

Nominal Expressions in Multilingual Corpora: Definites and Demonstratives

Susanne Salmon-Alt*, Renata Vieira**

* ATILF – CNRS
44 avenue de la Libération, B.P. 30687
54063 Nancy Cedex, France
Susanne.Salmon-Alt@inalf.fr

** Centro de Ciências da Comunicação e Centro de Ciências Exatas e Tecnológica,
Universidade do Vale do Rio dos Sinos
São Leopoldo, Brazil
renata@exatas.unisinos.br

Abstract

This paper presents the results of a multilingual corpus study on definite descriptions and demonstrative noun phrases. The analysis made on a parallel corpus (French and Portuguese) reinforces previous findings regarding the predominance of non-anaphoric uses of definite descriptions in English corpus. It is also shown that the use of demonstrative noun phrases, on the other hand, is more regularly based on discourse salient entities. The analysis involves syntactic issues and is oriented to the design of natural language processing tools.

1. Introduction

Knowledge about the interpretation of nominal expressions (anaphora resolution, co-reference, discourse status) is relevant to a variety of applications in the field of computational linguistics, including information extraction, natural language understanding and generation, machine translation and human-machine dialogue.

Over the last decade, corpus annotation with such information have been developed and used in different projects (Chinchor and Hirschman, 1997; Poesio, 2000; Vieira and Poesio, 2000). More recently, some projects are producing or making use of multilingual annotated corpora (Gundel et al., 1993; Harabagiu and Maiorano, 2000; Mitkov et al., 2000).

In this paper we present a multilingual corpus study on special types of nominal expressions and languages that have not been particularly addressed: definite descriptions and demonstrative noun phrases in a parallel corpus of Portuguese and French.

The paper is organized as follows. In section 2 we underline the need of studying nominal expressions in multilingual corpora. Section 3 presents the linguistic hypothesis guiding our studies of definite descriptions and demonstrative noun phrases. As our studies are also connected to systems evaluation, in section 4, we discuss some proposed systems for noun phrases interpretation to introduce the kind of computational work that could be evaluated for different languages on the basis of our corpus studies. In section 5, we describe the corpus, choice of annotation tools, and the annotation task. The corpus analysis comparing definite descriptions and demonstrative noun phrases in French and Portuguese languages is presented in section 6. Finally, we summarise our conclusions in section 7.

2. Motivations for studying nominal expressions in multilingual corpora

2.1. Validating linguistic hypotheses

From a linguistic point of view, multilingual corpora containing information about the linguistic realisation of nominal expressions and the relations between them play a fundamental role in validating models of coreference, anaphora and reference resolution.

If reference is considered as a function assigning, within a given context, an extra-linguistic entity (a referent) to a linguistic expression, then coreference can be defined as the relation between linguistic expressions referring to the same extra-linguistic entity (van Deemter and Kibble, 2000). A slightly different topic is anaphora resolution, concerned with the interpretation of expressions depending in various ways on other expressions within the same discourse. Here, 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).

Most of the traditional descriptive linguistic framework concerned with reference, coreference and anaphora (Corblin, 1997; Hawkins, 1978) is indeed based on observation of (sometimes construed) examples and intuition. However, in order to get a precise idea about the suitability of those theories for designing NLP tools and their coverage of phenomena that have to be treated in real corpora, they have to be evaluated on larger data. Moreover, to do this evaluation not only on monolingual, but also on multilingual corpora would open various interesting perspectives which we think are still open issues for linguistic theories as well as for the NLP community.

One of these perspectives is evaluating the generality and thereby cognitive plausibility of linguistic theories. The multilingual corpus study presented by Gundel et al. (1993) illustrates this point on a model of reference, showing how different cognitive statuses are related to different types of noun phrase realisation in English, Chinese, Japanese, Russian and Spanish. However, since

Gundel et al. (1993) didn't address the comparison of multilingual coreference chains, we think that a second perspective could be the question whether different languages use different features (e.g. use of specific anaphoric markers such as pronouns or definite descriptions) for realizing text cohesion within those chains. As a third perspective, in the context of anaphora resolution, multilingual corpora will be useful for studying the cross-language nature of other relations than coreference, especially bridging relations.

2.2. Design and evaluation of NLP tools

Natural language processing tools able to treat correctly coreferential or anaphoric noun phrases are relevant to a variety of applications, including information extraction and retrieval, automatic topic detection, terminology extraction, text summarization, natural language generation, machine translation and human-machine dialogue.

The development and evaluation of such tools are closely related to availability of corpora annotated with coreferential or anaphoric links. For the development phase, such corpora have been used for statistical approaches (Burger and Connolly, 1992), training of machine learning algorithms (Aone and Bennet, 1995) and rule-based systems (Vieira and Poesio, 2000). At the evaluation step, following the experience of the MUC coreference task¹ and measures, corpora became also crucial for testing the performance and robustness of the systems.

