
Coreferential Definite and Demonstrative Descriptions in French: A Corpus Study for Text Generation

HÉLÈNE MANUÉLIAN

LORIA, Nancy, France

helene.manuelian@loria.fr

ABSTRACT. This paper presents a new classification for the use of definite and demonstrative descriptions, its application in a corpus analysis and the results of this analysis. The proposed classification is based on existing literature and extended to support the generation of definite and demonstrative NPs. The corpus analysis shows in particular, that subsequent mentions of a referent can perform two functions (repeating given information and/or introducing new information). The comparison between definite and demonstrative determiners leads to preliminary data for generation algorithms.

1 Introduction

Algorithms for the generation of referring expressions [Dale et Reiter, 1995, Van Deemter, 2001, Van Deemter, 2000, Krahmer et al., 2001] essentially generate definite descriptions. The purpose of these algorithms is to produce given a referent that is already present in the context a referring expression that is informative enough for the identification of the intended referent by the listener. In a different perspective, [Gardent, Striegnitz, 2000] present an algorithm for the generation of bridging descriptions based on a structured model of the context and a tight interleaving between inference and generation. All these algorithms generate definite descriptions referring to objects already present in the context and sometimes handle the definite / pronoun opposition. However, they never handle the distinction between definite and demonstrative descriptions. A good way to extend these algorithms could be to introduce the distinction between the definite and the demonstrative. Both determine anaphoric noun phrases, so we need to know how to choose between them [Kleiber, 1986, Kleiber, 1988, Corblin, 1987].

This paper is a first step in this direction: we propose to explore this distinction through a corpus study of the anaphoric uses of definite and demonstrative. We show that the literature about French definite and demonstra-

tive is not precise enough to handle the generation of coreferential expressions. We further present a corpus study based on this literature from which we derived a new classification of determiner use, taking into account important data for generation which is not mentioned in the literature namely, the global content of the noun phrase. We give in conclusion some directions for the generation of referring expressions.

2 A First Corpus Study

The corpus studied is a part of the PAROLE corpus¹ which contains 65 000 words, 8777 definite noun phrases and 555 demonstrative noun phrases. The corpus is composed of articles from the French newspaper *Le Monde* which are taken from every section (national and international politics, economics, culture, sports, and leisure). It is annotated at the morphosyntactic level according to the Multext / Multitag annotation scheme [Lecomte, 1997, Beaumont et al. 1998].

Our goal was to maximally automatize the corpus processing, so we applied much preprocessing before performing the annotation described in this section. We used the G-search tool [Corley et al., 2001] to identify the noun phrases in the corpus and we wrote several filters to adapt the resulting output to the format used by our annotation tool, MMAX [Muller, Strube, 2001]. The annotation with MMAX is completely manual.

2.1 Annotation Scheme

The first distinction between demonstrative and definite is linked with the process of referent identification. The definite determines an expression whose referent is unique in the context with respect to the description contained in the noun phrase. The demonstrative NP denotes a referent which is highly focused, and the semantic content of the description is not used to identify the intended referent [Kleiber, 1986, Kleiber, 1988, Corblin, 1987]. The literature identifies three main uses for both determiners: first mention, anaphoric mention or bridging. Our annotation identifies all these uses for both determiners.

First mention This is the case when the referent has not been mentioned before in the context and cannot be inferred from any antecedent. With the definite, the referent must be uniquely identifiable in the context. The demonstrative must be used with a gesture ("deictic use" of the demonstrative).

¹This corpus is shared with the ATILF research unit (Analyse et Traitement Informatique de la Langue Française) in the context of the regional collaboration "CPER Intelligence Logicielle".

Anaphoric Noun Phrases Both determiners can be used in coreference. The difference is explained by the process of referent identification. The demonstrative requires the referent to be highly focused whereas it is not the case with the definite. Three types of anaphoric uses are found and annotated in our corpus. For each anaphoric use, an antecedent has to be identified in the previous text.

