the chance to revisit Webber's (1988) assumptions (Section 2) by commenting on those examples arising most questions among annotators, thus taking a bottom-up perspective. Throughout the discussion linguistic problems that have not been accounted for become apparent.

4.1 Preference for demonstratives

Webber (1988) states that there is a preference to use demonstratives *this* and *that* vs. the pronoun *it* to refer to a previous discourse segment. To test whether this preference also holds for Catalan and Spanish, discourse-deictic NPs were extracted and sorted by morphological form. Figures for absolute and relative frequencies are presented in Table 2. Given that the antecedents of discourse deixis are usually not longer than one sentence, I focus on this group. As far as pronouns are concerned, Catalan makes a slightly greater use of demonstratives (15.04%) than personal (13.16%) pronouns. No preference for demonstratives, however, is observed in Spanish, where personal pronouns (13.64%) are twice as much used as demonstratives (6.17%). With regard to full NPs, these are the forms that participate most frequently into discourse deixis, both in Catalan and Spanish (50%). This high percentage calls

[8] For this study, the anotation of two texts required 10 hours per coder.

ISSN 1736-6305 Vol. 2
http://dspace.utlib.ee/dspace
/handle/10062/7129

Coreferent NP	AnCora-Ca				AnCora-Es			
	≤ 1 sentence		> 1 sentence		≤ 1 sentence		> 1 sentence	
	#	%	#	%	#	%	#	%
Full NP								
Definite	80	30.08	4	1.50	78	25.32	13	4.22
Demonstrative	47	17.67	9	3.38	52	16.88	3	0.97
Possessive[a]	–	–	–	–	15	4.87	0	0
Pronoun								
Personal (neuter)	35	13.16	1	0.38	42	13.64	1	0.32
Zero	30	11.28	1	0.38	11	3.57	0	0
Relative	17	6.39	0	0	71	23.05	1	0.32
Demonstrative (neuter)	40	15.04	2	0.75	19	6.17	2	0.65
Total	249	93.61	17	6.39	273	93.51	20	6.49

[a] Given that possessive determiners are always preceded by the definite article in Catalan, possessive full NPs are included in the definite group.

Table 2: Distribution of discourse deixis in AnCora

thus for their inclusion in a coreference resolution system.

4.2 Referent coercion

The assumption that a discourse segment turns into a discourse entity when it is referred to by a demonstrative (Webber, 1988) suggests that the sequence

segment ... this ... it

is the prototypical one. Such a pattern, however, needs to be extended to allow for full NPs, which broadens the range of possible patterns. AnCora includes instances of:

- segment ... full NP ... full NP (3)
- segment ... full NP ... segment (5)[9]

(5) "El movimiento de las arenas hace difícil *saber dónde están enterradas las minas*. No es una cuestión de mapas el saber dónde están *0* y cuál es el estado de las minas", añadió *0* ... retirar estas minas, *de las que no se sabe la situación exacta*. (S)
"'The movement of the sand makes it difficult *to know where the mines are buried*. The knowledge of where (they) are and what is their state is not a matter of maps", (he) added ... removing these mines, *of which the exact situation is unknown*.'

[9]The zero pronoun is marked with *0* and with the corresponding pronoun in brackets in the English translation.

- full NP ... segment ... NP (6)

(6) Dos arqueòlegs nord-americans acaben de muntar un gran enrenou amb una nova teoria ... *Els primers pobladors del continent americà podrien haver estat habitants de la península Ibèrica que fa 18.000 anys van travessar l'Atlàntic.* Aquesta és la provocativa teoria de dos arqueòlegs nord-americans... (C)
'Two North American archaeologists have just caused quite a commotion with a new theory ... *The first inhabitants of the American continent could have been inhabitants of the Iberian Peninsula that crossed the Atlantic 18,000 years ago.* This is the provocative theory of two North American archaeologists ...'

A usual way of elaborating on a previous NP, providing additional information, is by using a clausal mention in a subsequent reference.

