An Evaluation of Inter-Annotator Agreement in the Observation of Anaphoric and Referential Relations

François Trouilleux, Gabriel G. Bès, Éric Gaussier

An Evaluation of Inter-Annotator Agreement in the Observation of Anaphoric and Referential Relations

François Trouilleux[*†], Gabriel G. Bès[*], Éric Gaussier[†]

[*] Groupe de recherche dans les industries de la langue (GRIL)
UFR-LACC, Université Blaise-Pascal, Clermont 2
34, avenue Carnot. 63037 Clermont-Ferrand - France
Firstname.Lastname@univ-bpclermont.fr

[†] Xerox Research Centre Europe
6, chemin de Maupertuis. 38240 Meylan - France
Firstname.Lastname@xrce.xerox.com

Abstract

When proposing a description of the data he observes, the linguist must make sure that his observations may be also regularly made by other persons. In this paper, we introduce a typology of anaphoric and referential relations and an experiment which aims at assessing that this typology is operational. Given three newspaper articles, five students were asked to identify anaphoric and/or referential relations between expressions and referents. This inter-subjectivity test confirms results already obtained: coreference is an operational notion, but the perspicuity of other relations is not obvious.

Linguists make use of many different notions which often overlap. It is the case for the notions of "coreference" and "anaphora" in the domain we are concerned with, namely text interpretation. Coreference, understood as the identity of reference of distinct expressions, includes cases of anaphora, but not all cases of anaphora (we have met the terms "associative anaphora", "indirect anaphora", "*one* anaphora", among others; none were used to refer to cases of coreference strictly speaking). Things get more complicated when coreference, for some, may extend beyond strict identity.[1] To get a grasp of this problem, we defined a typology of anaphoric and referential relations, which is presented in the first part of this paper.

The goal of this paper, however, is not the presentation of this typology *per se*. Through an experiment aiming at assessing that our typology of anaphoric and referential relations was operational, in the sense that different persons could use it to make the same observations, we would like to show that the sophistication of linguistic descriptions may find its limits in the testing of their operationality. In the perspective of a natural language processing system, it is important to select phenomena for which there is an operational description, in particular as such a description is necessary to evaluate the system.

The first part of the paper presents some preliminary notions and our typology of anaphoric and referential relations (sections 1 and 2). Section 3 introduces the experiment which aims at assessing the operationality of our typology. Evaluation criteria for the different types of observations to be made are given in section 4, then evaluation measures in section 5. The next sections introduce the results (section 6), organized with respect to the different types of relations to be observed, and a discussion of these results (section 7).

1. Preliminary notions

Denotation world. We take for granted that to a text is associated what we call a "denotation world", the world denoted by the text, and we say that expressions in this text "describe" this denotation world. Basically, the denotation worlds which we assume are associated to texts are "the vast complex of things and situations that sentences can be about" (Dowty et al., 1981). A given text may eventually describe several denotation worlds.

Referents. We assume the denotation world associated to a text is populated with "beings" or "referents": the objects, persons, events, facts, situations, etc. the text talks about. The sentence *John loves Mary*, for instance, describes a denotation world which contains three referents: the one denoted by *John*, the one denoted by *Mary* and the one denoted by the whole sentence, *i.e.* the fact that John loves Mary.[2]

Referential and non-referential expressions. In a given text, some expressions have the particularity to point or refer to referents of the denotation world while others do not. We say that the former are "referential" and the latter "non-referential". In the sentence *This shirt is blue*, we consider that *This shirt* is a referential expression; it points to an object in the denotation world associated to the sentence. The expressions *blue*, or *is blue*, or even *shirt* alone, on the other hand, are non-referential; they do not point to a referent but only *describe* the referent pointed to by *This shirt*.

Anaphora. As a rule, expressions in a text are to be interpreted in a context.[3] The expression *the president of the United States*, for instance, may denote Bill Clinton in one context, or Georges Washington in another, or Franklin Roosevelt in yet another context, etc. Similarly, the proper name *Bill Clinton* may be used to refer to the man who is at present president of the United States, but it might very well be used to refer to some other man

[1] It is the case in (Chinchor and Sundheim, 1995).

[2] A given text may describe a denotation world which may or may not be the real world. In both cases, we will talk about "referents"; our referents are *discourse* referents (cf. (Karttunen, 1976)).

[3] By "context", without further qualification, we always mean any kind of context, text or situation. References to a specific kind of context will be made clear by appropriate description (e.g. "textual context", "non-textual context").

bearing this name. The precise reference of expressions, while being usually unique in a given context, varies from context to context.[4]

One note, however, that different expressions impose different constraints on the way they may possibly be interpreted, depending on the richness of their descriptive content. For instance, we would say that the descriptive content of *the president of the United States* is richer than that of *the president*, which is richer than that of *he*. Any man referred to by the first expression may also be referred to by the two others; any man referred to by the second expression may also be referred to by the third; but the inverse of these two statements is not true.

We call "anaphoric expressions" expressions which have some incomplete descriptive content and as such are interpreted in relation to some other contextual expressions, *i.e.* a textual context. More specifically, we will consider as possible anaphoric expressions pronominal noun phrases, possessive determiners, temporal adverbial expressions, or noun phrases determined by a definite article or a demonstrative. In the following sentence,

1. *Comme elle l'avait laissé entendre en février, la BNP a décidé de rapprocher ses filiales de crédits spécialisés.*
   [As it let it be understood in February, the BNP has decided to merge its subsidiaries of specialized leasing.]

we consider that the personal pronoun *elle* and the possessive determiner *ses* are anaphoric with respect to *la BNP*, that the clitic pronoun *l'* is anaphoric with respect to the clause *la BNP a décidé de rapprocher ses filiales de crédits spécialisés*. Allowing some extension of the notion of anaphora to non-textual contexts in the case of temporal expressions, we will also consider that the phrase *en février* is anaphoric in that it does not provide a complete description of its referent, in this case the month of February 1998, and the identification of this referent relies on contextual information, in this case the date of the article this example has been taken from.

We will use the term "anaphora" both for cases where the contextual expression precedes or follows the anaphoric expression.

2. A typology of anaphoric and referential relations

Taking as a starting point the interpretation of anaphoric expressions, we propose a typology of the different relations between expressions which we think play a role in the interpretation of texts, in particular in the interpretation of noun phrases. We are here concerned with relations between expressions insofar as these relations are not expressed syntactically. The term "expression" here is also to be considered in a broad sense; it includes expressions proper (the words which are actually in the text), but also "elliptical expressions", which may be inferred from the resolution of ellipsis, ellipsis being induced by some expressions proper.

We distinguish two main types of relations depending on whether the relation involves a relation between the referents of the two expressions, or whether it only involves their description. The former are called "referential relations", the latter "description relations".

2.1. Referential relations

In referential relations, both related expressions are referential. In our typology, five types of relations are considered: coreference, distinction, set-membership, part-of and a fifth underspecified relation for relations which are not any of the four other types.