However, the availability as well as the coverage of such corpora is still a problem: whereas the size of annotated data for pronominal expressions and/or coreference (identity) relations only is slowly increasing, there are only a few data for other nominal expressions, such as definite descriptions and demonstrative noun phrases and other relations than identity (Garside et al., 1997; Poesio and Vieira, 1998; Poesio, 2000). This picture looks still worse for other languages than English, and multilingual corpora are still an important outstanding issue: the only parallel corpora annotated with anaphora we are aware of are Harabagiu and Maiorano (2000) for English and Romanian, Mitkov et al. (2000) corpus for English and French (10.000 words) and Vieira et al. (2002) for French and Portuguese (5000 words).

However, the need of such corpora is obvious for developing and testing the increasing number of multilingual tools for coreference and anaphora resolution (Azzam et al., 1998; Mitkov et al., 2000). Harabagiu and Maiorano (2000) and Mitkov and Barbu (2000) showed indeed that the performance of monolingual systems can be increased by "mutual enhancement strategies".

2.3. Our motivations

The multilingual corpus study presented in this paper aims to bring together the two previous research topics.

Concerning the issue of validating linguistic theories on reference, we are investigating here the hypothesis that different types of linguistic markers (referring noun phrases) have different functions, i.e. different coreferential or anaphoric behavior. Since much previous

work focus on pronominal anaphora, mainly in English, we choose to work on different languages (French and Portuguese) and on a different types of expressions (definite descriptions and demonstratives).

Regarding the second issue ? design and evaluation of NLP tools ? the main motivation for our multilingual corpus study of definite and demonstrative descriptions is indeed the development of a multilingual tool able to resolve reference for demonstratives and definite descriptions. Starting from a rule-based tool developed for English definite descriptions (Vieira and Poesio, 2000), we are investigating to which extend the basic principles implemented in this tool are suitable for other languages (French and Portuguese) and other types of expressions (demonstrative noun phrases).

As a "by-product", we intend to produce and make available multilingual corpora annotated with coreferential and anaphoric links for noun phrases for other languages than English. In order to make the corpus easier reusable for other research purposes, we spend particular attention to use annotation principles and tools that are compatible with recent work towards coding standards for linguistic resources (Isard et al., 2000; Poesio, 2000).

3. Linguistic Hypotheses

The basic hypothesis we are investigating here is the assumption that different types of linguistic markers, in our case definite and demonstrative noun phrases, have different linguistic functions and are therefore used in different contexts. Whereas the use of both types of noun phrases in discourse has often been considered under the general phenomenon of "anaphora", we expect them to have specific coreferential or anaphoric properties.

3.1. Definite noun phrases

As opposed to the intuition that definite noun phrases are primarily anaphoric, i.e. refer to familiar entities previously introduced into the discourse (Bosch and Geurts, 1990), several linguistic studies on various languages (English, French, Swedish, Portuguese) tend to reconsider this position in favor of the uniqueness theory, by enlarging the interpretational context with larger situations or domains (Fraurud, 1990; Poesio and Vieira, 1998; Rossi et al., 2001).

Indeed, according to Hawkins (1978), the (English) definite article may be used on the basis of a discourse antecedent (anaphoric and associative uses) as well as independently from the previous discourse (situational and unfamiliar with explanatory modifiers). For French, the descriptive linguistic framework on determiners proposed by Corblin (1987) shows that the different uses of definite descriptions (generic, associative and anaphoric) cannot be subsumed under a primarily anaphoric function. The fundamental interpretation principle of definites has therefore been redefined as being the identification of a unique referent within a given domain, based on the lexical content of the (eventually modified) head noun. This idea is also presented in (Löbner, 1986), definite descriptions are said to be "functional concepts" that take explicit (linguistic) or implicit arguments (from situational context).

Additionally, a purely anaphoric account of definites is unable to predict the distribution of definite noun phrases in corpora. Fraurud (1990) presented a corpus

¹http://www.itl.nist.gov/iaui/894.02/related_projects/muc/proceedings/muc_7_toc.html#coreference

study of definite noun phrases of Swedish texts and found that 60% of them were not used anaphorically (first mentions). Poesio and Vieira (1998) confirmed this result for English texts where about 50% of definites were discourse new. In Rossi et al (2001), similar numbers were found for Portuguese.

Based on these findings, the hypothesis we adopted for our multilingual corpus study is the following: definite noun phrases identify their referent on the basis of semantic information, but not necessarily within the previous discourse ? they are not primarily anaphoric.

