Coreference and anaphoric relations of demonstrative noun phrases in a multilingual corpus

Susanne Salmon-Alt, Renata Vieira, Caroline Gasperin

COREFERENCE AND ANAPHORIC RELATIONS OF DEMONSTRATIVE NOUN PHRASES IN MULTILINGUAL CORPUS

RENATA VIEIRA*, SUSANNE SALMON-ALT**, CAROLINE GASPERIN*

** UNISINOS São Leopoldo, Brazil*
{renata, caroline}@exatas.unisinos.br

*** ATILF – CNRS Nancy, France*
Susanne.Alt@loria.fr

Abstract

We present a corpus study regarding the use of demonstrative noun phrases in Portuguese and French. The motivation for this study is to verify specific features related to the coreferential and anaphoric role of such expressions in written texts. These features serve as background knowledge for the development of a multilingual tool for coreference and anaphoric resolution.

1 Introduction

Recent work on anaphor resolution is pointing to the fact that different types of referring expressions (pronouns, definite descriptions, demonstratives) are based on different features or require different knowledge for reference resolution (Strube, Rapp and Müller 2002; Sant'Anna and Lima 2002; Salmon-Alt and Vieira 2002; Poesio et al 2002).

In this work, motivated by rising background knowledge for the design of a multilingual tool for anaphora resolution, we analyze in detail syntactic, discourse and semantic features specifically related to the use of demonstrative noun phrases. As primary data, we use Portuguese and French corpora of written texts.

Section 2 defines the main concepts (*coreference*, *anaphora* and *demonstrative noun phrases*) used in this study. Section 3 gives a detailed overview of the features we investigated. Section 4 describes the annotation task, the corpora and the

annotation tool. A discussion of the results is given in section 5, and section 6 presents conclusions and future work.

2 Coreference and anaphoric relations of demonstrative noun phrases

According to related work on demonstratives in the area of descriptive linguistics (Corblin, 1987), demonstrative noun phrases are considered to be interpreted based on salience of the referent. A referent can for example be salient because of a pointing gesture or a previous mention. The fact that salience based on pointing gestures is excluded in our corpus study of written discourse implies that the interpretation of demonstratives should tend to be more closely related to previous text, as the only source of salience.

Having this in mind, we designed a corpus study focusing on coreference and anaphoric relations of demonstrative noun phrases. Coreference has been defined by van Deemter and Kibble (2000) as the relation holding between linguistic expressions that refer to the same extra-linguistic entity. A slightly different discourse relation is anaphora. In an anaphoric relation, the interpretation of an expression is dependent on previous expressions within the same discourse, but the anaphor and its antecedent may refer to different referents. Therefore, an anaphoric relation may be coreferential or not, and as it is known, a particularly difficult question is to determine the relation holding between the anaphor and its antecedent. (Strand 1996; Vieira and Teufel 1997; Poesio and Vieira 1998).

An expression may be anaphoric in the strict sense that its interpretation is only possible on the basis of the antecedent, as it is in general the case of pronouns in written discourse. On the other hand, it might be coreferential without being anaphoric, in the sense that the entity has been mentioned before in the text, as it is the case of subsequent mentions of self explaining expressions such as *the champion of the 2002 world cup – the team that won the 2002 world cup championship*.

In this work, we are interested in both coreferential and anaphoric relations. The analyses have been made regarding several features of the textual antecedents of given expressions, such as verifying whether the antecedent is coreferential or not, its syntactic structure as well as certain semantic properties.

In this study, we consider demonstrative noun phrases (NPs) in Portuguese and French. These are noun phrases starting with a demonstrative determiner (Table 1) and having a head noun, such as (*cette région, esta região, this region*). In both French and Portuguese, demonstrative determiners vary in gender and number. We are not

considering demonstrative pronouns being full nominal constituents such as *este, esta, isto, aquele* (Portuguese) or *celui-ci, ceux de gauche* (French).

	Masculine		**Feminine**	
	Portuguese	**French**	**Portuguese**	**French**
Singular	este esse aquele	ce(t)	esta essa aquela	cette
Plural	estes esses aqueles	ces	estas essas aquelas	ces

Table 1: *Demonstrative determiners*

3 Criteria for the corpus analysis

3.1 Types of coreferential and anaphoric uses

One goal of our classification experiments was to investigate coreferential and anaphoric demonstratives. Relations between a demonstrative description *d* and its textual antecedent *a* (if any) were, therefore, classified depending on different categories of use.