2.1.1. Coreference

If two referential expressions denote the same referent, the relation is coreference. In the following sentence,

2. *Allianz présente son nouveau visage.*
   [Allianz introduces its new face.]

the expressions *Allianz* and *son* refer to the same referent; they are coreferential.

An expression e_i may corefer with a set of expressions e_j, ..., e_n if it denotes a set *S* and the expressions e_j, ..., e_n each denote a distinct element or subset of *S* and the full extension of *S* is specified by e_j, ... e_n. In the sentence:

3. *Allianz rappelle que ses 10% d'Ergo ne sont pas stratégiques et que les deux compagnies restent des concurrentes acharnées.*
   [Allianz insists that its 10% of Ergo are not strategic and that the two companies remain fierce competitors.]

the expression *les deux compagnies* denotes a set composed of two entities, the referent of *Allianz* and the referent of *Ergo*. We consider that *les deux compagnies* is coreferent with the set composed of the two expressions *Allianz* and *Ergo*.

2.1.2. Distinction

The relation called "distinction" chiefly aims at giving an account of the anaphoric relation in the use of the adjective *autre* ("other"). The adjective *autre* may be seen as a two place predicate: if some referent o_i is described as *autre*, it is so in relation with another referent o_j. We say that o_i is distinguished from o_j. In the following sentence,

4. *Cette annexe parisienne du Palais de justice […] devrait rapidement être suivie d'autres pôles en province.*
   [This Parisian annex of the Law Courts should be soon followed by other poles in the provinces.]

the referent of *d'autres pôles* is distinguished from the referent of *Cette annexe parisienne du Palais de justice*.

The adjective *autre* expresses the negation of coreference; it also induces a description relation (see below): the two referents in a distinction relation are of the same type, *i.e.* there is a description which applies to both referents. Rules may be formulated. Let us assume we have a referent o_i on the one hand and a referent o_j on the other hand, with o_j described as *autre* in comparison to o_i.

If both o_i and o_j are member of a referent o_k (in a sense defined below), the description which is common to the two referents is expressed in e_k, which denotes o_k.

[4] This presentation of the notion of anaphora is partly inspired from (Ranta, 1994).

If, in addition to *autre*, o_j is described by a nominal description, this description also applies to o_i (this is the case in example 4 where the referent of *Cette annexe* is indeed a *pôle*).

If neither of the two conditions above hold, either there is in e_i, which denotes o_i, a nominal description which applies to o_j or this description is implicit. This is the case in the following text, where *les autres* denotes a set of beings which are neither Pierre, nor Juliette, but which are of the same type, namely persons.[5]

5. *Seules Pierre et Juliette sont venus. Les autres étaient en vacances.* ♦
[Only Pierre and Juliette came. The others were on holiday.]

2.1.3. Set membership

We distinguish in the denotation world referents which are singular beings and referents which are sets of at least two singular beings. The relation labeled "set-membership" in our typology is a relation between either a singular being or a set o_i on the one hand and a set o_j on the other hand. We say that o_i is a member of o_j if and only if:

- o_i is a singular being and is an element of o_j,
- or o_i is a set and is a proper subset of o_j.

This relation so covers both the set membership and the set inclusion relations of set theory.

It must be noted that a set membership relation involves a description relation (see below): to observe a set membership relation "o_i is member of o_j" it is necessary that the description the text provides of the referent o_j be applicable to its element or subset o_i (except if o_j is described by a noun which is typically used in the singular to refer to a group, e.g. *gouvernement* = a set of persons, in which case the description relation is implicit). The requirement that a relation of type "set membership" be doubled with a description relation follows from the two possible ways of defining a set: either by enumerating its elements (e.g. {1, 2, 3, 4}), or by providing a description of its elements (e.g. {x | x is a positive integer inferior to 5 }). In the following text,

6. *Le pôle financier bénéficiera d'équipements informatiques. Les magistrats auront notamment à leur disposition des logiciels d'instruction assistée par ordinateur*
[The financial pole will receive computer equipments. In particular, the magistrates will have at their disposal software for computer-aided judgment preparation.]

the referent of *des logiciels d'instruction assistée par ordinateur* stands in a relation of type "set-membership" with the referent of *équipements informatiques*. The description in the latter expression is applicable to the referent of the former; in other words, software *is* computer equipment.

2.1.4. Part-of

An anaphoric expression may denote a referent which is part of another referent. In the following text,

7. *Si vous allez à Clermont-Ferrand, visitez la cathédrale.* ♦
[If you go to Clermont-Ferrand, visit the cathedral.]

the referent of *la cathédrale* is part of the referent of *Clermont-Ferrand.*

Contrary to the set-membership relation, the part-of relation does not involve a description relation. If one considers our example text, it is not the case that *Clermont-Ferrand* denotes a set of cathedrals the referent of *la cathédrale* would be an element of. This difference in the presence or absence of a description relation is a criterion to distinguish the set-membership and part-of relations.

2.1.5. Unspecified relation

We did not try to specify all possible relations which could be observed between two referents, but rather left this issue open and used an unspecified relation for all cases where a relation can be seen but it cannot be analysed as one of the four previous types. In the following text,

8. *Cet immeuble [...] accueillera sur 6.400 mètres carrés, d'ici à la fin de l'année, 274 magistrats et fonctionnaires [...]. Montant du loyer : 21,6 millions de francs par an [...]*
[This building will host on 6400 square meters, by the end of the year, 274 magistrates and civil servants. Amount of the rent: 21.6 million francs a year.]

[le] loyer is anaphoric with respect to *Cet immeuble*; there is a relation between the referents of the two expressions, but this relation is not one of the previously defined relations. The unspecified relation is to be used in this case.

2.1.6. A special case for dates

When it comes to temporal objects (days, months, years, etc.), we have at our disposal a language to describe this kind of referents unambiguously. The expression *January 1st, 2000*, for instance, would presumably denote the same object in any context; there is no need for contextual information to identify the referent of this expression.

As a rule, one may unambiguously refer to a specific year using the number which is used as its name (e.g. *1999*); one may unambiguously refer to a specific month using its name (e.g. *January*) and the name of the year it is part of; one may unambiguously refer to a specific day using its number (e.g. *1st*) together with an unambiguous description of the month it is part of (e.g. *January 2000*); the same holds for centuries, decades, weeks, etc.

Having at our disposal this system of unambiguous descriptions, we make a special case for temporal expressions. We will consider as anaphoric any temporal expression which does not provide a complete description of its referent, regardless of the nature (the text or the situation) of the contextual information needed to identify this referent. As soon as we meet an incomplete date expression, we will consider that there is a referential

[5] Examples are all taken from the texts used for the experiment unless marked with ♦.

relation between this expression and the contextual element(s) which makes the interpretation of this expression possible.

In the following text,

9. *L'assureur allemand a engagé un tournant stratégique avec le rachat des AGF au début de l'année. Le développement de sa présence en France l'amènera se faire coter à Paris le 12 juin prochain.*
   [The German insurance company engaged in a strategic turn with the takeover of the AGF in the beginning of the year. The development of its presence in France will lead the company to quotation in Paris on this coming 12th of June.]