Direct Coreference Both determiners can be used in direct coreference situations. This is the case when the phrase head noun is the same in the antecedent and in the anaphor.

Indirect Coreference This is the case when the head noun of the anaphoric phrase is different from the head noun of the antecedent. The indirect coreference is also found with both determiners. Several ways of realising coreference are found and subtyped in the annotation, but this is out of the scope of this paper (for more details, see [Manuélian, 2003]).

Bridging This category of referring expression is essentially used with the definite [Clark, 1977]. This is the case when the antecedent and the anaphor do not corefer, but the anaphor is interpreted as a part of the antecedent or as an object linked to the antecedent by the world knowledge. We can find cases of demonstrative in bridging descriptions but these are rather rare.

Attributes and appositions We classified in a separated category definite noun phrases which corefer to another noun phrase in apposition, or attributes coreferring with the subject of the verb to be, considering that it was a particular case of coreference, expressed explicitly, and using different mechanism from classical coreference.

2.2 Results and Analysis

The table (1.1) shows the annotation results for both determiners. They confirm the results of previous theoretic and empirical studies [Corblin, 1987, Kleiber, 1986, Kleiber, 1988, Poesio, Vieira 1998, Vieira et al. 2002].

We can see clearly that the definite can be used in first mention cases, which is less the case for demonstrative. The proportion of first mention for the definite is very high compared with other empirical studies [Poesio, Vieira 1998], because it also contains the "containing inferrable" which are usually distinguished from other first mention uses (cases as *the light of the Sun*, that are uniquely identifiable in their first mention). Indeed, we can see that the most important use for demonstrative is the anaphoric use, which is not the case for definite (about 75% of the uses of the demonstrative are - directly or indirectly - coreferential against 16% for the definite).

We can also find an illustration of the reclassification power of the demonstrative if we look at only the coreferential uses: 312 cases of coreferential

uses of the demonstrative are indirect, which represent 77% of its coreferential uses (direct and indirect). 819 cases of indirect coreferential uses of the definite are found, which represents 57% of its coreferential use.

The bridging phenomena is not so important for both categories and we can confirm that it is more frequent with the definite than with the demonstrative (For more details see [Gardent et al. 2003]). This annotation nonetheless confirmed the existing results but also permitted the extraction of coreferential noun phrases (1400 definite and 400 demonstrative), and their more detailed study presented in the next section of this paper.

Relation	demonstrative number	demonstrative proportion	definite number	definite proportion
First Mention	113	20,36%	6893	78,53%
Association	9	1,62%	417	4,75%
Direct Coreference	94	16,94%	607	6,92%
Indirect Coreference	312	58,02%	819	9,33%
Appositions / attributes	17	3,06%	41	0,47%
TOTAL	555	100	8777	100

Figure 1.1: First Annotation Results

3 A New Classification of Determiner Use

3.1 Motivations

The classification used in our annotation is helpful if we want to study the linguistic uses of coreferential descriptions. The problem for generation is that this classification does not handle certain elements which are essential for the generation of anaphoric expressions. In particular, modifiers are not taken into account, and one of the problems for generation is to decide the semantic content of a complete description, not only the semantic content of the head noun. Moreover, studying our data, we found that the information contained in the anaphoric noun phrases is not necessarily given information (i.e. explicitly or implicitly entailed by the context). Indeed, in some cases, we found that the anaphoric noun phrase introduces new information about the referent. Given these observations, questions that need to be addressed to support better generation are:

- What is the communicative function of the noun phrase to be generated: does it convey given information or does it introduce new information?
- If the information contained in the anaphoric noun phrase is given, does it always come from the antecedent, or from the context?
- If the information contained in the anaphoric noun phrase is new, which linguistic means are used to express it?

3.2 Classification and Examples

We first distinguished anaphoric expressions repeating information (Information Repeating Anaphors) from anaphoric expressions which add information (Information Adding Anaphors). The first category is divided into five subcategories, according to the source from which the information is taken (explicitly or inferred). The second category is divided into four classes according to the linguistic mean used to carry the new information. All the examples here are taken from the corpus.