3.2. Demonstrative noun phrases

Our hypothesis for the interpretational specificity of demonstrative noun phrases is based on descriptive linguistics for French (Corblin, 1987). As opposed to definite noun phrases identifying their referent based on semantic content, 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 (the only source of salience). Demonstratives should therefore be used in a more anaphoric way than definites. Moreover, since Corblin (1987) assumes the referent of demonstratives to be already salient, we hypothesize a clear preference for coreferential demonstratives (identity between antecedent and demonstrative) over other types of anaphoric relations (associative use).

An additional interesting hypothesis ? developed for French demonstratives² ? is the idea that the specific function of demonstratives is to bring new information about the referent. This idea follows from the previous point: since a demonstrative expression is said to identify the referent based on salience and independently of the semantic content of the noun, this semantic content is available for something else, i.e. give new information about the referent, for example by reclassifying it, as in the following example:

- | | | |
|----|--------------------------------------------------------------------|----------------------------|
| a. | fabricants de plates-formes
levantes destinées au | <i>elevating platforms</i> |
| d. | à ce que ces produits soient
conçus | <i>these products</i> |

In summary, the hypothesis we adopted for the interpretation of demonstrative noun phrases in our corpus study is the following: demonstrative identify their referent on the base of salience. From our type of material (written texts) follows that they should be necessarily related to previous discourse? therefore they are expected to be primarily anaphoric and preferentially coreferential.

4. Computational processing of noun phrases

As opposed to linguistic studies and hypotheses such as presented in the previous section, the most common approach for implementing systems for processing noun

phrases is still to consider a general class of anaphoric expressions. Moreover, much work centers on pronominal reference in English (Lappin and Leass, 1994; Grosz et al., 1995). As a result, the treatment of other types of expressions is typically seen as an extension of or variation on the basic coreferential mechanism involved in pronominal reference. Such an approach, however, does not predict essential differences between the use of pronouns, definite descriptions and demonstratives in contexts where human users would have clear preferences.

We present briefly some systems and models. Sidner (1978) main contribution is a theory of focus and its role in resolving definite noun phrases. The problem for the implementation of her models is the dependency on a knowledge network and associated inference mechanism. Although she acknowledges the occurrence of non-anaphoric definite noun phrases, her emphasis is on anaphoric relations and associations. Carter's system (Carter, 1987) implements a modified version of Sidner's algorithm. He proposes a shallow processing anaphor resolver in which reasoning is minimally considered. Again, definite descriptions are just one type of anaphoric expression among several dealt with his system.

The Core Language Engine (CLE) (Alshawi, 1990) is a domain independent system that translates English sentences into formal representations. Referential readings of definite descriptions are handled by proposing referents from the external application context as well as the CLE context model. Attributive readings may also be proposed, in this case, the identification of an external or contextual referent is not necessary for the resolution. Both referential reading resolution with the external application context and the attributive reading seem to account for discourse new descriptions.

Statistical approaches have being also tested for the problem. In Burger and Connolly (1992), a Bayesian network is used to resolve anaphora by probabilistically combining linguistic evidence. Their sources of evidences are: c-command (syntactic constraints), semantic agreement, discourse focus, discourse structure, recency, centering. Aone and Bennet (1995) propose an automatically trainable anaphora resolution system. They train a decision tree using the C 4.5 algorithm by feeding feature vectors for pairs of anaphor and antecedent. They use 66 features, including lexical, syntactic, semantic, and positional features.

As most of these systems deal with anaphora resolution in general or are based on knowledge representation and inference, quantitative evaluation for the particular cases of definite descriptions and demonstratives is not available. The system proposed in Vieira and Poesio (2000), however, is specially concerned with definite descriptions. As the system deals mainly with descriptions which are not dependent on inter-sentential lexical inference or reasoning, heuristics are proposed to identify: 1) antecedents with same head noun and 2) descriptions that are not based on textual antecedents. These latter are identified on the basis of their syntactic structures (noun phrases that provide arguments for the conceptual function). These heuristics, however, were only evaluated for English.

In our work we are interested in the performance of these heuristics to other languages than English. In Rossi et al (2001) these heuristics were first applied to

² We are not aware of work investigating this hypothesis for other languages.

Portuguese language. Also for the Portuguese language, Sant’Anna and Lima (2001) have adapted the heuristics for English definite descriptions to a system dealing with demonstrative noun phrases.

By developing the multilingual studies presented here we are setting the basis to evaluate these heuristics in a parallel corpus of Portuguese and French.

5. Corpus annotation

5.1. Corpus

As a working corpus for studying definite and demonstrative descriptions, we choose 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.

We undertook two corpus annotation experiments related to the use of definite descriptions and demonstrative. For a study of approximately 500 definite descriptions we needed a corpus of 5000 words, for the study of demonstratives we needed a corpus ten times larger in order to have approximately 250 demonstratives. Tables 1 and 2 give an overview of the resources we used.