4.3 Required presence

Although Webber (1988) claims that a discourse-deictic pronoun must point to something which explicitly appears in the discourse, the fact that text comprehension is highly constructive accounts for counterexamples in which the antecedent cannot be easily recovered from the preceding context.

(7) Para el presidente, "es evidente" que el PSOE no llega a nuevos sectores de la población por lo que debe hacerse "un gran esfuerzo de cambio, de mensajes claros y de valores

que permitan que ese mensaje sea asumido por los nuevos componentes de una sociedad española que ha cambiado mucho". (S)
'For the president, "it is evident" that the PSOE does not arrive to new sectors of the population, so that there is the need for "a big effort of change, of clear messages and of values that allow this message to be assumed by the new components of a Spanish society that has changed a lot."'

These cases resemble "bridging" (Clark, 1977) in that the reader has to carry out a process of inference to arrive at the antecedent, since this does not appear explicitly. Clark's original idea of brigding needs to be extended twofold: (i) it implies not only NP antecedents (*her house ... the door*) but also discourse segments, and (ii) both definite and demonstrative NPs are possible bridging anaphors.

4.4 Non-specificity

Webber (1988) points out the ambiguous nature of discourse deixis, especially with respect to the extension of the antecedent (already highlighted by the reliability study, Section 3.2). From her point of view, different extensions of a discourse segment might imply different referents. Hence, the use of the term "ambiguity." More accurately, however, the point at issue is the non-specific nature of the antecedent (8). Webber proposes the "right frontier" as a cue, according to which only discourse segments on the current right frontier of the discourse tree can yield referents for abstract anaphors. The problem lies in choosing *which* segment on the right frontier[10]. This non-specificity also applies to full NPs, especially those of the kind *la situación* 'the situation' in (4).

(8) Agassi insistió que *0* puede ser mejor jugador para volver a tener un gran año, aunque *0* no le garantice los triunfos que *0* tuvo en 1999. (S)
'Agassi stressed that (he) could be a better player to have a great year again, although (it) does not guarantee him the victories that he had in 1999.'

5 Making sense of the data

It follows from the discussion in the previous section that, from a computational perspective, the automatic resolution of discourse deixis can profit from insights on:

- Detecting the kind of NPs that can be abstract anaphors.
- Accounting for the inherent non-specificity of abstract antecedents.

I turn now my attention to these two issues. First, I suggest linguistic cues for detecting abstract anaphors by focusing on nominalizations and *labels* in Francis' (1994) terms. Second, I draw on the theory of *underspecification* (Poesio et al., 2006) to account for the continuum of specificity on which antecedents of abstract anaphors seem to lie.

5.1 Detecting abstract anaphors

Although both pronouns and full NPs can be abstract anaphors, the former amount to no more than a reduced set (neuter, relative and zero pronouns) while the latter constitute an infinite set, thus posing greater difficulties. An analysis of the Catalan and Spanish forms observed in discourse deixis suggests that three specific groups of nouns are potential candidates to be abstract anaphors. Table 3 illustrates absolute and relative frequencies as well as examples (from AnCora-Es) of each group.

- Nominalizations
  - Deverbal nouns (e.g. *exportación* 'exportation') can be detected by the presence of a nominalizing affix (Spanish *-ción, -miento, -cia, -aje*, etc.).
  - Verbal forms converted into nouns (e.g. *apoyo* 'support') can be identified by extracting from the morphological parser those pairs of tokens that can be either a noun or a verb. Only a limited set of verbal forms can undergo such conversion: first-person indicative, first- or third- person subjunctive, and past participle.
- "Cousins"
  They are non-deverbal abstract nouns denoting things that are conceptually event-like. E.g. *éxito* 'success' (no verb such as the English *succeed* exists in Spanish).
- Labels
  The term is borrowed from Francis (1994): labels[11] are nominal groups that function as pro-forms used "to encapsulate or package a stretch of (written) discourse." They

[10] "When there is more than one [discourse segment], ... I will have nothing to say here about how the choice between them is made." (Webber, 1991)

[11] Also known as *anaphoric nouns* or *shell nouns* (Schmid, 2000).

