Taking Genre into account when Analyzing Conceptual Relation Patterns

Anne Condamines

Taking *genre* into account when analysing conceptual relation patterns

Anne Condamines[1]

Abstract

This paper uses a corpus study to investigate the influence of text genre on the frequency and semantic interpretation of certain pattern/concept relations. In linking pattern/concept relations to text genre, this study identifies three types of dependency: weak dependency, where the relation appears in almost any kind of text; complete dependency, where it is strongly linked to a particular text or group of related texts; and dependency in terms of text genre. The particular examples that form the basis of the study are meronymic *chez*, which is found to have a significant dependency in didactic texts in the natural sciences; *comme* as a marker of hypernymy and co-hyponymy, which has a weaker, but observable dependency in technical and didactic genres; nominal anaphora involving hypernyms, where no consistent conclusions can be reached; and meronymic *avec*, where the significant factor is shown to be communicative objective rather than domain (subject matter). I discuss the relevance of such studies to Natural Language Processing, and indicate the potential for further research.

1. Introduction

Within corpus linguistics, genre is one of the most crucial but also one of the most difficult problems to be tackled. Its importance derives from the possibility of categorising texts as a way of detecting and explaining variation in large corpora. Taking genre into account within corpus linguistics, therefore, mainly implies a quantitative approach that is intended to identify, rapidly, those linguistic regularities which, in turn, make it possible to place different texts within the same class. But genre may also be used to construct more detailed, qualitative descriptions. Such qualitative descriptions are particularly necessary when the aim is not only to categorise texts but also to identify those regularities which depend on a purpose that is clearly identified at the outset of the study. This is

[1] Cognition, Langues, Langage, Ergonomie, Equipe de Recherche en Syntaxe et Sémantique, (CLLE-ERSS), UMR 5263, CNRS et Université Toulouse, Maison de la Recherche, 5 allées Antonio Machado, F-31055 Toulouse Cedex, France.
Correspondence to: Anne Condamines, *e-mail*: anne.condamines@univ-tlse2.fr

particularly true when one is elaborating terminologies or ontologies based on texts. The importance of the task of elaborating terminologies will be presented below but what should be kept in mind is that it is linked to the possibility of constructing a conceptual network from texts using conceptual relation patterns. Genre plays a role in this issue, but the way in which it does so depends on the pattern: sometimes what prevails is simply the quantitative point of view, when a particular pair of pattern and relation is more frequently found in texts belonging to a certain genre. Yet, sometimes, as shown by this study, understanding the mechanisms of a given pattern calls for a more fine-grained analysis – that is, a qualitative one.

In this paper, I present four conceptual relation patterns and emphasise the ways in which genre can be taken into account from both a quantitative and a qualitative point of view. At this stage in our thinking, the aim of the study is not to propose a complete method for taking genre into account in every analysis of a pattern, and even less a method for taking into account genre for learning ontology from texts. The aim of the paper is, rather, to show that, for some potential relational patterns, the nature of texts plays an important role in the possibility that the patterns are useful for identifying a certain relation. But the study has another ambition. Whenever possible, I try to give an explanation for the link between the possibility that the pattern identifies a relation, and the specificity of relevant texts, mainly in terms of the contextual features of the writing. Thus, I try to explain, at least partially, the semantics of patterns in terms of extra-linguistic elements. One of the extra-linguistic features examined concerns the specialised nature of texts. Such a feature seems relevant since terminologies as well as ontologies are more often built for specialised domains. Moreover, intuitively, we can perceive that, in specialised fields, conceptual relations are more present than in non specialised ones and, therefore, that patterns are more often used for marking conceptual relations than for other kinds of objectives. But it is well known that it may be very difficult to define what is a specialised domain. This paper presents some cases of patterns in which the role of domain is clear and other ones in which it is much more difficult to identify which extra-linguistic features are relevant.

2. Presentation of the problem

The results given in this paper concern both genre and conceptual relation patterns. I present the two phenomena in this section.

2.1 Conceptual relation patterns

A conceptual relation pattern may be defined as any linguistic form – whether morphological, lexical or discursive – that can be used to identify a segment of text that can be understood as describing a relation between two nouns. For example, in:

(1) *Tous les invertébrés sauf les insectes se rangent dans les vers.*

[All invertebrates except insects are worms.]

the pattern [tous les N1 sauf N2] ([all N1 except N2]) suggests that there is a hypernymic relation between *vers* ('worms') and *invertébrés* ('invertebrates').

Some patterns were studied by linguists twenty or thirty years ago. Lyons (1977) spoke of 'formulae' and Cruse (1986) of 'diagnostic frames'. More recently, several teams of linguists have tried to understand how these discourse structures may be used to identify conceptual relations. Meyer (2001: 290) believes that there are three kinds of conceptual relation patterns (which he calls 'knowledge patterns'). These are lexical patterns ('involving one or more specific lexical items'), grammatical patterns and paralinguistic patterns ('which include punctuation, as well as various elements of the general structure of a text'). He states that these patterns are, 'complex in their nature, and in the way they can be realised in text': they are sometimes unpredictable, polysemic, and/or domain-dependent. We will see these three features again in the description of patterns below.