We will consider that *l'année* and *le 12 juin prochain* are anaphoric expressions. The expression *le début de l'année* will not be considered as anaphoric in that provided that *l'année* is unambiguously interpreted *le début de l'année* is also unambiguously interpreted (in this case *l'année* denotes the year 1998; we consider that *le début de 1998* would provide a complete description of its referent).

The fact that anaphoric temporal expressions may relate to a context which is not expressed linguistically will lead us, in the experiment to be described, to adopt a different strategy to study the interpretation of these expressions: the point will not be so much to identify the necessary contextual information, but to provide a complete description of the referent of anaphoric date expressions.

2.2. Description relations

The relations we have described so far all involve the denotation of the related expressions. There are cases where the expression to which an anaphoric expression is related is not referential, or cases where the relation between a referential anaphoric expression and a referential antecedent only involve the description of their respective referents and not their denotation. These cases are instances of "description" relations.

2.2.1. Basic description relations

Some pronouns in French may be non-referential. In the following sentence, the clitic pronoun *l'* is non-referential:

10. *Marie est intelligente, Juliette ne l'est pas.* ♦
    Marie is intelligent, Juliette is not.

It is used to describe the referent denoted by *Juliette*, yet it does not have a descriptive content. This descriptive content is to be taken from the expression *intelligente*.

Verbal ellipsis is also a case where a description is to be taken from the context. The following sentence is a conjunction of two clauses, where the description *s'élève* in the first conjunct is to be taken in the second conjunct (*ses encours à 54 milliards*) in order to interpret it as "ses encours s'élèvent à 54 milliards".

11. *Sa production annuelle s'élève à 20 milliards de francs et ses encours à 54 milliards.*
    [Its annual production amounts to 20 thousand million francs and its debts to 54 thousand million.]

A referential anaphoric expression may be related to another referential expression in a relation which only involves the description of the two referents. In the following text, the clitic pronoun *en* is to be interpreted as anaphoric to *des logiciels d'instruction assistée par ordinateur*. It is not the case that this pronoun denotes the same object as its antecedent. The expression *des logiciels d'instruction assistée par ordinateur* denotes the software the magistrates will have at their disposal, while *en* denotes the software which the judges Eva Joly and Jean-Pierre Zanotto have at present. The relation between the anaphoric expression and its antecedent only concerns the description of the two referents.

12. *Les magistrats auront notamment à leur disposition des logiciels d'instruction assistée par ordinateur. Actuellement, seuls les juges Eva Joly et Jean-Pierre Zanotto en disposent à la galerie financière de Paris.*
    [The magistrates will have at their disposal software for computer-aided judgment preparation. At present, only the judges Eva Joly and Jean-Pierre Zanotto have some in the Paris finance hall.]

Noun phrases whose head is an adjective (e.g. *le premier* [the first]) or a cardinal number (e.g. *deux* [two], *un million* [one million]) are also analysed as involving a description relation, unless there is a set-membership or a distinction relation between the referents of the two expressions. In the following sentence, *15 millions* is to be considered anaphoric to *21,6 millions de francs* in a description relation. The two expressions denote two distinct objects and none is an element, nor a subset, nor part of the other.

13. *Montant du loyer : 21,6 millions de francs par an auxquels s'ajoutent 15 millions de travaux spécifiques pour sécuriser les lieux que prend en charge le propriétaire de l'immeuble.*
    [Amount of the rent: 21.6 million francs a year, to which must be added 15 million for alterations to increase the security of the place, which will be borne by the owner of the building.]

2.2.2. Reference to a description

Another type of description relation appears when an anaphoric expression actually refers to a description mentioned elsewhere in the text, as in the sentence

14. *Si Marie est intelligente, Juliette n'a pas cette qualité.* ♦
    [If Marie is intelligent, Juliette does not have this quality.]

where the anaphoric referential noun phrase *cette qualité* refers to the description *intelligente*. In this case, the antecedent expression is non-referential, according to our use of this term. With this type of relation, the anaphoric expression gives the description the status of referent.[6]

2.3. Paraphrase

[6] There is no instance of this relation in the texts used for the experiment to be described.

Finally, we observe with reported speech a third type of relation, which is in a sense midway between a referential and a description relation. In the following sentence,

15. *Jacques ne viendra pas. C'est Marie qui me l'a dit.* ♦
[Jacques won't come. It is Marie who told me so.]

we do not analyse the relation between the clitic pronoun *l'* and *Jacques ne viendra pas* as one of type "coreference". Rather, we will consider that (1) *Jacques ne viendra pas* denotes a referent o_i (the fact that Jacques will not come), that (2) the denotation of the clitic pronoun is a discourse D_i (possibly identical to the antecedent sentence but not necessarily; Marie may have said something like "cet idiot ne veut pas venir" [this idiot does not want to come]) and that (3) the denotation of D_i is the same as that of *Jacques ne viendra pas*. This type of relation is labelled "paraphrase", as a sentence s_i which describes a denotation world w_i in terms different from those of another sentence s_j which also describes w_i may be called a paraphrase of s_j. The observation of this relation is limited to cases where a clitic pronoun is the object of a verb which belongs to the class of speech verbs (e.g. *dire, raconter,* etc.). This relation is so dependent on the semantics of the verb the pronoun is the object of.

3. Questionning the operationality of linguistic descriptions

We have presented a typology of relations which aims at describing the different relations which may be observed between the referents of expressions or between expressions themselves. When we observe texts, we are able to identify such relations, but it is necessary to question the operationality of our system. It is necessary to make sure that the observations we make may be also regularly made by other persons. If, given the same data, different persons make different observations while using the same descriptive system, then this descriptive system is not operational.

In the remainder of this paper, we will introduce an experiment aiming at assessing the operationality of our typology of anaphoric and referential relations. The results of this experiment will show that linguistic descriptions may well be too sophisticated when it comes to practical application.

Five annotators were given three texts from *La Tribune des Fossés* (a French newspaper mainly concerned with finance) and asked to annotate the anaphoric and referential relations they identified in these texts according to the typology presented and with a specifically designed annotation scheme. In the mean time, an "expert" provided his own interpretation of the text, an annotation which will stand as the key (the gold standard) against which the five response annotations will have to be evaluated. The expert is the person who took the greater part in the definition of the typology and who trained the annotators to the task.

The five annotators were five students working at GRIL, Université Blaise-Pascal, Clermont-Fd. They so all had some competence in linguistics, even if to different degrees. In addition to this competence, they were specifically trained to the task they had to do. The training was organized as follows:

- a two hour course introducing the typology,
- first annotation exercise: the five annotators, and the expert carry out, together, the annotation of an article from *La Tribune*,
- second annotation exercise: the five annotators, divided in two teams, carry out the annotation of another article from *La Tribune*, this time on their own,
- the training ends with the discussion and correction of these annotations.