Corpus	Language	Size (in words)	Number of Definites
MLCC	French	~ 5000	461
	Portuguese		541

Table 1: Corpus for the study of definite descriptions

Corpus	Language	Size (in words)	Number of Demonstratives
MLCC	French	~ 50000	291
	Portuguese		243

Table 2: Corpus for the study of demonstrative NPs

5.2. Annotation task

Our aim was to analyze the use of definite and demonstrative descriptions in order to verify whether the hypotheses presented in section 3 could be validated for French and Portuguese. Therefore, we ran two annotation experiments, a first one for definites, a second one for demonstratives.

As a first annotation step, all definite and demonstrative nouns phrases have been marked up. For definites, we had approximately 500 definite descriptions over the first 10 question-answer pairs. For demonstratives, we had approximately 250 cases over 90 question answer pairs. Regarding prepositional and relative clause attachment, we choose the “maximal” option, including them systematically into the noun phrase.

For our first experiment, the annotation task, done by two other annotators, was the classification of definite descriptions in one of four classes and identification of antecedents when appropriate. We proposed classification

schemes inspired by linguistic work. They are presented more in detail in section 6.2.1.

We started our experiments with annotating definite descriptions. As a consequence of our first results (low inter-annotator agreement), we decided to change slightly the annotation process for demonstratives. We intended to make the task easier for the annotators by avoiding too many different decisions to be taken at the same time (additional problems were due to the manipulation of the annotation tool, see section 5.3). Then, for the annotation of demonstratives, we separated clearly the antecedent finding task from the classification task, whereas these two tasks have been done in one (complex) step for definites.

For our second experiment, with demonstratives, the annotation task was the mark-up of antecedents for the noun phrases. By letting them do so, we allowed antecedents of any type and size (head nouns, noun phrases, other chunks of text). However, we imposed as a maximal limit a sentence. It means that antecedents greater than a sentence (a few cases for demonstrative noun phrases) were not marked in this experiment. The last annotation step, for the second experiment, was then to classify the relation between the antecedent and the noun phrase.

5.3. Annotation tool

Building corpora that will be useful for a broader community implies to fulfill various requirements on the architecture and the annotation scheme.

Concerning the architecture, the principle of a stand-off-annotation seems to be more and more supported. The main advantages of such an annotation are support of different annotation levels without mixing them up and support of alternative annotations of the same coding level, thereby making easier the evaluation of inter-annotator agreement (Carletta, 1996). The stand-off principle is also suitable for multilingual corpora, allowing to externalize the information about aligned text chunks.

Among several free annotation tools for coreference or anaphora in monolingual corpora, we found two of them suitable from a conceptual point of view ? the MATE workbench (Isard et al., 2000) and MMAX (Müller and Strube, 2001). From a practical point of view however, only MMAX seems to be at the same time a performant and customizable annotation tool. Furthermore, it would be easier to extend it in order to fit the requirements related specifically to parallel corpora (see section 7). For these reasons, we chose the MMAX tool for our experiment.

6. Comparative results for definites and demonstratives

6.1. Syntactic structure and determination

After the first annotation step ? mark up of definite and demonstrative noun phrases ? we classified the resulting markables depending on their internal syntactic structure. This has been done in order to ask whether there could be established a relation between preferences for specific syntactic structures for definites and demonstratives in written corpora and different referential functions, as hypothesized in section 3.

6.1.1. Internal syntactic structure of noun phrases

For the syntactic classification of French and Portuguese definite and demonstrative noun phrases, we choose the classes summarized in table 3. The main features distinguishing these classes are presence or lack of (adjectival, prepositional and relative) modifiers and the type of the modifier, if present.

class	Syntactic structure	Examples (Fr/Pt)
1	Det N	<i>cette région</i> <i>esta região</i>
2	Det (adj N N adj)	<i>ces pratiques abusives</i> <i>estas práticas abusivas</i>
3	Det N of N	<i>ce parc d' éoliennes</i> <i>esta ajuda de emergência</i>
4	Det N rel_pro	<i>ces oiseaux que la loi protège</i> <i>este cidadão que a lei protege</i>
5	Det (adj N N adj) of N	<i>ces usages vulnérables de la route</i> <i>esta sociedade gestora de participações sociais</i>
6	Det (adj N N adj) rel_pro	<i>ce grave problème social dont souffre l' Achaïe</i> <i>este grave problema social que sofrem os cidadãos</i>
7	Other	<i>ce domaine clé</i> <i>motivos eses</i>

Table 3: Internal syntactic structure of noun phrases

Class 1 stands for noun phrases containing only a head noun without modifiers (and includes a few cases of Portuguese or French elliptical noun phrases such as *le dernier / o último – the latter one*). Class 2 contains noun phrases with adjectival modifiers; class 3 contains noun phrases with a prepositional complex introduced by the preposition *of*; class 4 contains nouns phrases followed by relative clauses; class 5 and 6 are complex noun phrases combining features of class 2/3 and of class 2/4, respectively. Tables 4 and 5 give an overview about the distribution (in percentage) of French and Portuguese definite and demonstratives over our 7 syntactic classes.