Information Repeating Anaphors (IRA)

The information is given by the antecedent only (AO)

Example: *Celle-ci, (...) aurait tissé un réseau de **liens ambigus** dans la gendarmerie, la sûreté de l'Etat, les clubs de tir. Le procès (...) avait permis de mettre **ces liens** en relief.*

Translation: She would have established **some ambiguous links in the police, and clubs**. The trial brought **these links** further.

The information is given by the antecedent and the context (A+C)

Example: *Le patronat avait très sensiblement modifié **son comportement**. (...) La clé de **ce nouveau comportement** tient en deux chiffres.*

Translation: The employers modified strongly **their behaviour**. The reason for **this new behaviour** (...).

The given information is inferred from a lexical relation with the antecedent (LR)

Example: *L'Inde paie un tribut sans cesse plus lourd à **la sécheresse**, (...). **Ce phénomène** a été accentué par des choix économiques erronés.*

Translation: India suffers more and more of **the drought**. **This phenomenon** has been provoked by wrong economical choices.

The given information is inferred from a lexical relation with the antecedent and from the context (LR + C)

Example: *La municipalité s'est dotée récemment d'un **somptueux Palais des concerts**. C'est dans **ce bâtiment confortable et flambant neuf** qu'a eu lieu l'inauguration.*

Translation: The city council build recently **a beautiful Palace of concert**. The inauguration took place in **this comfortable and very new building**.

The given information is inferred by world knowledge from the antecedent and from the context (WKL)

Example: *Les journalistes ne feront pas de reportage sur **la visite de M. Honecker au cimetière de Neunkirchen, dans la Sarre, où sont enterrés ses parents**. (...) après que le chef d'Etat eut requis la "tranquillité"*

pour cette partie "privée" de son voyage en République fédérale.

Translation: The journalist won't make reports about **the visit of M. Honacker in the Neunkirchen graveyard where his parents are buried.** (...) after the head of state asked for quietness during **this private part of his travel in the Federal Republic.**

Information Adding Anaphors

The new information is introduced by a specifying lexical relation (SLR)

Example: *Ce document souligne (...) les conséquences médicales de la consommation de tabac, (...). Les auteurs de ce rapport (...).*

Translation: **This document** stresses the consequences of consuming tobacco. The authors of **this report** (...).

The new information comes from a specifying lexical relation and from modifiers: (SLR + mod)

Example: *Mais à Roubaix (...), le personnel a l'impression de compter les points. (...) Pour ces ouvrières du bassin houillier dont quelques-unes ont déjà trois heures de transport par jour, la nouvelle (...).*

Translation: But in Roubaix **the staff** has the feeling to count points. For **these workers from the coalfield who travel three hours a day** the news (...).

The new information comes from modifiers: (mod)

Example: *L'aviation israélienne a effectué (...) un raid sur le camp de réfugiés palestiniens d'Ain-el-Heloue, dans les faubourgs de Saida, chef-lieu du Liban-sud. Les chasseurs-bombardiers israéliens ont effectué (...) plusieurs attaques sur ce camp qui compte soixante-mille habitants, (...).*

Translation: The Israelian air force attacked **the palestinian refugee camp of Ain El Heloue in Saida suburb.** The israelians bombers made several dive attacks on **this camp which counts sixty thousand of inhabitants.**

The new information is in the whole phrase and no lexical relation is used: (No LR)

Example: (...) *je tombe sur un article intitulé " Pourquoi les maris prennent le large". Je me dis : cherche pas, ils se débinent pendant que tu t'échines à faire des pompes et des flexions, ces salauds-là.*

Translation: I find a paper which title is "Why **husbands** escape?". I say "They escape when you are practising sports, **these bastards.**"