In addition, the annotators were given a document presenting the typology and the annotation scheme (29 pages), together with a complementary document (3 pages) which clarified some points raised in the course of the training period.

The three texts from *La Tribune* contain respectively 115, 272 and 676 words. They had not been seen by any of the participants, expert included, before the end of the training period.

4. Observation and evaluation predicates

The annotation of anaphoric and referential relations in a text induces a set of "observation predicates", an observation predicate being a formula stating something like "I see a relation of type *x* between this expression and this other expression". The different types of relations defined in our typology lead to different observation predicates. The main distinction between referential relations and description relations leads to observation predicates which will sometimes be about referents and sometimes about descriptions.

To compare a response observation predicate about referents to the observation predicates of the key, we will have to figure out what is the correspondence between the referents of the key and those of the response. To this end, we will use the method presented in (Trouilleux et al., 2000). We will not present this method here, as full details are given in this paper, which the interested reader is invited to read. The important point is that we have a correspondence between the referents of the key and those of the response and so will be able to compare all observation predicates, be they about referents or about descriptions.

To perform the comparison of the observation predicates formulated by the annotators (the response) to the observation predicates formulated by the expert (the key), we will need "evaluation predicates", *i.e.* predicates which state whether a response observation predicate is "correct" or not considering the observation predicates of the key. A perfect response annotation with respect to the key will contain all and only the observation predicates of the key. The evaluation predicates will take into account not only the correctness or incorrectness of the response observation predicates, but also the fact that key observation predicates may be missing in the response or the fact that the response annotation may contain spurious observation predicates.

This section introduces the different types of observation predicates an annotator could formulate depending on the types of relations he/she observed and

the evaluation predicates for each of these observation predicates.

4.1. Coreference

We consider that identifying coreference chains is to map the expressions in these chains to referents. A coreference chain C_i is associated to a referent o_i. The identification of a coreference chain induces a set of observation predicates of the form "e_i denotes o_i", where e_i is an expression of the coreference chain and o_i is the referent associated to the coreference chain. We call such an observation predicate a "denotation assignment".

To a coreference chain of cardinality n in the key correspond n - 1 denotation assignments. Referents in our case only exist insofar as they are denoted by an expression. It follows that for a single non-coreferring expression, the mapping from this expression to its referent is trivial. Similarly, in a coreference chain, there is one expression for which the denotation assignment is trivial. This expression may be seen in our system as representing the referent of the coreference chain.

An observation predicate of the form "e_i denotes o'_i" in a response annotation is *correct* if "e_i denotes o_i" exists in the key and o'_i and o_i have been analysed as corresponding referents. It is *incorrect* if there is an observation predicate of the form "e_i denotes o_j" in the key and o_j corresponds to a response referent o'_j different from o'_i (*i.e.* in the response, e_i has been placed in a coreference chain but in the wrong one).

An observation predicate of the form "e_i denotes o'_i" in a response annotation is *spurious* if there is no observation predicate of the form "e_i denotes o_i" in the key (e_i does not belong to a coreference chain in the key).

An observation predicate of the form "e_i denotes o_i" in the key is *missing* in a response annotation if there is no observation predicate of the form "e_i denotes o'_i" in the response (e_i belongs to a coreference chain in the key, but not in the response, or e_i has been taken as representing the referent of a response coreference chain which does not correspond to a coreference chain of the key).

4.2. Other referential relations

The observations predicates for referential relations other than coreference, namely the "set-membership", "part-of", "distinction" and "unspecified" relations, are predicates over referents, not expressions. Again, the method we use to establish the correspondence between the referents of two annotations enables us to deal with this problem, as we know the correspondence between the referents of the key and that of the response.

We will say that a response observation predicate of the form "o'_i relation o'_j", where *relation* stands for any of the four possible referential relations other than coreference, is *correct* if there is in the key "o_i relation o_j" and the response referents o'_i and o'_j correspond to the key referents o_i and o_j, respectively, except for the unspecified relation where the order of the arguments is irrelevant.[7]

A response observation predicate of the form "o'_i relation o'_j" is *incorrect* if there is in the key an observation predicate of the form "o_i relation o_j", o'_i and o'_j correspond to o_i and o_j, respectively, and the relation used in the response observation predicate is different from that used in the key observation predicate. It is also *incorrect* if there is in the key "e_i denotes o_j" and "e_i denotes o'_i" in the response, or if there is in the key a paraphrase or description relation between e_i and e_j and these two expressions denote o'_i and o'_j in the response. Intuitively, what is incorrect is the use of a relation x instead of a relation y.

A response observation predicate of the form "o'_i relation o'_j" is *spurious* if there is no observation predicate over the two corresponding referents in the key, or over the expressions which are said to denote o'_i and o'_j.

A key observation predicate of the form "o_i relation o_j" is *missing* if there is no observation predicate over the two corresponding referents/expressions is formulated in the response using the same relation.

4.3. Dates

For expressions denoting dates, the annotators were required to supply a specific description of the date if such a description could be inferred. Otherwise they had to use the unspecified relation. The observation predicates for dates expressions can be seen as having the form "e_i denotes the date denoted by the specific description d_i".

We will say that an observation predicate in the response is *correct* if the key contains the same description d_i for e_i and *incorrect* if the key contains for e_i a description d_j different from d_i.

An observation predicate of the form "e_i denotes the date denoted by the specific description d_i" in the key is *missing* in the response if no description is given for e_i. The same kind of observation predicate in the response is *spurious* if there is no description for e_i in the key.

4.4. Description and paraphrase relations

Description and paraphrase relations are relations between two expressions: an anaphoric expression e_i and its antecedent e_j, *i.e.* the expression through which e_i is to be interpreted. Given these two expressions, let us note the observation of a description relation as a predicate of the form "e_i description e_j" where e_i is the anaphoric expression and e_j its antecedent. Similarly, let us note the observation of a paraphrase relation as a predicate of the form "e_i paraphrase e_j".

We will say that a response observation predicate of the form "e_i relation e_j", where *relation* stands for the predicate "description" or "paraphrase" is *correct* if the same predicate exists in the key.

It is *incorrect* if there is in the key an observation predicate which relates e_i to an expression e_k distinct from e_j, or which relates the referent of e_i to a referent which is not denoted by e_j.

A response observation predicate of the form "e_i relation e_j" is *spurious* if it exists in the response and there is no observation predicate over the two expressions or the two corresponding referents in the key, or if a different relation has been observed in the key.

A key observation predicate "e_i relation e_j" in the key is *missing* in the response if either there is no observation predicate over the two expressions or the two corresponding referents in the response, or there is an

[7] This latter qualification holds for every instance of the unspecified relation; we shall not make it explicit in the remainder of the text.

observation predicate which links the two expressions by a different relation.

5. Evaluation measures

Our goal is to assess that our typology is operational. We want to evaluate whether what we say we see in texts, which we describe the way we did above (section 2), can actually also be seen by other observers and described in the same way.