Direct coreference: *d* corefers with a previous nominal expression *a*; *d* and *a* have the same nominal head:

(1) *a. às autoridades gregas (the greek authorities)*
d. essas autoridades (these authorities)

Indirect coreference: *d* corefers with a previous nominal expression *a*; *d* and *a* have different nominal heads:

(2) *a. a Albânia (Albania)*
d. este país (this country)

Other anaphora: the antecedent is not a nominal expression or the relation between demonstrative and its antecedent is not a coreference relation:

(3) *a. adoptar medidas de âmbito nacional (to adopt measures)*
d. essa adopção (this adoption)

These classes, based on previous work on computational processing of definite descriptions (Vieira & Poesio, 2000), enable us to evaluate the proportion of

coreferential relations and of noun phrase antecedents for demonstrative noun phrases. The reason for isolating nominal antecedents from other expressions such as verb phrases, sentences or paragraphs is to evaluate how well a system for anaphora resolution of demonstratives can perform on the basis of nominal expression relations only, a fact which seems to be reasonable within the context of the current state of the art of automatic anaphora resolution (Mitkov, 2002). The distinction between *same nominal head* and *different nominal head* allows us to observe the frequency of semantic bridging between a demonstrative and its antecedent, and gives therefore an idea about the need of additional lexical knowledge sources.

The *other anaphora* class represents the uses of demonstratives that require special techniques to identify antecedents that are not noun phrases (sentences, paragraphs or sets of those) and antecedents that do not refer to the same entity as the anaphoric demonstrative.

3.2 Syntactic structure of demonstrative noun phrases

French and Portuguese demonstrative noun phrases have been classified according to the presence or not of adjectival, prepositional and relative-clause modifiers. Each demonstrative NP belongs to one of the following classes, growing in terms of complexity:

Noun phrases containing only a head noun without modifiers (DET N), also including a few cases of Portuguese or French elliptical noun phrases such as *ce dernier – esse último* (*this latter one*):

(4) *cette région – esta região (this region)*

Noun phrases with adjectival modifiers (DET (ADJ N | N ADJ)):

(5) *ces pratiques abusives – estas práticas abusivas (these abusive practices)*

Noun phrases with prepositional phrases introduced by *de (of)* and perhaps adjectival modifiers (DET (N | ADJ N | N ADJ) OF (N | N ADJ | ADJ N)):

(6) *ces usages vulnérables de la route (these vulnerable uses of the road)*

(7) *esta ajuda de emergência (this help of emergency/emergency help)*

Nouns phrases with relative clauses and perhaps adjectival modifiers (DET (N | ADJ N | N ADJ) REL_PRO):

(8) *ces oiseaux que la loi protège (these birds that the law protects)*

(9) *este grave problema social que sofrem os cidadães (this serious social probem that suffer the citizens)*

The reason therefore was to explore a possible relation between complexity of syntactic structures and discourse roles of demonstrative NPs, traditionally considered as being predominantly coreferential or anaphoric (Corblin, 1987). Our underlying hypothesis is that demonstratives, whose interpretation is mainly context dependent, are preferably realized through simple noun phrase structures. In other terms, following (Löbner 1985), the arguments for their semantic function are provided mainly by textual antecedents and not through noun phrase complements.

3.3 Size of antecedents

Also important for resolving anaphora is knowledge about certain characteristics of the antecedents. In preliminary analyses of the corpus, we noticed that demonstrative expressions tend to refer to ideas expressed throughout the texts (cases such as *this problem, this situation, these facts*). These abstract concepts have as antecedents not just clearly defined entities such as those referred to by noun phrases, but whole sentences or paragraphs as well as disjoint parts of texts.

To check the frequency of these cases in our corpus, we divided the antecedents into four categories:

Antecedents that were NPs (for which a single head noun can be clearly identified):

(10) *a. a substituição da fuligem por um produto menos nocivo (the substitution of the soot by another less harmful product)*

d. este problema (this problem)

Antecedents identified as being part of a sentence (bigger than an NP but not a complete sentence):

(11) *a. estas taxas são aumentadas periodicamente (these taxes are increased periodically)*

d. este procedimento do Governo italiano (this procedure of the Italian government)

Antecedents that were full sentences:

(12) *a. A Comissão das Comunidades Europeias declarou pretender investir no transporte ferroviário de mercadorias, principalmente para distâncias de pelo menos 500 quilómetros e, se possível, superiores a 1 000 quilómetros. (The European Community Comission declared its intention of investing on rail transport for goods, mainly for distance greater than 500 km and, if possible, greater than 1000km.)*

d. esta posição (this position)

Antecedents that were larger than one sentence (or not clearly identifiable by only one linguistic expression).