3.3 Comparison between Definite Use and Demonstrative Use According to the New Classification

The two top tables of figure 1.2 show the source of the information used in IRA according to the categories defined in section (3.2), in relation to categories defined in the annotation scheme (section 2). For each category of annotation we distinguished in the table the modified anaphors by " + mod". The bottom tables in 1.2 synthetize the results by showing the number of demonstrative and definite IRA in regard to the necessity of relating them to their antecedent with inferences. We assume that it is not necessary to make inferences if the information contained in the anaphor is explicitly given in the antecedent. The tables (1.3) shows the linguistic means for adding information for IAA, according to the classification defined in section (3.2) in relation with the annotation categories. As for tables in (1.2), we indicate when the anaphoric noun phrase is modified.

We removed from these results the NPs which have non-nominal antecedents. From these two tables which classify 1412 coreferential definite NPs and 352 coreferential demonstrative NPs, we can say as a first result that for both determiners, the most frequent use is the information repeating anaphoric one (about 75% of the noun phrases belong to the category of IRA). However, this leaves at least one fourth of definite and demonstrative descriptions which contain new information, a fact which should be taken into account in text generation. We will now compare the definite and the demonstrative inside each category of use.

Information Repeating Anaphors From the tables (1.2) we can observe the following facts: First, the information comes from the antecedent (directly or not - results in AO and LR column of the tables) in 70% of the cases for definite and in 57% of the cases with the demonstrative. This means that the demonstrative is probably more able to allow inferences from the context. Second, the content of IRA must be inferred with 69% of the demonstrative NPs and with 51% of the definite NPs.

These data are important because the existing generation algorithms only generate anaphoric noun phrases from the explicitly given information in the antecedent. These results show that if we introduce the possibility of generating demonstrative anaphoric noun phrases we will have to allow inference from other sources than the antecedent.

Information Adding Anaphors For noun phrases which add information, we found (1.3) that the most IAA demonstrative fall in the "modifier" category and most of the definite description fall into the No Lexical Relation one. We should approach this result with caution because the high proportion of proper nouns, combined with the fact a proper noun never have a lexical relation with a common noun, strongly biases the result,

Demonstrative	AO	A+C	LR	LR+C	WKL
Dir+mod	7	4	0	0	0
Dir	72	0	0	0	0
Ind+mod	1	0	10	15	17
Ind	0	22	57	10	44
total	80	26	67	25	61
proportion	21%	10%	26%	10%	23%

Definite	AO	A+C	LR	LR+C	WKL
Dir+mod	90	59	0	0	4
Dir	410	0	0	0	0
Ind+mod	4	6	72	31	53
Ind	0	9	136	22	126
total	511	73	208	53	183
proportion	50%	7%	20%	5%	18%

Demonstrative	Without inferences	With inferences
Direct	72	0
Direct + mod	7	4
Indirect	0	133
Ind + mod	1	42
total	80 (21%)	179 (69%)

Definite	Without inferences	With inferences
Direct	410	0
Direct + mod	90	63
Indirect	0	299
Ind + mod	4	162
total	504 (49%)	524 (51%)

Figure 1.2: Results for Information Repeating Anaphors

thus hiding the possibility that if the content of the anaphor has no lexical relation with a common noun antecedent, the demonstrative is used very frequently. Moreover, 43% of the demonstrative add information by this mean and most of them have a common noun as antecedent which is not the case for definite. So, the reclassifying power of demonstrative is illustrated by this data, and is obviously a mean to add information about a referent with the demonstrative.