We would say our description of the phenomena is fully operational if and only if

- the annotators observed all and only the relations we observed,
- and they classified these relations the same way as we did.

However, it is most unlikely that a human annotator will produce a perfectly correct annotation. The annotator may be tired, bored with the task, have something else to do, or fail to concentrate on the task for whatever reason: errors are expected. We will so consider that our typology is operational if the observations we made are also made by the annotators as a whole or in their majority.

In order to assess the operationality of our typology, we so will use evaluation measures which take the group of annotators as a whole. These measures are based on the evaluation measures we obtain independently for each of the five annotations. We first present the evaluation measures for a specific annotation (individual measures), then the measures used to assess operationality (global measures).

5.1. Individual measures

We will use, classically, recall and precison as the evaluation measures. In addition, we will use three measures for error analysis. These measures have all been presented in (Trouilleux et al., 2000) where they were applied to coreference resolution.

We assume we want to evaluate the set of observations O_i made by an annotator i with respect to the key.

5.1.1. Recall and precision

Let *possible* be the number of observation predicates in the key and *actual* the number of observation predicates actually made by annotator i (the cardinal number of O_i). Let C_i be the number of observation predicates made by annotator i which have been judged *correct* by the evaluation predicates. We have the following recall and precision measures for the observations of annotator i:

- $recall_i = C_i / possible$
- $precision_i = C_i / actual_i$

5.1.2. Error analysis

In addition to recall and precision, we will use three measures for error analysis: substitution, over-generation and under-generation. These measures are obtained using as operand the number of observation predicates which are judged *incorrect*, *spurious* or *missing*, respectively.

Let I_i, S_i and M_i be respectively the number of observation predicates judged *incorrect*, *spurious* and *missing* in the evaluation of the observations made by annotator i. The evaluation predicates considered identify three different types of errors. Let E_i be the total number of errors made by annotator i:

- $E_i = I_i + S_i + M_i$

Given these operands, the three error analysis measures for annotation i are:

- $substitution_i = I_i / E_i$
- $over\text{-}generation_i = S_i / E_i$
- $under\text{-}generation_i = M_i / E_i$

These measures are to be interpreted as follows. A high value for under-generation will indicate that the errors made by the annotator consist chiefly in not making an observation where the expert does. A high value for over-generation will indicate that the errors made by the annotator consist chiefly in making an observation where the expert does not. The substitution measure will have different semantics depending on the type of relation considered. For coreference, it will indicate that expressions which should be included in a coreference chain have been identified by the annotator but placed in the wrong coreference chain. For other referential relations, it will indicate that the annotator rightly saw a relation between two referents or expressions observed in the key, but wrongly classified this relation.

5.2. Global measures

The measures presented above have been defined to evaluate a single annotation with respect to the key. However, in order to assess that our typology is operational or not, we need global measures, which take into account the group of annotators as a whole, rather than just each annotator in turn.

5.2.1. Average

The simplest global measures are obtained by computing the average measures out the five individual measures. Average recall (A-recall) and average precision (A-precision) will then be:

- A-recall = $1/5 \Sigma_i R_i$
- A-precision = $1/5 \Sigma_i P_i$

Average precision and recall can be computed for all the observation predicates, or for observation predicates divided by types of relations and expressions. In order to compute average precision and recall, we first compute an average annotation for reasons that will become clear below. The precision and recall of this average annotation correspond to the average precision and recall defined above. When evaluating a particular relation type and/or expression type, the average annotation is defined over the union, for all annotations and the key, of the observation predicates for that particular relation(s) and/or expression(s). For each such observation predicate, the average annotation corresponds to the mean of the different annotations.

In addition to our use of an average annotation, we compute the variance of the different annotations, in order to see how spread the different annotations are around the average one. The variance may be used as a way to measure the inter-subjectivity of the observations, since when all the annotators agree the variance equals 0, whereas when there are strong disagreements between annotators, the variance takes on high values. However,

since it is difficult to define thresholds under which to consider that annotators disagree too strongly (this is true for the variance and for other measures, as shown in (Di Eugenio, 2000) for Kappa), we will use the variance as a complement to other measures to compare how different elements of our typology are observed by different annotators.

Finally, inasmuch as our average recall and precision are sensitive to extreme values, *i.e.* annotations which strongly differ from the average one, we make use of an additional measure, the "majority opinion".

5.2.2. Majority opinion

We will use "majority opinion" measures to avoid some shortcomings of the average measures.

Let us assume that the key contains 3 observation predicates A, B and C, and that each of the five annotators only made one correct observation and missed the two other. Both average measures will yield a score of 1/3 recall. This score, however, would not differentiate a situation in which the annotators all made the same observation, say A, from a situation in which for instance two annotators observed A, two annotators observed B and one annotator observed C. The difference in these two situations is that in the former case the observation A is inter-subjective and B and C are not, while only doubt comes out of the latter case.

The majority opinion simply consists in seeing the group as a single individual. Let us call this individual the "ideal observer". For an observation predicate op_i made by the expert, we have five evaluation predicates, one for each annotation. The value of the evaluation predicates may be one of *correct*, *missing* or *incorrect*. If three annotations are judged by the same evaluation predicate for op_i, this evaluation predicate is the one which judges the observation predicate made by the ideal observer. If only two annotations have the same evaluation predicate, we consider that the value *incorrect* has precedence over the value *correct*, which has precedence over the value *missing*. Let us assume, for instance, the key contains the observation predicate "*son* denotes *Allianz*" and this observation predicate is *missing* in three of the five response annotations; the observation predicate would then be considered *missing* in the annotation of the ideal observer. If this observation predicate were *correct* in two annotations, *incorrect* in two other and *missing* in the remaining one, the observation predicate would be considered *incorrect* in the annotation of the ideal observer.

The same holds for *spurious* observation predicates. In this case, the ideal observer is attributed a *spurious* observation predicate if the observation has been made by at least three annotators.

5.2.3. Comparing relations

The different measures we have presented will help us analyse the results in detail. However, in order to avoid arbitrariness, we will, in a first step, use these measures as a way to compare the different relations of our typology. In particular, we will sort the different relations according to the precision and recall figures for the majority opinion. The combination of recall and precision in a single measure is done via the F1-measure:

- F1 = 2(PxR) / (P+R)

6. Results

This section presents the results of our operationality test. As there has been important differences in the way the annotators observed the different relations, we have not tried to compute global measures for the whole typology, but rather analysed the results for each relation in turn. After an inventory of the observations made by the expert and an overview of the results, detailed results will be presented in turn for description relations, paraphrase relations, coreference, dates and, finally, referential relations other than coreference.