Class	1	2	3	4	5	6	7
French	35,4	22,6	24,3	1,5	6,5	0,75	8,9
Portuguese	40,8	22,7	25,5	1,8	3,2	0,5	5,3

Table 4: Syntactic structure of definite noun phrases

Class	1	2	3	4	5	6	7
French	80,4	10,3	6,2	0,3	1,0	0,7	1,0
Portuguese	80,2	7,4	6,5	0,4	0,8	0,4	4,1

Table 5: Syntactic structure of demonstrative noun phrases

6.1.2. Discussion

As the most important result, we noticed the difference between definites and demonstratives regarding the proportion of noun phrases belonging to class 1 (head noun without modifiers). Whereas this proportion is about 37% for definites, it is about 80% for demonstratives, in the two languages.

One possibility of establishing a relation between this result and our hypothesis about the differences in the referential behavior of definite descriptions and demonstratives could be the following: In section 3, we expected definite descriptions to be interpreted on the basis of semantic information, but not necessarily anaphorically to entities introduced within the previous discourse. If one considers that the quantity of semantic information increases with the adjunction of modifiers, then the fact that they belong mainly to classes 2 to 7 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. If such a correlation between syntactical complexity and textual antecedent independency could be confirmed based on the classification results, it would be another argument in favor of the assumption that definite descriptions are not primarily anaphoric expressions (see section 6.2.3).

Regarding demonstratives, in French as well as in Portuguese, we have few modified demonstrative NPs (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.

6.2. Classification and determination

In order to evaluate, on the one hand, the hypothesis about the anaphoric character or not of definites and demonstratives and to measure, on the one hand, the agreement which could be reached by human annotators on this subject, we asked two annotators per language to find an antecedent for each definite and demonstrative expression and to classify the relation holding between these expressions and their antecedents.

6.2.1. Categories of relations

The relations between definite or demonstrative expressions and their textual antecedents have been defined depending on different categories of use. In the linguistic literature there is a large variety of classifications of definite description uses: they vary from coarse-grained classifications such as first and subsequent mentions (Fraurud, 1990) to fine-grained classifications such as Strand's taxonomy of linking relations (Strand, 1996). Our classes are mainly inspired by work of Hawkins (1978), Fraurud (1990), Prince (1981, 1992) and Poesio and Vieira (1998). For each definite or demonstrative description *d*, we proposed one of the following four classes:

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

- | | | |
|----|-----------------------------------------|------------------------------|
| a. | tem conhecimento do livro | <i>of the book</i> |
| d. | que o livro não se debruça sobre | <i>the book</i> |
| a. | e prestar às autoridades gregas | <i>the greek authorities</i> |
| d. | para essas autoridades | <i>these authorities</i> |

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

- a. A circulação **dos cidadãos** que *the citizens*
- d. controle **das pessoas** nas fronteiras *the people*

- a. À **Albânia** *Albania*
- d. ajudar **este país** a atingir *this country*

Other kind of anaphora: the interpretation of *d* depends on a previous expression *a*, but either *d* does not corefer with *a* or *a* is not a nominal expression:

- a. O **recrutamento** de pessoal *the recruitment*
- d. **as condições de acesso à carreira científica** *the conditions of employment for scientific jobs*

- a. **foi várias vezes condenada a má aplicação desta Convenção na CEE** *.was several times condemned to wrong application of this convention in the CEE*
- d. Salientou-se que **esta situação** originava um *this situation*

Discourse new: the interpretation of *d* does not depend on any previous expression:

- d. sobre **a actividade das várias organizações internacionais** *the activity of various international organisation*

6.2.2. Classification results

Tables 6 to 9 show the results for the classification of French definites (6), Portuguese definites (7), French demonstratives (8) and Portuguese demonstratives (9).