Demonstrative	SLR	Mod	SLR +mod	No LR
Direct mod	0	11	0	0
Indirect mod	2	33	8	22
Indirect	4	0	0	13
total	6 6,3%	40 42%	8 8,4%	41 43%

Definite	SLR	Mod	SLR +mod	No LR
Direct mod	0	43	0	0
Ind. mod	1	7	0	108
Indirect	12	38	14	61
total	13 4,6%	81 28,5%	21 7,4%	169 59,5%

Figure 1.3: Results for Information Adding Anaphors

Synthesis We have now to observe the distribution of the different phenomena among all the anaphoric noun phrases. We present it in the table (1.4). The top table presents the proportion of the different anaphoric noun phrases within the category of IRA, and the bottom table the proportion of each category for IAA (each category is abbreviated as before and preceded by the type of determiner). Because of the bias mentioned in the previous section we divided each table into two parts, one for anaphoric noun phrases with proper nouns as antecedent, the second for common nouns as antecedent.

We propose the following basis for a generation algorithm. It is based on the salience criteria proposed in the literature and on the frequency of occurrences in the corpus which is the only parameter available at this mo-

IRA	dem AO	dem A+C	dem LR	dem LR+C	dem WKL	def AO	def A+C	def LR	def LR+C	def WKL
proper N	0	0	0	0	25	3	1	0	0	91
proportion	0%	0%	0%	0%	20,8%	2,5%	0,8%	0%	0%	75,8%
common N	80	26	67	25	36	508	72	208	53	92
proportion	6,8%	2,2%	5,7%	2,1%	3,1%	43,5%	6,2%	17,8%	4,5%	7,9%

IAA	dem SLR	dem mod	dem SLR+mod	dem NoLR	def SLR	def mod	def SLR+mod	def NoLR
proper N	0	10	0	11	0	15	0	160
proportion	0%	5,1%	0%	5,6%	0%	7,6%	0%	81,6%
common N	6	30	8	30	13	66	21	9
proportion	3,3%	16,4%	4,4%	16,4%	7,1%	36%	11,5%	4,9%

Figure 1.4: Functions of anaphoric noun phrases in the whole corpus

ment:

If the function of the description is IRA

If antecedent = proper noun

If the referent is focused, use a demonstrative and infer the content of the anaphor from the world knowledge

If the referent is unique in the context, use the definite and infer the content from:

- world knowledge
- the antecedent (give the type of the antecedent)
- the antecedent and the context

If antecedent = common noun:

If the referent is focused, use the demonstrative and infer the content from:

- the antecedent
- a lexical relation
- world knowledge
- antecedent and context
- lexical relation and context

If the referent is unique, use the definite and infer the content from:

- antecedent
- lexical relation
- world knowledge
- antecedent and context
- lexical relation and context

If the function of the description is IAA

If antecedent = proper noun

If the referent is unique in the context, use the definite and

- a NP without lexical relation
- a noun describing the type of the referent and modifiers

If the referent is focused use the demonstrative and :

- a NP without lexical relation
- a noun describing the type of the referent and modifiers

BIBLIOGRAPHY

If antecedent = common noun

If the referent is unique in the context, use the definite and :

- modifiers
- specifying lexical relation and modifiers,
- specifying lexical relation
- a NP without lexical relation

If antecedent = common noun and if focused, use the demonstrative and

- equally use modifiers or a NP without lexical relation
- a specifying lexical relation and modifiers
- lexical relation

4 Conclusion and Future Work

In this paper we laid the ground for an extension of the existing generation algorithms for referring expressions which would encompass not only anaphoric definite descriptions but also anaphoric demonstrative descriptions. Based on the classification proposed in the literature, we described the results of a first corpus analysis. We then proposed a more detailed classification whose classes are arguably needed for a better specification of the different uses of definite and demonstrative. We applied this classification to the corpus thereby extracting from the resulting analysis interesting facts about both definite and demonstrative. We ended by sketching the basis of an algorithm supporting the choice between the two determiners. This study must be completed by adding parameters to lead to a real algorithm. These parameters could be discourse-linked restrictions (position in the anaphoric chain for example) or syntactic restrictions (distance from the antecedent, grammatical functions of the antecedent...).

5 Acknowledgements

I would like to thank Claire Gardent for her precious help and support, and Eric Kow for all the scripts and programmes he wrote for the corpus exploitation.