6.1. Observations made by the expert

The expert formulated 202 observation predicates which divide up as follows, from the most to the less frequent:

- coreference: 129 (63.8%)
- set-membership: 26 (12.9%)
- dates: 23 (11.4%)
- unspecified relation: 11 (5.4%)
- description relation: 7 (3.5%)
- paraphrase: 4 (2%)
- part-of relation: 1 (0.5%)
- distinction relation: 1 (0.5%)

One notes a much higher frequency of referential relations (94.6% in all) over description (3.5%) and paraphrase relations (2%). Some referential relations are also poorly represented ("part-of" and "distinction" relations). It is clear that for these relations we will not have enough data to reach any conclusion. Finally one notes that coreference is by far the most frequent relation

6.2. Overview of the results

Table 1 shows the average and majority opinion measures, together with the variance measures, for each type of relation. As the number of referential relations other than coreference is not very important, results for these four relations are grouped together. One observes that temporal expressions and coreference relations are by far the best observed, since the F1-measure for the average, respectively the majority opinion, is above 0.75, respectively 0,85, whereas for other types of relation the F1-measures are below 0.5.

	AR	*AP*	AF_1	*MR*	*MP*	MF_1	*V*
TE	0.78	0.90	0.84	0.91	1	0.95	0.65
CO	0.72	0.82	0.77	0.84	0.92	0.88	2.08
D	0.37	0.70	0.48	0.29	1	0.45	0.50
P	0.25	1	0.40	0	-	-	0.27
RR	0.12	0.20	0.15	0.10	0.50	0.16	3.39

Table 1. Evaluation measures by type of relation.[8]

The variance is in the low range for temporal expressions as well as for description and paraphrase

[8] Glossary for Table 1. Columns: *AR*, *AP*, AF_1 = average recall, precision and F-measure respectively; *MR*, *MP*, MF_1 = majority opinion recall, precision and F-measure; *V* = variance. Lines: *TE* = temporal expressions, *CO* = co reference, *D* = description relations, *P* = paraphrase relations, *RR* = referential relations other than identity.

relations, indicating agreement among annotators for these relations. For temporal expressions, the high scores obtained for the F1-measures show that annotators consistently and accurately observed the relation. However, for descriptive and paraphrase relations, the F1-measures are low, and the detailed analysis presented below shows that annotators "agreed in not observing" these two relations. The variance for coreference relations is in the middle range, indicating partial disagreements between annotators. As explained below, this is mainly due to one annotator who did a poor job on this relation (the F1-measure for majority opinion is high indeed: 0.88). Lastly, the variance for other referential relations is high, indicating strong disagreements between annotators on these relations.

The following sections introduce the results for the different types of relations, starting with description and paraphrase relations, as the way the annotators observed these cases results in errors in the identification of coreference relations. For each type of relation, a table gives the complete evaluation measures, both individual and global. [9]

6.3. Description relations

Average recall and precision for description relations amount respectively to 0.37 and 0.70, with a variance of 0.5. Errors are mostly absence of observations, as indicated by the under-generation measures. If one considers the majority opinion, perfect precision is reached but recall is low. Out of the seven description relations to be observed, only two have been seen by the majority. Four have not been seen; they all involve monetary expressions in situations exemplified by (13): *21,6 millions de francs... 15 millions.* These seem not obvious to spot for an untrained observer. One may note, however, that two of the five annotators correctly identified these cases. The remaining case is the description relation in (12). The relation between the clitic pronoun *en* and its antecedent has been seen by the annotators, but they all considered the relation was coreference. It seems that the distinction between coreference and description relation is not obvious to make. One also notes that the fact that coreference was much more frequent than description relations may have led the annotators to overlook this distinction.

	a1	a2	a3	a4	a5	av	mo
R	0.71	0.14	0	0.14	0.86	0.37	0.29
P	0.83	0.50	-	0.50	1	0.70	1
S	0	0	0	0	0	0	0
O	0.33	0.14	0	0.14	0	0.12	0
U	0.67	0.86	1	0.86	1	0.88	1

Table 2. Results for description relations.

6.4. Paraphrase relations

Average recall is 0.25 for paraphrase relations. Three annotators did not use this relation at all, one saw all relations. The majority opinion recall (0), as well as low variance (0.27), reflect an agreement not to use the paraphrase relation. It is not the case, however, that the annotators did not link the four pronouns in question to their antecedents; rather, they judged that the relation was coreference.

The distinction between the paraphrase relation and coreference is probably too subtle. In particular, when the antecedent is *quoted* speech, the distinction may cease to hold, as the words are precisely those which were pronounced, e.g.:

16. *L'assureur allemand [...] considère que cette prise de contrôle lui confère « une très forte position dans le secteur de l'assurance mondiale, avec un pied particulièrement solide dans notre marché domestique qu'est l'Europe », comme l'explique son président.*
[The German insurance company considers that this takeover gives it "a strong position in the sector of world insurance, with a particularly strong foothold in Europe, which is our domestic market", as its president explains.]

The expert also has some responsibility in the failure to correctly analyse these relations. Between the time when he wrote the documentation for the annotators and the time of the experiment, the expert, feeling that the distinction might not be clearly observed, revised its definition of the paraphrase relation type to all cases where the clitic pronoun *le* can be paraphrased as a sentence. The annotators were informed of this change during their training period, but the documentation had not been modified accordingly. This indicates further that there is some confusion on this relation.

	a1	a2	a3	a4	a5	av	mo
R	0	0.25	0	0	1	0.25	0
P	-	1	-	-	1	1	-
S	0	0	0	0	-	0	0
O	0	0	0	0	-	0	0
U	1	1	1	1	-	1	1

Table 3. Results for paraphrase relations.

6.5. Coreference

Coreference is one of the best observed phenomena, with 0.72 average recall and 0.82 average precision. The variance is important (2.08); this is partly due to the fact that annotator 3 did quite a poor job and that the evaluation measures we used tend to strongly penalize errors.

	a1	a2	a3	a4	a5	av	mo
R	0.78	0.83	0.51	0.63	0.87	0.72	0.84
P	0.90	0.84	0.67	0.83	0.84	0.82	0.92
S	0.11	0.24	0.28	0.20	0.30	0.23	0.17
O	0.20	0.35	0.15	0.11	0.43	0.25	0.17
U	0.69	0.41	0.57	0.69	0.27	0.52	0.66

Table 4. Results for coreference.

The number of coreference relations is important enough to allows us to further analyse the results for this relation by type of expression. Table 5 shows the average

[9] Glossary for these tables. Columns: a*i* = annotator *i*, av = average, mo = majority opinion. Rows: R = recall, P = precision, S = substitution, O = over-generation, U = under-generation.

and majority opinion measures for four different types of expressions:

- proper names (15.5% of all coreferring expressions),
- pronouns (49.6%),
- descriptive noun phrases (as opposed to proper names and pronouns, 32.6%),
- clauses and sentences (2.3%).

	AR	*AP*	AF_1	*MR*	*MP*	MF_1	*V*
pn	0.77	0.99	0.87	0.90	1	0.95	2.50
pro	0.76	0.82	0.79	0.90	0.89	0.89	3.68
dnp	0.69	0.76	0.72	0.79	0.97	0.87	2.37
st	0.07	1	0.13	0	-	-	1.67
all	0.72	0.82	0.77	0.84	0.92	0.88	2.08

Table 5. Average and majority opinion recall and precision for coreference, by type of expression.