As systems for anaphor resolution usually consider only relations holding between noun phrases, our analysis will shed some light on how this assumption may influence the performance of such systems.

3.4 Semantic Analysis

Finally, certain basic semantic features (*concreteness* vs. *abstractness* and well-defined lexical relations) were analyzed for the head nouns of demonstrative NPs and their antecedents.

First, the head nouns of both demonstratives and their antecedents were classified manually as abstract or concrete nouns according to distinctions presented in (Cegalla, 1996; Cunha & Cintra, 1985):

Concrete nouns refer to real existing beings (names of people, places, institutions, species), or else, things that imagination considers like that (fary).

Abstract nouns refer to notions, actions, states and qualities. They are nouns referring to things that do not exist in the world by themselves; they depend on other beings to exist: *beauty*, love, trip, life.

This enabled us to compare the matching between concrete and abstract features of demonstrative and their antecedents. We also verified the syntactic structure of the antecedents for concrete demonstratives to test our hypothesis that concrete demonstratives have a tendency to have noun phrases as antecedents instead of more complex structures such as sentences or paragraphs.

Second, we analyzed the semantic relation holding for those cases classified as indirect coreference, that is

H ypernymy:

(13) *a. Angola (Angola)*
d. esse país (this country)

Synonymy:

(14) *a.. o período de 1991/1995 (the period of 1991/1995)*
d. essa altura (this time)

Discourse deictic (anaphora that rely on particular positions within the text, as in *este último* (*this last one)*, analyzed in Corblin, 1999):

(15) *a. o Conselho de Estado grego (the Greek State Council)*
d. este último (this latter)

Other semantic relations (less well defined relations):

(16) *a. a proteção das aves (the birds protection)*
d. neste domínio (this domain)

As these semantic relations were observed within the context, pairs such as *obras cinematográficas - aquele tipo de criação artística / cinematographic works – that kind of artistic creation* were considered as synonymy. Also, the analysis was mainly made regarding the semantic relations holding between the head nouns of the two noun phrases (exceptions are special cases such as the previous examples *that kind of*). Therefore the while the relation holding between *1989* and *that time* was considered as hypernymy, the one holding between *the period of 1991/1995* and *that time* was considered as synonymy.

4 Corpus annotation

4.1 Corpus

The corpus of our study consists of French and Portuguese texts from the MLCC corpus. This multilingual parallel corpus contains written questions asked by members of the European Parliament and corresponding answers from the European Commission, published in the Official Journal of the European Commission, C Series, Written Questions 1993.

In order to have about 250 demonstratives for each language, we had to select a corpus of approximately 50.000 words, corresponding to 90 question-answer pairs. Table 2 presents a description of the resources we used. Although the texts are parallel texts, the French version has a greater number of demonstratives (291) than the Portuguese version (243).

Corpus	Language	Nb words	Demonstratives
MLCC	French	~ 50000	291
	Portuguese		243

Table 2: Corpus for the study of demonstrative NPs

4.2 Annotation tool

MMAX[1] is a tool for corpus annotation (Müller & Strube, 2001), supporting annotation of electronic corpora, providing an interface for creating markables, annotating relations between markables, and browsing the annotation. It allows the specification of user-definable attributes for the markables and computes the Kappa reliability measure for different annotations. All data is represented in XML format. To annotate the corpus with the MMAX tool, we first transformed the corpus from its original SGML TEI standard to XML MMAX format, generating MMAX words and text files.

```
<words>
   <word id="word_49">milhares</word>
   <word id="word_50">de</word>
   <word id="word_51">refugiados</word>
</words>
```

Figure 1: *Words basic file*

```
<markables>
   <markable classification="indirect"
     id="markable_3"   pointer="markable_8"
     np_form="demNP" span="word_135..word_136"/>
</markables>
```

Figure 2: *Markables output file*

The basic input format contains *word* elements as shown in Figure 1. The output of the annotation process is an XML file, containing a list of markables and their attributes as shown in figure 2.

4.3 Annotation task

The annotation procedure was divided into three phases: selecting the markables,assigning the antecedents, and classifying the uses. We separated the task of selecting an antecedent from that of classifying types of use, according to previous experience (Vieira, Salmon-Alt & Schang, 2002). suggesting that low inter-annotator agreement was at least partly due to the complexity of the task. We considered that a

[1] http://www.eml.org/english/Research/NLP/Downloads

native speaker identifies an antecedent in a more intuitive way if the task does not include classification at the same time. Phase 1 was done by one annotator for each language and the annotations of phases 2 and 3 were done by two subjects for each language.