Category	Annotator1	Annotator2	Average
Direct coref.	132	96	24,7%
Indirect coref.	23	27	5,4%
Other anaphora	63	26	9,7%
Discourse new	216	241	49,6%
Not classified	27	71	10,6%
Total	461	461	100,0%

Table 6: Classification of French definites

Category	Annotator1	Annotator2	Average
Direct coref.	96	179	25,4%
Indirect coref.	51	45	8,9%
Other anaphora	46	77	11,4%
Discourse new	266	198	42,9%
Not classified	82	42	11,5%
Total	541	541	100,0%

Table 7: Classification of Portuguese definites

For the first experiment on definites, they give, for each annotator, the distribution of definite descriptions over the four categories (direct coreference, indirect coreference, other type of anaphora, discourse new) as well as the number of descriptions which have not been classified. For the second experiment on demonstratives, they give at a first level the distribution over three classes (direct coreference, indirect coreference, other). The "other" class includes other kind of anaphora such as

defined above and cases where demonstratives were indeed anaphora, but not classified as such because the antecedent was not marked up (greater than a sentence or split over several sentences). It contains also few cases which have been classified as discourse new by one annotator and as anaphoric by the other one.

Category	Annotator1	Annotator2	Average	
Direct coref.	102	95	33,8%	
Indirect coref.	60	45	18,0%	
Other	Anaphora	118	111	39,3%
	Disc. New	0	7	1,2%
	Not marked	11	33	7,6%
Total	291	291	100,0%	

Table 8: Classification of French demonstratives

Category	Annotator1	Annotator2	Average	
Direct coref.	80	74	31,7%	
Indirect coref.	60	49	22,4%	
Other	Anaphora	77	66	29,4%
	Disc. New	0	0	0%
	Not marked	26	54	16,5%
Total	243	243	100,0%	

Table 9: Classification of Portuguese demonstratives

6.2.3. Discussion of our linguistic hypotheses

The first reason for doing our classification experiments was to evaluate the hypotheses about the anaphoric character or not of definites and demonstratives.

Regarding definite descriptions, we hypothesized them to be not primarily anaphoric. Rather, they were expected to identify a referent on the basis of semantic information, but not necessarily within the previous discourse. Demonstrative noun phrases were seen as identifying their referent on the base of salience. Given our material (written texts), we expected them to be necessarily related to previous discourse, and to be primarily anaphoric and preferentially coreferential. Our classification results do support these hypotheses.

Regarding the distribution of definite descriptions over the four classes, we found a great number of descriptions classified as discourse new. For the two languages, this number ? over 40% ? is closed to the 50% threshold observed in previous annotation experiments carried out on English by Poesio and Vieira (1998). This observation not only confirms the hypothesis that definite descriptions are not primarily anaphoric (Fraurud, 1990, Poesio and Vieira, 1998), but brings also new information about a possible cross-language and cross-genre stability of this feature.

Additionally to these global results, in order to get an idea about the distribution of syntactically complex definite noun phrases (see section 6.1.2) over the different categories, we report in table 10 a study carried out for one French annotator. Table 10 tends to confirm the hypothesis of a correlation between syntactic structure and context (in)dependency of definite noun phrases: whereas the most important proportion (over 50%) of simple definite noun phrases (category 1: head noun only) are directly coreferential with their antecedents and therefore context dependent, the majority (over 50%) of

syntactically complex definites (categories 2 to 7) are discourse new descriptions, i.e. context independent. The great number of syntactically complex definite noun phrases found in section 6.1.2 can therefore be taken as an indicator for context independency accordingly with the hypothesis that definite descriptions are not primarily anaphoric.

Syntactic structure	Direct coref	Ind. coref	Other anaph.	Disc. New	Not class.	Total
1	51,0	3,5	9,6	29,3	6,6	100,0
2-7	16,4	6,8	18,3	51,7	6,8	100,0

Table 10: Syntactic structure and anaphoric relations of French definites for annotator 1

Regarding our initial hypothesis on demonstratives ? preference for context dependency and anaphoric behavior ? the results in tables 8 and 9 show clearly that only 13% of them (9% for French and 17% for Portuguese) are not related to a discourse antecedent, i.e. not context dependant. The remaining demonstratives are indeed anaphoric in a broader sense (context dependent), with 50% of them being coreferential with previous noun phrases. As for definite descriptions, the linguistic hypothesis seems to be confirmed for our French and Portuguese corpora.

6.2.4. Evaluation of agreement

In order to evaluate the inter-annotator agreement on the classification task, we calculated the Kappa (Carletta, 1996) for each experiment. This measure establishes $K = 0.8$ as good agreement.

For the experiment on definites, Kappa was calculated for four classes (direct coreference, indirect coreference, other anaphora, discourse new). For the French definites, we found $K = 0.52$ and for the Portuguese definites $K = 0.48$. For our experiment on demonstratives, we calculated Kappa for only three classes (direct coreference, indirect coreference, other). Here, we found $K = 0.79$ for French and $K = 0.65$ for Portuguese demonstratives. Globally, this means that it is impossible to take the classification task such as carried out here as a key task for evaluating systems of processing definite descriptions. For demonstratives however, the results allow tentative conclusions.