ISSN 1736-6305 Vol. 2
http://dspace.utlib.ee/dspace
/handle/10062/7129

	Nominalizing affix	Noun / Verb	Cousins	Labels	
				Neutral	Evaluative
#	39 (34%)	32 (28%)	7 (6%)	13 (12%)	23 (20%)
e.g.	*concentración* 'concentration', *pensamiento* 'thought', *entrenamiento* 'training'	*acuerdo* 'agreement', *visita* 'visit', *accidente* 'accident'	*éxito* 'success', *desventaja* 'disadvantage', *presencia* 'presence'	*asunto* 'issue', *caso* 'case', *cuestión* 'matter'	*actitud* 'attitude', *dificultades* 'difficulties', *objetivo* 'objective'

Table 3: Typology of abstract anaphors (from AnCora-Es)

have both a naming and encapsulating function, and are extremely common in the press, summarising the preceding co-text. Two criteria can be used to recognize a label: (i) the head noun is non-specific, and (ii) it requires lexical realization in its co-text. Depending on the semantics, they fall into two main groups:

- Neutral labels (e.g. *situación* 'situation'): they simply build a package from a stretch of discourse.
- Evaluative labels (e.g. *razón* 'reason'): they inform the reader how a chunk of discourse is to be interpreted, thus adding a positive/negative evaluation to the "package."

A list of labels can be extracted by mining the annotated corpus. Most of the labels we obtain from AnCora-Es and AnCora-Ca coincide with those reported in Francis (1994), and the rest show semantic similarities. Although Francis (1994) points out that modification is a device for adding extra meaning to labels, modified labels were the minor group in AnCora (18%), which supports that they are pro-forms.

5.2 Underspecified abstract antecedents

The end of Section 4 argued for the non-specificity of abstract antecedents, which becomes evident as soon as one attempts to delimit their exact boundaries. This point is supported by Botley (2006), who reports that "indirect anaphora definitely poses difficulties for corpus-based linguistics, in that almost 30% of abstract reference cases analysed were hard to classify straightforwardly. This is because antecedents lack clear surface linguistic boundaries." Francis (1994) also comments on the same: "Labels do not necessarily refer to a clearly delimited or identifiable stretch of discourse. It is the shift in direction signalled by the label and its immediate environment which is of crucial importance for the development of the discourse."

Both the reliability study and the corpus analysis seem to suggest that rather than a dichotomy, specificity constitutes a continuum, extending along the range of boundaries that the antecedent can take, from specific boundaries (1) to fuzzy ones (8). I believe that the theory that best accounts for this reality is that of underspecification which has been provided for some lexical ambiguities like homonymy and polysemy. Poesio et al. (2006) have extended it to anaphora: "With certain types of ambiguity the ambiguous expressions may be left unresolved in the right context." They argue that the final interpretation of some (mereological) pronouns in dialogues is not fully specified, but only "good enough" for the listener's purposes. They give (9) as an example, where it is not clear whether the pronoun *that* refers to *the orange juice* which has been loaded into *the tanker car*, or the tanker car itself, or indeed whether that matters. This leads them to formulate the Justified Sloppiness Hypothesis.

(9) so then we'll
... we'll be in a position to
load the orange juice into the tanker car
... and send *that* off

According to this hypothesis, there are cases when an anaphor has two potential antecedents x and y, which are elements of an underlying mereological structure with summum $x \oplus y$. There is still a fourth interpretation that can be derived from such a summum: $z \lhd (x \oplus y)$, which is a p-underspecified interpretation in which the anaphor is interpreted as denoting an element

z included in the summum. What is crucial is that all four possible interpretations are equivalent for the purposes of the plan.

Following this line, I take the account further by adapting the Justified Sloppiness Hypothesis to cover discourse deixis for both pronouns and full NPs. Instead of four we have three possible interpretations:

(i) x is the largest/maximal discourse segment.

(ii) y is the shortest/minimal discourse segment.

(iii) $y \triangleleft z \triangleleft x$, in which z is a p-underspecified interpretation denoting a discourse segment whose extension lies between the minimal y and the maximal x.