Phase 1 - Selection of markables: In this phase, one annotator uses MMAX to mark the demonstrative descriptions in the corpus. Each demonstrative NP corresponds to a markable to be analyzed in the following phases.

Phase 2 - Identification of textual antecedents: Two annotators (native speakers) mark the antecedents of the previously selected demonstratives[2].

Phase 3 - Classification of the coreference and anaphoric relations: In the third phase of the annotation, the relationship between demonstratives and their textual antecedents were classified, according the uses defined in section 3.1. Additionally, we checked the values for the syntactic and semantic features also introduced in the previous section.

5 Results

Here we show the resulting analysis of the features described in section 3: general distribution of coreferential and anaphoric use of demonstrative NPs (5.1), their syntactic structure (5.2), the type of antecedents for demonstrative anaphora (5.3) and some basic semantic characteristics of demonstrative NPs head nouns (5.4). In section 5.5 we correlate some of these properties.

5.1 Types of coreferential or anaphoric uses

Since demonstratives are likely to identify their referent on the basis of salience, and given our material (written texts), we expected them to be necessarily related to previous discourse, and preferentially in a coreferential way. Our classification results do support these hypotheses for both French and Portuguese corpora.

Category %	**French**	**Portuguese**
Direct coreference	32	34

[2] Antecedents greater than one sentence as well as antecedents not clearly identifiable by a single text chunk were not marked due to practical reasons related to the tool (the selection of such long markables would prevent the visual distinction of markables and antecedents in the texts).

Indirect coreference	21	18
Other anaphora	47	48
Total	100	100

Table 3: *Classification of French and Portuguese demonstratives*

The results in table 3 show that demonstratives are context dependent, with more than half of them being coreferential with previous NPs. The other half are either coreferential with antecedents which are not NP or not coreferential.

Demonstratives whose antecedents were not explicitly marked are also included in the *other anaphora* class. The fact that we observed a high number of abstract head nouns for demonstratives of this group (*manner, range, problem, reason, purpose, situation, case, decision, context, …*) led us to investigate further correlations between concreteness/abstractness of head nouns and type of anaphoric use (section 5.5).

5.2 Syntactic structure

Table 4 presents the distribution of French and Portuguese demonstratives over th|rench as well as in Portuguese, present few modified structures: only 20 % in both languages are subject to adjectival, prepositional or relative clause modification.

Syntactic structure %	**Demonstrative NPs**		**Definite NPs**	
	French	**Portuguese**	**French**	**Portuguese**
DET N	80,4	80,2	35,4	40,8
DET(ADJ N | N ADJ)	10,3	7,6	22,6	22,7
DET (N |ADJ N | N ADJ) OF N	7,2	7,3	30,0	28,7
DET (N | ADJ N | N ADJ) REL_PRO	1,1	0,8	2,3	2,3
Other	1,0	4,1	9,7	5,5
Total	100	100	100	100

Table 4: *Syntactic structure of demonstratives, compared to definites*

When compared to the structure of definite descriptions investigated in previous work (Vieira, Salmon-Alt & Schang, 2002), we noticed the difference between definites and demonstratives regarding the proportion of noun phrases belonging to class 1 (head noun without modifiers). This proportion is about 37% for definites in the two languages, whereas for demonstratives this structure is verified for about 80% of the cases. One possibility is that definite descriptions are more often interpreted on the basis of semantic information, but not necessarily anaphorically to entities introduced within the previous discourse, as first observed in (Poesio & Vieira, 1998). If one considers that the quantity of semantic information increases with the

adjunction of modifiers, then the fact that they belong mainly to complex classes would confirm this hypothesis. Moreover, one can suppose that the more semantic information is given within the definite noun phrase itself, the less important is the interpretational dependency on information provided by previous discourse.

Regarding demonstratives, in French as well as in Portuguese, we have few modified demonstrative NPs (only about 20%). As opposite to the explanation for definites, this small proportion can be seen as a confirmation of the interpretational property of demonstratives to refer to something already salient through previous discourse. Indeed, the lack of modifiers and therefore less semantic information about the referent increases the need of supplying this information by the discourse context and might be seen as a confirmation for considering demonstratives as mainly anaphoric expressions rather than discourse new, according to the Giveness Hierarchy model (Prince, 1981; Prince, 1992; Gundel et al 1993).