The important difference for Kappa on definites and on demonstratives has different reasons. First, given the low agreement for the first experiment on definite descriptions, we adopted another annotation order with a clearer task separation for the experiment on demonstratives (see 5.2). Second, we calculated the Kappa for demonstratives on only three classes, merging "other kind of anaphora" with "not classified" and the few cases of disagreement where one annotators considered the demonstrative as "discourse new". Finally, informal feedback from the annotators suggests that the annotation task was inherently easier for demonstratives than for definites. This is an interesting point, since it goes into the direction of our initial hypothesis about the specific referential behavior of definites and indefinites: if demonstratives are more likely to be used as discourse anaphora, then it seems plausible that the annotation task was easier for human annotators.

7. Conclusions and perspectives

Linguistic work as well as previous corpus studies have shown that definite descriptions are commonly used to introduce new discourse elements. Our studies extend these findings for two other languages and for a different corpus. We also compared the use of definite descriptions to another type of noun phrases, commonly considered as anaphorically: demonstrative noun phrases. As opposite to definite descriptions, they presented a distribution of types of use that shows predominance of anaphoric use. These findings confirm the linguistic hypotheses raised in this paper, in which one of the main issues is the reliability on previous discourse for the interpretation of these nominal expressions: definite descriptions are not mainly anaphoric expressions, demonstrative noun phrases are.

Whereas the agreement on classifying demonstratives allows at least tentative conclusions, the low agreement for definites does not allow us to take the annotated corpora as it is as a key for designing and evaluating tools for processing definite descriptions. To decide exactly what such a system should do, we plan new annotation experiments, starting with two classes only and to refine the definition of not coreferential anaphora, possibly by restricting it first to well-defined relations such as part-whole or set-subset relations.

In order to avoid disagreement related to the use of the annotation tool, we would also like to propose improvements to the MMAX interface, especially by analyzing the user feedback we collected from the annotators. Another change we propose for the tool is making it able to handle parallel corpora. The optimal data architecture for such a direct parallel annotation would take as an input three XML files: word files (list of words for each language), text structure files (defining which parts of the texts are titles, paragraphs, sentences etc.) and an alignment file, defining which text chunks are aligned. These three files would be referred to by an XSL style sheet, used for building a user-friendly graphical annotation interface. During the annotation process, three XML output files would be generated: two of them keep track of the annotated expressions for each language, the third one keeps track of relations between expressions of different languages referring to the same referent.

Such a parallel annotation will allow us to study the distribution of nominal expressions in parallel corpora in order to evaluate how consistently definites and demonstratives are used in different languages, whether there are invariables concerning the linguistic realisation of other anaphora than coreference and to which extent parallel resources annotated with anaphoric relations for definites and demonstratives can be useful for enhancing current heuristics of anaphora resolution.

8. References

- Alshawi, H., 1990. Resolving Quasi Logical Forms. *Computational Linguistics*, 16(3):133-144.
- Aone, C. and Bennet, S. W., 1995. Automated acquisition of anaphora resolution strategies. *Proc. AAAI Spring Symposium on Empirical Methods in Discourse Interpretation and Generation*, 1-7, Stanford.
- Appelt, D., et al. 1995. SRI International FASTUS system MUC-6 test results and analysis. *Proc. of the 6th*