Again, the three interpretations are equivalent for the purposes of communication. It might be possible to establish a mapping between anaphor form and specificity of antecedent, e.g. complex NPs usually specify the antecedent whereas labels tend to leave their antecedent p-underspecified.

Underspecification provides the theoretical framework for which we do not have to consider instances of discourse deixis with a non-well delimited discourse segment as linguistically incorrect, but as wholly legitimate references whose role in the discourse does not require that they be fully specified. Therefore, I conclude that annotation efforts should not assume that anaphoric expressions referring to an abstract object always have a clearly identifiable antecedent.

Concerning the automatic resolution of discourse deixis, the fact that only 5% of all coreference links involve discourse deixis has a twofold effect: on the one hand, automatically delimiting exact abstract antecedents will not be very helpful; on the other hand, successfully detecting discourse-deictic NPs can stop them from being included in a wrong chain, and thus have beneficial effects on the overall performance of the coreference resolution system. With the help of a morphological parser and the extracted list of labels, if a nominalization or a label is encountered by the system whose head does not match any previous NP but matches a previous verbal form, then it is more likely to be discourse deictic than non-coreferent or coreferent with an NP.

6 Conclusion

This paper takes an empirical approach to discourse deixis in Catalan and Spanish, opening the field to these two languages. Emphasis was put on the need for complementing theoretical accounts based on a limited set of examples. The coreference annotation in the AnCora corpora includes discourse segments and thus offers a significant amount of data on the basis of which we can approach this topic from a bottom-up perspective and from the point of view of coreference resolution. Such an approach lays bare the complexities of abstract anaphora and shows the shortcomings of former theoretical claims.

The corpus study presented here differs from former work in several aspects. Apart from obtaining data for two languages not studied in this respect, it is not limited to pronouns or demonstrative NPs, but includes both pronouns (personal, demonstrative, relative and zero) and full NPs (definite, demonstrative and possessive). It deals with written discourse, unlike most existing work on dialogues. Finally, the annotated dataset is much larger than those used so far.

The coding scheme is in accordance with real data, thus covering discourse deixis as it occurs in the two Romance languages under analysis. Two datasets of 100 000 words each served to return to four assumptions made by Webber (1988) and point out the problematic issues with the help of real examples. On the basis of the annotated corpora, the resolution of discourse deixis was divided into two different tasks: detecting the units that are likely to be abstract anaphors, and delimiting the boundaries of abstract antecedents. Given that discourse-deictic NPs are found to represent less than 5% of all coreferent NPs, I stressed that the main focus of attention should be the detection of abstract anaphors rather than the delimitation of the exact boundaries of their antecedents.

My main claim was that abstract anaphors conform to certain criteria: they are either one of a specific set of pronouns or full NPs in the form of nominalizations, cousins, or labels. A set of labels was extracted by mining the corpus, which largely overlapped with those reported by Francis (1994). As for abstract antecedents, the non-specificity that makes its delimitation so difficult was accounted for with the semantic theory of underspecification according to Poe-

ISSN 1736-6305 Vol. 2
http://dspace.utlib.ee/dspace
/handle/10062/7129

sio et al. (2006). Hence, I proposed that they lie on a continuum, from fully specified to p-underspecified, and that it is legitimate for them to be left underspecified.

From a computational perspective, the data discussed and suggested insights open new avenues for the automatic resolution of discourse deixis and coreference resolution by extension. Computational approaches should bear in mind that not all references require to be fully specified for successful communication, and so annotation efforts must not insist on setting fixed boundaries in every case. Whereas it is possible for a coreference resolution system to detect abstract anaphors with the help of a morphological parser and an extracted list of labels, there is no point in trying to delimit the exact antecedent when it is underspecified. Detection alone can result in a statistically significant improvement on the overall performance of the system by reducing the number of false positive links.

Acknowledgements

Thanks to all the annotators who participated in the reliability study: Irene Carbó, Sandra García, Iago González, Esther López, Jesús Martínez, Laura Muñoz, and especially to Isabel Briz and Montse Nofre for annotating the AnCora corpora.