5.3 Size of antecedents

Type of the antecedent %	**French**		**Portuguese**	
	Ann1	**Ann2**	**Ann1**	**Ann2**
NP	81	68	62	65
< Sentence	9	7	22	9
Sentence	6	10	4	1
Not marked	4	15	12	25
Total	100	100	100	100

Table 5: *Type of antecedent for demonstrative anaphora*

The results in table 5 show that the antecedents for demonstrative NPs were noun phrase structures at least in 62% for all annotators. In the remaining cases the antecedents were identified as one single sentence, part of a sentence or paragraphs (which accounts for most cases of antecedents not marked). This gives us an idea of the limitation of systems that work on anaphor resolution based on NP structures only. Such a system is likely to fail on about 30% of the cases on the basis of this assumption.

From the results shown in section 5.1 (table 3), we could see that nearly 50% of the demonstratives were coreferential with previous NPs. However the number of NP antecedents identified by the annotators (table 5) sum up to 81 % of the cases,

therefore at least 30% of the demonstratives stand in other kind of anaphoric relation with previous NPs. An example is:

(17) *a. l' installation, dans la forêt pétrifiée, de neuf aérogénérateurs (the installation, in the petrified forest, of nine wind generators)*

d. cette atteinte portée à un monument d' histoire naturelle d' importance considérable (this considerable attack to a monument of natural history)

Examples of demonstrative NP head nouns, for which antecedents were not marked are *point*, *interpretation*, *efforts* or *sense*. Again, we have mainly abstract nouns, for which a specific textual antecedent is hard to identify in the text. Therefore, the relation between the semantics of the demonstrative head noun and the size or type of antecedent were investigated, as presented in section 5.5.

5.4 Semantic analysis

Concrete vs. abstract demonstratives and antecedents

Semantic classification %	French	Portuguese
Concrete	21	22
Abstract	79	78
Total	100	100

Table 6: *Demonstrative NP head nouns*

Table 6 shows the results regarding the semantic analyses of demonstrative head nouns, according to the abstract and concrete distinction (section 3.4). Regarding their distribution, the results confirm our hypothesis: there is a clear predominance of abstract head nouns in demonstrative noun phrases (near 80 %). Another positive point is the equal distribution of concrete and abstract head nouns in French and Portuguese since the classification was done manually by different annotators. Table 7 shows the semantic classification of the antecedent head nouns, for each annotator and for both languages. Whereas demonstrative noun phrases were predominantly abstract for both languages, the classification of the antecedents were found to be less consistent. In Portuguese, the antecedents were mainly concrete (57%) and for French, mainly abstract (67%).

Semantic Classification %	French Ann. 1	Ann. 2	Average	Portuguese Ann. 1	Ann. 2	Average
Concrete	32	33	33	66	49	57
Abstract	68	66	67	34	51	43
Total	100	100	100	100	100	100

Table 7: *Semantic classification of antecedent head nouns*

Given the classification results for the demonstrative NPs (table 6), this means also that demonstrative anaphora are sometimes used to re-classify the entity referred to by the antecedent by a more abstract noun, this observation being consistent with previous linguistic analyses of discourse roles of demonstrative NPs (Corblin, 1987). An example for such a case is:

(18) *a. une essence super à teneur en octane plus élevée (a super benzine with higher octane)*
d. cette dernière qualité (this latter quality)

Furthermore, we also investigated the correlation between concrete and abstract demonstratives and their antecedents as well as the relation between concrete and abstract demonstratives with the size of the antecedents. The results are reported in section 5.5.

Semantic relations

Another semantic feature we analyzed was the semantic relation holding between indirect coreferential demonstratives and their antecedents. Table 8 shows the distribution over the semantic relations presented in section 3.4. Concerning well-defined semantic relations, there is a clear predominance of hypernymy. However, other frequent type of relation is the *other semantic relations* class, referring to cases often based on general semantic inference, which do not correspond to a precise lexical semantic relation.

Semantic relation %	Portuguese Ann 1	Ann 2	French Ann 1	Ann 2
Hypernymy	41	65	33	40
Synonymy	5	4	7	10
Discourse deictic	3	4	15	19
Other semantic relations	51	27	45	31
Total	100	100	100	100

Table 8: *Semantic relations for demonstratives (indirect coreference)*

5.5. Cross feature analyses

Concreteness/abstractness and anaphoric relations

Semantic classification %	French		Portuguese	
	Concrete	Abstract	Concrete	Abstract
Direct coreference	50	28	64	24
Indirect coreference	34	11	31	23
Other anaphora	16	61	5	53
Total	100	100	100	100

Table 9*: Semantic of head nouns vs. anaphoric relation*

The observation of many abstract head nouns for non coreferential demonstratives (section 5.1) raises the question of whether the semantic features of demonstrative head nouns (i.e. abstract or concrete) allow predictions about the type of the anaphoric relation between the demonstrative NP and its antecedent.