Analysing the results by types of expressions, one first notes that the coreference relations which involved sentences (3 cases) have not be observed.

Omissions are also what affect the majority opinion recall measure for the identification of coreference involving descriptive noun phrases, as indicated by the high majority opinion precision value (0.97).

Not surprisingly, coreference relations between proper names are the best observed. The identity of the expressions makes errors quite unlikely. The majority opinion recall measure for this type of expressions indicates that two errors were made. The first one consists in not seeing two occurrences of the expression *France* as coreferring. The other error is more interesting in that it points out the limits of the notion of identity. Given

17. *Fort du rachat des AGF, Allianz présente son nouveau visage [...] C'est un nouveau groupe Allianz qui naîtra d'ici à la fin de l'année.*
[In a strong position thanks to its takeover of the AGF, Allianz introduces its new face.[...] It is a new group Allianz which will be born by the end of the year.]

the expert considered that *Allianz* and *un nouveau groupe Allianz* denote the same referent, while the annotators considered in their majority that the two expressions denote two distinct referent, which are however related to each other. There is clearly a link between the two; whether it is identity is indeed questionable; one may argue that the company named Allianz which will exist by the end of the year is not the same as the one which exists at present, especially if the former is described as "new", but one may also point out that the former will be the result of the evolution of the latter and so that they are the same entity.

The results for pronominal expressions are affected by several types of errors. Three errors made by the ideal observer affect both the recall and precision measures. Two appear in example 13, in which the reflexive pronoun *s'* has been analyzed as coreferent with *auxquels* rather than with the inverted subject *15 millions de travaux* and the relative pronoun *que* has been wrongly analysed as coreferent with *les lieux*. We view these two examples as indicators of the lack of attention in the annotators' work.

The precision measure for the majority opinion is further affected by the four incorrect interpretations of pronoun as linked by a description or paraphrase relation evoked in the previous sections.

Finally, three omissions, which affect the majority opinion recall measure, are to be noted; they concern a first person possessive determiner, and two relative pronouns.

6.6. Dates

The expert identified 23 expressions denoting a date which had incomplete descriptive content. These expressions have been well observed by the annotators, with an average recall and precision of 0.78 and 0.90 and, most remarkably a majority opinion recall and precision of 0.91 and 1. The variance value is 0.65. Only two omissions are to be noted: the adverb *hier*, occurring in two different texts, has been overlooked by three of the five annotators. The two other annotators, however, did provide the correct description for this expression.

	a1	a2	a3	a4	a5	av	mo
R	0.91	0.70	0.78	0.65	0.87	0.78	0.91
P	1	0.73	0.95	0.88	0.95	0.90	1
S	0	0.86	0.20	0.25	0.33	0.33	0
O	0	0	0	0	0	0	0
U	1	0.14	0.80	0.75	0.67	0.67	1

Table 6. Results for date expressions.

6.7. Referential relations other than coreference

The referential relations other than coreference the expert observed in the texts have to a large extent not been observed by the five annotators. Furthermore, the annotators used the "set-membership", "part-of" and "unspecified" relations interchangeably. Average recall and precision amount respectively to 0.12 and 0.20, with a variance of 3.39. The majority opinion recall and precision amount to 0.10 and 0.50, respectively, which means that some relations have been observed by the majority. Let us consider these observations.

Out of 39 observation predicates involving referential relations in the key, only 4 have been correctly observed by the majority of annotators.

There is in one of the texts a reference to a specific set of magistrates, and to the set of magistrates in general. The former set is a subset of the other:

18. *Les magistrats auront à leur disposition [...] On n'échappera pas à un besoin de spécialisation croissant des magistrats*
[The magistrates will have at their disposal [...] We won't avoid a growing need for specialization of magistrates.]

In the following sentence, the referent of *les 25 % de la Coface* is a subset of the "actifs" which are to be sold. A relation indicated by the adverb *notamment:*

19. *Aucune décision n'a été prise quant à d'éventuelles cessions d'actifs, et notamment les 25 % de la Coface…*
[No decision has been taken concerning possible transfers of assets, in particular the 25 % of Coface.]

In the following text, the referent of *le premier* is an element of the set denoted by *d'autres pôles*, and there is a relation of type "distinction" between the referent of *d'autres pôles* and that of *Cette annexe*:

20. *Cette annexe parisienne du Palais de justice […] devrait rapidement être suivie d'autres pôles en province. Le premier sur la liste du gouvernement est le pôle corse.*
[This Parisian annex of the Law Courts should be soon followed by other poles in the provinces. The first on the government's list is the Corsican pole.]

Besides these four relations, the annotators also observed in their majority four other relations between referents, but failed to correctly identify the type of this relation. One of these relations is actually the relation between the referent of *Allianz* and the referent of *un nouveau groupe Allianz* which we already discussed (section 6.5). Here are the three other relations.

In the following text, the expert observed a coreference relation between *6.400 mètres carrés* and *quelque 23 mètres carrés par personne* (two ways of denoting the same surface area); the annotators considered this relation either as a description, a set-membership or a part-of relation. The expert also observed an unspecified relation between *moitié moins* and *quelque 23 mètres carrés par personne* (the two expressions denote different surface areas, and it is not the case that the former is part of the latter, the two areas being in two different places); four annotators observed either a set-membership or a part-of relation.

21. *Cet immeuble […] accueillera sur 6.400 mètres carrés, d'ici à la fin de l'année, 274 magistrats et fonctionnaires […]. Ils auront à leur disposition quelque 23 mètres carrés par personne, contre pratiquement moitié moins auparavant au Palais de justice.*
[This building will host on 6400 square meters, by the end of the year, 274 magistrates and civil servants. They will have at their disposal some 23 square meters per person, as opposed to half as much before at the Law Courts.]

Finally, the last relation observed by the annotators is the one between *des logiciels* and *équipements informatiques* which we used to illustrate the set-membership relation in (6). Two annotators correctly viewed this relation as set-membership, two others viewed it as part-of, the last one as identity. One may note in this example the presence of the adverb *notamment* as an indicator of the set-membership relation.

	a1	a2	a3	a4	a5	av	mo
R	0.08	0.15	00.08	0.10	0.18	0.12	0.10
P	0.18	0.21	0.23	0.22	0.17	0.20	0.50
S	0.13	0.19	0.11	0.16	0.26	0.17	0.11
O	0.17	0.28	0.11	0.16	0.37	0.22	0
U	0.70	0.53	0.77	0.67	0.37	0.61	0.89

Table 7. Results for all referential relations other than identity.

To conclude these remarks on the way the annotators observed the referential relations other than identity, it must be noted that they never agreed on a spurious relation, as indicated by the null majority opinion over-generation measure. As majority opinion errors are mostly due to absence of observation (0.89 under-generation), we considered counting as *correct* majority opinion observations cases where two observations were made, together with three absences of observation. With this new criterion, eight observation predicates originally evaluated as *missing* became *correct* (2/8) or *incorrect* (6/8), but seven *spurious* observations were also added. The referential relations other than identity have definitely been badly observed.