Bibliography

- [Beaumont et al. 1998] Beaumont C., Lecomte J., et Hatout N., (1998) *Etiquetage morpho-syntaxique du corpus "Le Monde" pour les besoins du projet PAROLE*, Technical Report, INALF, Nancy.
- [Clark, 1977] Clark H.H., Bridging *Thinking: Readings in Cognitive Science*, Johnson-Laird P.N., Wason P.C. (eds), Cambridge, Cambridge University Press.

- [Corblin, 1987] Corblin F. (1987), *Indéfini, Défini et Démonstratif*, Genève, Paris, Droz.
- [Corblin, 1999] Corblin F. (1999), Les références mentionnelles : le premier, le dernier, celui-ci. In *La référence (2) Statut et processus*, Mettouch A. and Quinyin H. (eds.), Travaux linguistiques du CERLICO, Rennes, PUF.
- [Corley et al., 2001] Corley S., Corley M., Keller F., Crocker M.W., et Trewin S., (2001) Finding Syntactic Structure in Unparsed Corpora : The Gsearch corpus query system, *Computer and Humanities*, 35(2), pp81-94.
- [Dale et Reiter, 1995] Dale R. Reiter E., (1995), Computational Interpretations of the Gricean Maxims in the Generation of Referring Expressions *Cognitive Sciences* 19(2), pp233-263.
- [Gardent, Striegnitz, 2000] Gardent C., Striegnitz K., (2000), Generating Indirect Anaphora, proceedings of *IWCS'00 (International Workshop on Computational Semantics)*.
- [Gardent et al. 2003] Gardent C., Manuélian H., Kow E., (2003), Which Bridges for Bridging Descriptions, in *EACL Workshop on Linguistically Interpreted Corpora* proceedings.
- [Kleiber, 1986] Kleiber G., (1986), Pour une explication du paradoxe de la reprise immédiate un N - le N / un N - Ce N *Langue Française*, 72, pp 54-79.
- [Kleiber, 1988] Kleiber G., (1988), Reprise immédiate et théorie des contrastes, *Studia Romanica Posnaniensa*, 13, pp 67-83.
- [Krahmer et al., 2001] Krahmer E., van Erk S., Verleg A., (2001), A Meta-Algorithm for the Generation of Referring Expressions, proceedings of *8th European Workshop on Natural Language Generation* pp 29-39.
- [Lecomte, 1997] Lecomte J., (1997) *Codage Multext - GRACE pour l'action GRACE / Multitag*, Technical Report, INALF, Nancy.
- [Manuélian, 2002] Manuélian H., (2002), Annotation des descriptions définies : le cas des reprises par les rôles thématiques, proceedings of *RECITAL 2002, Nancy, France*, pp455-467.
- [Manuélian, 2003] Manuélian H., (2003), Une analyse du démonstratif en corpus, proceedings of *TALN 2003, Batz sur Mer, France*.
- [Muller, Strube, 2001] Muller C., Strube M., (2001) Annotating Anaphoric and Bridging Relations with MMAX, proceedings of *2nd SIGDial Workshop on Discourse and Dialogue*, pp90-95.
- [Poesio, Vieira 1998] Poesio M., Vieira R., (1998), A Corpus Based Investigation of Definite Description Use, *Computational Linguistics*, 24-2 pp183-216.
- [Vieira et al. 2002] Vieira R., Salmon-Alt S., Gasperin C., Schang E., Othero G., (2002), Coreference and Anaphoric Relations of Demonstrative Noun Phrases in a Multilingual Corpus, proceedings of *DAARC*.
- [Van Deemter, 2000] Van Deemter K., (2000), Generating Vague Descriptions, proceedings of *First International Conference on Natural Language Generation*, pp 179-186.

BIBLIOGRAPHY

- [Van Deemter, 2001] Van Deemter K., (2001), Logical Form Equivalence : the Case of Referring Expressions Generation, proceedings of *8th European Workshop on Natural Language Generation* 21-29.