Table 9 shows this relation for French and Portuguese demonstratives. They confirm our intuition by showing that more than 80% of demonstratives with a concrete head noun enter in a coreference relation with their antecedent, whereas it is the case for only 40% of demonstratives with an abstract head noun. This observation could be used as a baseline for evaluating demonstrative anaphora resolution separately for concrete and abstract head nouns.

Concreteness/abstractness of demonstratives and antecedents

Dem NP	Antecedents %			
	Concrete	Abstract	not NP	Total
Concrete	94	2	4	100
Abstract	30	25	45	100

Table 10: *Semantics of demonstratives and antecedents (Portuguese)*

Dem NP	Antecedents %			
	Concrete	Abstract	not NP	Total
Concrete	92	8	0	100
Abstract	7	67	26	100

Table 11*: Semantics of demonstratives and antecedents (French)*

In section 5.4 we presented the classification into *concrete* or *abstract* for the head nouns of demonstrative NPs and antecedents. Here, we analyze the interconnection between these features. Tables 10 and 11 show the percentage of concrete and abstracts antecedents, depending on concreteness or abstractness of the demonstratives, according to one annotator for each language. Demonstratives were considered to be either concrete or abstract, but antecedents are sometimes not expressed as NPs.

For concrete head noun demonstratives, the antecedent head noun is concrete as well most of the times (over 90 % for both languages). This observation could be important for anaphor resolution heuristics, since it allows excluding less plausible antecedent candidates for concrete demonstratives, provided a suitable lexicon containing the needed semantic information. An example follows:

(19) *a. associations ecologists (ecologist associations)*
d. ces associations (these associations)

Cases where concrete demonstratives are anaphoric to abstract head noun antecedents are rare in both languages. We found here cases of metonymy (20) and process-result polysemy (21). In both cases, the relation could not be said coreferential in a strict sense.

(20) *a. le vol Air Lingus EA 643 (the flight Air Lingus EA 643)*
d. cet avion (this plane)

(21) *a. une demande d'information (a request for information)*
d. cette letter (this letter)

For demonstratives with abstract head nouns, things are less straightforward. It seems however that the probability that they refer to entities introduced previously by concrete head nouns is low (between 0.07 and 0.3, depending on the language), although it is still higher than the inverse case (abstract antecedent for concrete demonstrative). This could be explained by the fact that additionally to result-process polysemy (*informatics, activity*), this configuration includes also generic anaphora (classes referred to by expressions like *this genre*, *this species*), as shown in the examples:

(22) *a.. des entrepises informatiques (informatics companies)*
d. cette activité industrielle (this industrial activity)

(23) *a.. les rares chèvres sauvages (the rare wild goats)*
d. cette espèce (this species)

Finally, we present an example of a demonstrative NP with abstract head noun whose antecedent has also an abstract head noun. However, this is a combination that cannot be predicted, since the antecedents of abstract demonstratives were non NPs in up to 50% of the cases.

(24) *a. l'exode de milliers d'Albanais (the outflow of millions of Albanians)*
d. cet afflux massif de réfugiés auxquels elles doivent fournir une assistance humanitaire (this massive influx of refugees to whom they should provide humanitarian assistance)

Semantics of demonstratives and syntactic structure of antecedents

Finally, we correlated semantics (*concrete* vs. *abstract*) of demonstratives with different syntactic structures of the antecedents (NP and not NP), investigating whether the semantic feature of a head noun makes it possible to predict the preferred syntactic structure of the antecedent. The results for one annotator per language are presented in table 12.

Demonstrative head noun	Antecedents % French		Portuguese	
	NP	not NP	NP	not NP
Concrete	100	0	94	6
Abstract	74	26	53	47

Table 12*: Semantics of demonstratives and type of antecedents*

As a result, concrete demonstratives were related to NP antecedents for the majority of the cases for both languages (94 to 100%). Again, for abstract head nouns, it is difficult to draw conclusions, since they seem to be generally distributed over NP and non NP antecedents.