- Message Understanding Conference*, 237-248. Columbia.
- Ariel, M., 1990. *Accessing Noun-Phrase Antecedents*. Routledge, London and New York.
- Azzam, S., Humphreys, K. and Gaizauskas, R. 1998. Coreference Resolution in a Multilingual Information Extraction System. In *Proc. of the 1st Language Resources and Evaluation Conference (LREC)*, Linguistic Coreference Workshop.
- Bosch, P. and Geurts, B., 1990. Processing definite NPs. *Rivista di Linguistica*, 2(1):178-199.
- Burger, J. D. and Connolly, D., 1992. Probabilistic resolution of anaphoric reference. *Proc. AAAI Fall Symposium on Probabilistic Approaches to Natural Language*, 17-24, Cambridge.
- Byron, D. and Allen, J., 1998. Resolving demonstrative anaphora in the TRAINS93 corpus. *Proc. Discourse, Anaphora and Reference Resolution Colloquium*. Lancaster.
- Carbonell, J.G. and Brown, R.D., 1988. Anaphora Resolution: a Multi-Strategy Approach. *Proc of the 12th International Joint Conference on Computational Linguistics*, 96-101.
- Carletta, J., 1996. Assessing agreement on classification tasks: the Kappa statistic. *Computational Linguistics*, 22(2):249-254.
- Carter, D.M., 1987. *Interpreting Anaphors in Natural Language Texts*. Chichester, UK, Ellis Horwood.
- Chinchor, N. A. and Sundheim, B., 1995. Message Understanding Conference (MUC) tests of discourse processing. *Proc. AAAI Spring Symposium on Empirical Methods in Discourse Interpretation and Generation*, 21-26, Stanford.
- Chinchor, N. and Hirschman, L., 1997. MUC-7 Coreference Task definition, Version 3.0. *Proc. of the 7th Message Understanding Conference MUC-7*, 1997
- Clark, H. H., 1977. Bridging. In Johnson-Laird, P.N. and Wason P.C. (eds.) *Thinking: Readings in Cognitive Science*. London New York: Cambridge University Press.
- Corblin, F., 1987. *Indéfini, Défini et Démonstratif*. Droz, Genève.
- Fraurud, K. 1990. Definiteness and the Processing of Noun Phrases in Natural Discourse. *Journal of Semantics* (7) 395-433.
- Garside, R., Fligelstone, S. and Botley, S. 1997. Discourse annotation: anaphoric relations in corpora, In: Garside, R., Leech, G. and McEnery, A. *Corpus annotation: linguistic information from computer text corpora*, Longman, Londres-Nova Iorque, pp. 66-84.
- Grosz, B.J., Joshi, A.K. and Weinstein, S., 1995. Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics*, 12(2):203-225.
- Gundel, J., Hedberg, N.A. and Zacharski, R., 1993. Givenness, Implicature and the form of referring expressions in discourse. *Language*, 69(2):275-307.
- Harabagiu, S. M. and Maiorano, S. J. 2000. Multilingual Coreference Resolution, in *Proc. of the Language Technology Joint Conference on Applied Natural Language Processing ANLP-NAACL2000*, Seattle WA.
- Hawkins, J.A. 1978. *Definiteness and Indefiniteness: a study in reference and grammaticality prediction*. London: Croom Helm.
- Isard, A., McKelvie, D., Mengel, A., Moller, B. M., 2000. The MATE Workbench Annotation Tool, a Technical Description. *Proc. of LREC 2000*, Athens.
- Lappin, S. and Leass, H., 1994. An algorithm for pronominal anaphora resolution. *Computational Linguistics*, 20(4) 535-561.
- Löbner, S., 1985. Definites. *Journal of Semantics*, 4:279-326.
- Mitkov, R. and Barbu, C. 2000. Mutual enhancement of performance: bilingual pronoun resolution for English and French. *Proc. Discourse, Anaphora and Reference Resolution Conference*. Lancaster.
- Mitkov, R. Evans, R., Orasan, C., Barbu, C. Jones, L. Sotirova, V. 2000. Coreference and anaphora: developing annotating tools, annotated resources and annotation strategies. *Proc. Discourse, Anaphora and Reference Resolution Conference*. Lancaster.
- Müller C., Strube M., 2001. Annotating anaphoric and bridging relation with MMAX. *Proc. of the 2nd IJCAI Workshop on Knowledge and Reasoning in Practical Dialogue Systems*, 45-50, Seattle, 5 Aug.
- Poesio, M. 2000. Coreference. MATE Dialogue Annotation Guidelines Deliverable D2.1, 126-182. (www.ims.uni-stuttgart.de/projekte/mate/mdag/)
- Poesio, M., and Vieira, R., 1998. A corpus-based investigation of Definite Description Use. *Computational Linguistics*, 24(2):183-216.
- 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.
- Rossi, D., Pinheiro, C., Feier, N. and Vieira, R., 2001. Resolução de correferência em textos da língua portuguesa. *REIC*. 1(2) (www.sbc.org.br/reic).
- Sant'Anna, V. and Lima, V. L. S., 2001. Cálculo de referências anafóricas pronominais demonstrativas na língua portuguesa escrita. In *Anais do XXI Congresso da Sociedade Brasileira de Computação*, Fortaleza-CE.
- Sidner, C. L., 1978. The use of focus as a tool for disambiguation of definite noun phrases. In Waltz, D. L. (Ed.), *Theoretical issues in natural language processing -2*. University of Illinois at Urbana-Champaign, 25-27, July.
- Strand, K., 1996. A Taxonomy of Linking Relations. *IndiAna Workshop*, Lancaster, July 19th 1996.
- van Deemter, K., Kibble, R., 2000. On Coreferring: Coreference in MUC and related annotation schemes. *Computational Linguistics* 26(4).
- Vieira, R., and Teufel, S., 1997. Towards Resolution of Bridging Descriptions. *Proc of the 35th International Joint Conference on Computational Linguistics*, Madrid, July, 7-12.
- 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. and Schang, E., 2002. Multilingual Corpora Annotation for Processing Definite Descriptions, *Proc. Portugal for Natural Language Processing PorTAL*, Universidade do Algarve, Faro, Portugal, June 23- 26.