7. Discussion

7.1. Operational notions

From the results of the experiment we have presented, we conclude that coreference is an operational notion, with the qualification that what is operational might be only the notion of identity, doubled with the notion of anaphora in the case of paraphrase and description relations, as these two relations may be mistaken for coreference. One also concludes to agreement on the interpretation of temporal expressions. Other notions in our typology have not proved to be operational.

These results confirm those which have been obtained in the development of the Message Understanding Conferences information extraction tasks (Chinchor and Sundheim, 1995). They will have the same consequence for us as the MUC results had for the participants in the evaluation campaign:[10] when its comes to developing natural language processing tools, we can and will only try to handle coreference and temporal expressions, not other relations. The fact that these relations or expressions have been regularly observed by a group of annotators gives us the guarantee that evaluation conditions for our would-be natural language processing system exist and that these conditions are *external* to the system, *i.e.* they have been shown to exist independently of the definition of the system.

7.2. Directions for future experiments

Looking back on the experiment, we feel that the task the annotators had to perform was too complicated to be done reliably. We see some directions for other future experiments, which would aim at assessing inter-subjectivity on some phenomena covered by our typology.

7.2.1. Restrictions on the form of expressions

A first direction is to restrict the typology to specific types of expressions, identified, for instance, on the basis of syntactical information (e.g. pronouns, definite noun phrases). While we focused primarily, in the definition of the different relations to be observed, on the interpretation of expressions, a typology such as the one proposed by Halliday and Hasan (1976) lays stronger focus on the form of the expressions involved in the different relations. This might be more operational, as these added

[10]The MUC tests have led to the definition of a task which was limited to coreference resolution only. See for instance (Hirshman and Chinchor, 1998).

restrictions would sometimes constrain the choice of a relation. To strengthen this hypothesis, we note that the referential relations other than identity which the annotators did observe all involve either expressions which are clearly anaphoric (*le premier*, *moitié moins*) or relations where strong linguistic clues were to be found (the adverb *notamment*, an identity of description, the adjective *autre*). It must also be noted that relations involving clauses or sentences have not been well observed.

Restriction of the observations to a specific type of expressions would also allow one to run a "forced choice" experiment, *i.e.* an experiment in which the annotators would be asked to provide an answer for every element of a fixed set of expressions. In a forced choice experiment, the answer can be negative (e.g. "I see no relation to a contextual expression here") as well as positive. With such a method, one would overcome the difficulty of interpreting an absence of observation either as an omission, or as a positive statement that there is no relation.

7.2.2. Restrictions to some type(s) of relation

A second direction for future experiments is to restrict the observations to a specific relation. The fact that the annotators had many distinct relations to observe probably led them to concentrate better on some of them. We think that if an annotator had only to observe relations of type "set-membership", the operationality of this relation could be shown, especially as there is some evidence, in our experiment, that the annotators overlooked the requirement that a description relation be also present in a set-membership relation.

7.2.3. Restriction to the interpretation of expressions

Finally, we would argue that rather than relating expressions to other contextual expressions, one should maybe focus on the interpretation of the expressions proper. For temporal expressions, the annotators had to give a specific description of the referents; they did not have to say which contextual information they used to identify these referents. It turned out that they agreed on the referents of these expressions. The same method could maybe be applied to all types of referents. Without any contextual information, the referent of an expression may be identified only if the expression is specific enough. A way to make an expression more specific is to add complements (e.g. *the PRESENT president of the United States* vs. *the president OF THE UNITED STATES* vs. *the president*). Such complements, in particular prepositional phrases, actually express a relation between referents (e.g. *two of the five annotators*: "set-membership", *the cathedral of Clermont-Ferrand*: "part-of", *the rent of the building*: "unspecified relation"). A new experiment could consist in asking annotators to provide more specific expressions for a given set of expressions. Presumably, the annotators would use prepositional phrases such as the one proposed above, and chances would be high that they actually refer, with these prepositional phrases, to an object already referred to in the text. They would hence specify a relation, but they would not need to concern with the type of the relation between the two referents. [11]

Different problems are addressed by our typology: the incomplete descriptive content of some expressions, the interpretation of these expressions proper (*i.e.* answering the question "what does these expressions mean or refer to?"), and the way this interpretation is performed (*i.e.* the way the expressions are related to their context). We believe the method we have superficially outlined here would allow one to keep these different aspects separate, the result of the interpretation on the one hand, the process which led to this interpretation on the other.

8. Conclusion

After having defined a typology of anaphoric and referential relations, we presented an experiment aiming at assessing that this typology was operational. Given three texts, five annotators were asked to observe and describe the phenomena in these texts which were presumably covered by our typology. Results show that coreference is an operational notion and that temporal expressions are correctly interpreted. However, there is strong doubt on the operationality of the other relations we have defined. Ultimately, this experiment shows that, when questioned, a linguist's descriptions of the data may fail to pass the inter-subjectivity test.

9. References

Chinchor, N. and B. Sundheim, 1995. Message Understanding Conference (MUC) Tests of Discourse Processing. In *Proceedings of the AAAI Spring Symposium on Empirical Methods in Discourse Interpretation and Generation.* Stanford.

Di Eugenio, B., 2000. On the Usage of Kappa to Evaluate Agreement on Coding Tasks. In *Proceedings of the Second International Conference on Language Resources and Evaluation.* Athens, Greece: ELRA.

Dowty, D., R. Wall, and S. Peters, 1981. *Introduction to Montague Semantics.* Dordrecht, Holland: D. Reidel Publishing Company.

Fligelstone, S., 1990. *A description of the conventions used in the Lancaster Anaphoric Treebank Scheme.* Lancaster: Department of Linguistics and Modern English Language.

Halliday, M. A. K. and R. Hasan, 1976. *Cohesion in English*. Longman.

Hirshman, L. and N. Chinchor, 1998. MUC-7 Co reference Task Definition. Version 3.0. In *Proceedings of the Seventh Message Understanding Conference (MUC-7).* hhtp://www.muc.saic.com/: Science Applications International Corporation.

Karttunen, L., 1976. Discourse Referents. In McCawley, J.-D., ed. *Syntax and Semantics 7: Notes from he Linguistic Underground*. New York: Academic Press.

Ranta, A., 1994. *Type-Theoretical Grammar*. Oxford University Press.

Trouilleux, F., E. Gaussier, G. Bès and A. Zaenen, 2000. Coreference Resolution Evaluation Based on Descriptive Specificity. In *Proceedings of the Second International Conference on Language Resources and Evaluation.* Athens, Greece: ELRA.

[11] This type of annotation, restricted to "inferable 'of-' complements", has been used in the annotation of the Lancaster Anaphoric Treebank (Fligelstone, 1990). We do not know to which extent it proved operational.