6. Agreement issues

We verified the inter-annotator agreement on classifications as well as on the identification of antecedents for each language. In order to evaluate the inter-annotator agreement on the classification task, we calculated Kappa (Carletta, 1996) for each experiment. This measure establishes $K = 0.8$ as good agreement. We calculated Kappa for the three classes (direct coreference, indirect coreference, other). We found $K = 0.79$ for French and $K = 0.65$ for Portuguese demonstratives. These results show better agreement than for previous experiments related to four different classes for

definite descriptions (Vieira, Salmon-Alt & Schang, 2002). The improvement might be related to the reduced number of classes as well as to the fact that we isolated in this experiment the identification of the antecedent from the classification task. Informal feedback from the annotators also suggests that the annotation task was easier for demonstratives than for definites. We have also compared the choice of antecedents for the two annotators of each language.

The results are presented in Tables 13 and 14. These tables show for annotators 1 and 2 in each language, cases where the antecedent was the same or not (A1=A2, A1≠A2) in correlation with the type of antecedent chosen (*direct*, *indirect*, *other* as well as those cases in which the antecedent was not marked ∅, because it was greater than a sentence). There was total agreement on the antecedents for 51% of the cases in Portuguese and 69,8% for French. Most cases of disagreement for Portuguese were related to cases where the antecedent was not marked. In some cases (around 4% in Portuguese and 9% in French) the antecedents chosen by the annotators are not the same but they are coreferential expressions themselves (Coreference(A1,A2)) which can be considered as partial agreement.

Agreement on antecedents		**#**	**%**
A1 = A2	**Direct**	61	25,1
	Indirect	31	12,7
	Other	20	8,2
	A1 = A2 = ∅	12	4,9
	Total agreement	124	51
A1 ≠ A2	**(A1 or A2) = ∅**	62	25,5
	Coreference (A1, A2)	10	4,1
	¬ Coreference (A1, A2)	47	19,3
	Total disagreement	119	49

Table 13: *Agreement on antecedents in Portuguese corpus*

Agreement on antecedents		**#**	**%**
A1 = A2	**A1 = A2 = ∅**	11	3,8
	Direct	76	26,1
	Indirect	43	14,8
	Other	73	25,1
	Total agreement	203	69,8
A1 ≠ A2	**(A1 or A2) = ∅**	29	10,0
	Coreference (A1, A2)	27	9,3
	¬ Coreference (A1, A2)	32	11,0
	Total disagreement	88	30,2

Table 14*: Agreement on antecedents in French corpus*

7 Conclusions and future work

This study investigated anaphoric and coreferential properties of demonstrative noun phrases in French and Portuguese. Having in mind the overall objective of designing a tool for definite and demonstrative noun phrase reference resolution, the main conclusions of this work are the following:

As suggested by linguistic description (Corblin 1987) and as opposed to definite descriptions (Poesio and Vieira 1998, Vieira et al. 2002), the interpretation of demonstrative noun phrases is mainly context dependant, in the sense that human annotators are able to find, for more than 80% of them, textual chunks as antecedents. Moreover, this hypothesis seems to be reinforced by the finding that over 80% of demonstrative NPs are noun phrases without any additional modifier, suggesting that this type of anaphora is less informative by itself and relies heavily on textual context.

However, the demonstrative NPs were identified as coreferential with previous NPs in about 50% of the cases only. This observation gives raise to two comments.

First, for all the cases were the antecedent is a non nominal text chunk, i.e. for more than 40% of demonstrative NPs in our corpus, it is difficult to select a precise portion of the text as an antecedent: the limits between verbal phrases, sentences and even paragraphs for presenting an idea recovered with abstract nouns such as *this manner, this situation* or *this point of view* are not easy to analyze.

Secondly, when the relation of a demonstrative and its antecedent is not a coreferential one, the amount of world knowledge and reasoning required for the resolution is very large. As for other types of nominal anaphora (Poesio et al. 2000), less than half of the cases enter in a well defined lexical relation and could therefore be resolved on the base of lexical resources such as WordNet. An additional problem is here the lack of a well developed WordNets for other languages than English.

However, as challenging as these problems may be seen, we raised several cross-language features specifically related to the discourse role of demonstrative expressions: there are not only mainly textual dependent for their interpretation (either coreferential or anaphoric), but in more than half of the cases, the antecedent is also an NP. Furthermore, classification experiments on basic semantic features of the head nouns involved in demonstrative anaphora and the related antecedents (abstract vs. concrete entity) have shown that concrete demonstratives have high tendency to take concrete NPs as antecedents (over 90%). Abstract demonstratives rely in a less strong

way on antecedent NPs (between 50% and 70%, depending on annotators and languages).