This work was supported by FPU-2006-08 grant from the Spanish Ministry of Education and Science, and Lang2World (TIN2006-15265-C06-06) – subproject of Text-Mess.

References

Ron Artstein and Massimo Poesio. 2006. Identifying reference to abstract objects in dialogue. In *Proceedings of the 10th Workshop on the Semantics and Pragmatics of Dialogue (BRANDIAL2006)*, pages 56–63.

Simon Botley. 2006. Indirect anaphora: Testing the limits of corpus-based linguistics. *International Journal of Corpus Linguistics*, 11(1):73–112.

Donna K. Byron and James F. Allen. 1998. Resolving demonstrative anaphora in the TRAINS93 corpus. In *Proceedings of the 2nd Discourse Anaphora and Anaphor Resolution Colloquium (DAARC1998)*, pages 68–81.

Donna K. Byron. 2002. Resolving pronominal reference to abstract entities. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 80–87.

Herbert H. Clark. 1977. Bridging. In P.N. Johnson-Laird and P.C.Wason, editors, *Thinking: Readings in Cognitive Science*. Cambridge University Press, Cambridge.

Miriam Eckert and Michael Strube. 2000. Dialogue acts, synchronising units and anaphora resolution. *Journal of Semantics*, 17(1):51–89.

Gill Francis. 1994. Labelling discourse: an aspect of nominal-group lexical cohesion. In M. Coulthard, editor, *Advances in Written Text Analysis*, pages 83–101. Routledge, London.

Kari Fraurud. 1992. *Processing Noun Phrases in Discourse*. Ph.D. thesis, Department of Linguistics, Stockholm University.

Lauri Karttunen. 1976. Discourse referents. In J. McCawley, editor, *Syntax and Semantics*, volume 7, pages 363–385. Academic Press, New York.

Costanza Navarretta and Sussi Olsen. 2008. Annotating abstract pronominal anaphora in the DAD project. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)*, Marrakech, Morocco.

Massimo Poesio and Ron Artstein. 2008. Anaphoric annotation in the ARRAU corpus. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC2008)*.

Massimo Poesio, Patrick Sturt, Ron Artstein, and Ruth Filik. 2006. Underspecification and anaphora: Theoretical issues and preliminary evidence. *Discourse Processes: a multidisciplinary journal*, 42:157–175.

Sameer S. Pradhan, Lance Ramshaw, Ralph Weischedel, Jessica MacBride, and Linnea Micciulla. 2007. Unrestricted coreference: Identifying entities and events in OntoNotes. In *Proceedings of the 1st IEEE International Conference on Semantic Computing (ICSC 2007)*, pages 446–453.

Marta Recasens, M. Antònia Martí, and Mariona Taulé. 2007. Text as scene: Discourse deixis and bridging relations. *Procesamiento del Lenguaje Natural*, 39:205–212.

Hans-Jörg Schmid. 2000. *English Abstract Nouns as Conceptual Shells. From Corpus to Cognition*. Mouton de Gruyter, Berlin.

Mariona Taulé, M. Antònia Martí, and Marta Recasens. 2008. AnCora: Multilevel Annotated Corpora for Catalan and Spanish. In *Proceedings of the 6th International Confer-*

ence on Language Resources and Evaluation (LREC 2008), Marrakech, Morocco.

Renata Vieira, Susanne Salmon-Alt, Caroline Gasperin, Emmanuel Schang, and Gabriel Othero. 2002. Coreference and anaphoric relations of demonstrative noun phrases in a multilingual corpus. In *Proceedings of the 4th Discourse Anaphora and Anaphor Resolution Colloquium*.

Bonnie L. Webber. 1988. Discourse deixis: reference to discourse segments. In *Proceedings of the 26th Annual Meeting of the Association for Computational Linguistics*, pages 113–122.

Bonnie L. Webber. 1991. Structure and ostension in the interpretation of discourse deixis. *Language and Cognitive Processes*, 6(2):107–135.

ISSN 1736-6305 Vol. 2
http://dspace.utlib.ee/dspace
/handle/10062/7129