As an overall conclusion, one might keep in mind two important points: on the one hand, most of the properties we investigated seems to be cross-language, since the results are similar in French and in Portuguese; on the other hand, the specific distribution of the syntactic and semantic features for demonstrative NPs seems to justify a specific treatment of this kind of anaphora as opposed to other anaphoric expressions, such as pronouns or definite descriptions. Further work is needed for the analysis of coreferent demonstrative with non NP antecedents as well as for non coreferent anaphoric demonstratives.

Acknowledgments

This work was developed with financial support of CNPq/ProTeM-CC, INRIA and FAPERGS. We also would like to thank Gabriel Ávila Othero and our annotators Cassiano Haag, Jean-Luc Benoit, Emmanuel Schang, and Margarete Silva.

References

Carletta J. (1996). Assessing agreement on classification tasks: the Kappa statistic. *Computational Linguistics*, 22(2):249-254.

Cegalla D. P. (1996). *Novísima gramática da língua portuguesa*. São Paulo: Nacional.

Corblin F. (1987). *Indéfini, Défini et Démonstratif*. Droz, Genève.

Corblin F. (1999). Les références mentionnelles : le premier, le dernier, celui-ci. In Mettouch A., Quinyin H.: *La référence (2). Statut et processus*. Travaux linguistiques du CERLICO, P.U. Rennes.

Cunha C., Cintra L. (1985). Nova gramática do português contemporâneo. Rio de Janeiro: Nova Fronteira.

Gundel, J.; Hedberg, N., Zacharski, R. 1993. Cognitive Status and the form of referring expressions in discourse. Language 69:274-307.

Löbner S. (1985). Definites. Journal of Semantics, 4 279-326.

Mitkov R. (2002). Anaphora Resolution (Studies in Language and Linguistics). Longman.

Müller C., Strube M. (2001). MMAX: A Tool for the Annotation of Multi-modal Corpora. *Proceedings of the 2nd IJCAI Workshop on Knowledge and Reasoning in Practical Dialogue Systems*. Seattle, Wash., 45-50.

Poesio M., Vieira R. (1998). A corpus-based investigation of Definite Description Use. *Computational Linguistics*, 24(2):183-216.

Poesio M., Ishikawa T., Walde S., Vieira, R. (2002). Acquiring lexical knowledge for anaphora resolution. *Language resources and evaluation conference LREC 2002*, Las Palmas, Spain.

Prince, E.F., 1981. Toward a taxonomy of given-new information. In P. Cole, ed., *Radical Pragmatics*. Academic Press, New York, 223-256.

Prince, E.F., 1992. The {ZPG} letter: subjects, definiteness, and information status. In Thompson, S. and Mann, W. (eds.) *Discourse description: diverse analyses of a fund-raising text*. J. Benjamins, 295-325.

Salmon-Alt S., Vieira R. (2002). Nominal Expressions in Multilingual Corpora: Definites and Demonstratives. *Language resources and evaluation conference LREC 2002*, Las Palmas, Spain.

Sant'Anna V., Lima V. (2002). Resolution of demonstrative anaphoric references in portuguese written texts. *Proceedings of Portugal for Natural language Processing PorTAL 2002*, Faro, Portugal

Strand K. (1996). A Taxonomy of Linking Relations. *IndiAna Workshop*, Lancaster, England.

Strube M., Rapp S., Müller C. (2002) The influence of minimum edit distance on reference resolution. *The 2002 Conference on Empirical Methods in Natural Language Processing*. Philadelphia, Penn., US.

van Deemter K., Kibble R. (2000). On Coreferring: Coreference in MUC and related annotation schemes. *Computational Linguistics* 26(4).

Vieira R., Teufel S. (1997). Towards Resolution of Bridging Descriptions. *35th International Joint Conference on Computational Linguistics*, Madrid, Spain.

Vieira R. (1998). *Definite description processing in unrestricted text*. PhD Thesis. Centre for Cognitive Science, Edinburgh University. Edinburgh, UK.

Vieira, R., and Poesio, M., 2000. An Empirically-Based System for Processing Definite Descriptions, *Computational Linguistics*, 26(4):525-579.

Vieira R, Salmon-Alt S., Schang E. (2002). Multilingual Corpora Annotation for Processing Definite Descriptions. *Proceedings of Portugal for Natural language Processing PorTAL 2002*, Faro, Portugal.