These portions of text have become particularly interesting for computer scientists seeking to build ontologies from texts. Neches *et al*. (1991) define an ontology as, 'the basic terms and relations comprising the vocabulary of a topic area, as well as the rules for combining terms and relations to define extensions to the vocabulary'.

Since they are used for specific applications (information retrieval, information extraction, translation…), ontologies are now frequently 'learnt' from texts. It is very difficult in this linguistics-orientated paper to present all the work done in the domain of learning ontology from texts, which has proved to be very fruitful (Aussenac-Gilles *et al*., 2000; Cimiano *et al*., 2004; Maedche *et al*., 2001). Suffice it to say that the main purpose of such studies, therefore, is to build knowledge representations which are close to what occurs in discourse, since they are very often used for identifying information in texts. Elaborating an ontology from a corpus requires the construction of a knowledge model of this corpus in a relational form (a conceptual network). Such a network is made up of nodes joined by arcs, each of which is labelled by a term.

Such a conceptual network underlies several representational languages such as conceptual graphs and terminological logic. What is interesting about such a knowledge representation is that it makes it possible to evaluate consistency and so is useable for automatic reasoning.

Computer scientists are highly interested in conceptual relation patterns for designing tools for the automatic detection of conceptual

relations in corpora; indeed, many descriptions of such patterns have been made by computer scientists, linguists and terminologists (Green *et al*., 2002; L'Homme, 2004; Meyer, 2001; Marshman *et al*., 2002; Pearson, 1998). Most often, the relations under investigation are hypernymy (Hearst, 1992), meronymy and causality (Cabré *et al*., 1997).

However, most of these studies assume that language is homogeneous: what is investigated is how a certain relation is expressed within texts (the approach is thus an introspective, top-down one). Very few studies focus on bottom-up approaches – from texts to relations. Such approaches take into account variations between texts since relational patterns do not appear in all texts in the same way, and it is necessary to understand why. This paper deals with the following question: how can the nature of a corpus play a role within the conceptual relation/pattern pair – in other words, what kind of link can there be between patterns and corpora?

2.2 What kinds of corpus/pattern dependencies can be identified?

Three kinds of dependencies between corpora and patterns may be identified (Condamines, 2002): weak dependency, complete dependency and dependency in terms of text genre.

2.2.1 Weak dependency

Certain patterns are likely to appear in any text. This is the case for hypernymic or meronymic patterns such as:

(2) *The house comprises a hall*, *two reception rooms*, *and a dining room.*

[N1 comprises N2, (N3) and N4] or [all N1 except N2] (as in 1).

These patterns are the ones usually identified by introspection (a top-down approach). Even if the relation/pattern pair seems very strong, certain difficulties may arise. First of all, these patterns may be polysemic. For example, *comme* in French (see Section 3.1.2) may be associated either with a hypernymic relation or a comparison relation. In:

(3) *Un département comme la Seine bénéficie à la fois d'arrivées d'enfants et de scolaires.*

[A department such as the Seine benefits from the arrival of both children and students.]

it is possible to identify a hypernymic relation between *département* and *la Seine*. But the same pattern is not used to express the same relation in:

(4) *On comprend que les lycées professionnels et d'enseignement général, comme l'Université, soient très peu tournés vers les formations scientifiques et technologiques.*

[We can understand that professional lycées and general lycées, as well as universities, are not much geared towards scientific and technological training programmes.]

Another difficulty with patterns (which is true of all patterns) is that determining whether the speaker using conceptual relation patterns is expressing his or her own point of view, or if he or she is subscribing to that of a group of speakers, may not be straightforward. With the goal of constructing ontologies in mind, only the second situation needs to be modelled since, generally speaking, ontologies are supposed to represent shared knowledge since such models must be acceptable and reusable for collective tasks. Thus, text genre may play a role in the interpretation of the pattern: it is probably more difficult to identify a collectively understood relation if the pattern appears in a novel rather than in a technical handbook.

2.2.2 Complete dependency

In some corpora, certain specific structures act as patterns for given relations. These structures are understandable as relational patterns, but they are unpredictable in terms of their structure (Meyer, 2001). I found such a structure in a corpus from Électricité de France (EDF) dealing with specifications and the writing of documents in computer science. It is a corpus of approximately 45,100 words, written by various experts but with the same purpose. In this corpus, the pattern for the relational condition was:

[(phase, étape) ou déverbal + … + (*lorsque*, *dès que*) + V au passif]
[(phase, stage) or nominalisation) + … + (*when*, *as soon as*) + passive V]

(5) *La phase d'intégration du composant peut commencer lorsque l'ensemble des éléments logiciels ont été codés*.

[The component integration phase may begin when all the software elements have been coded.]

This example should be understood to mean that 'software elements' must first have been coded for the 'component integration phase' to begin.

Therefore, the relation expressed looks like a temporal one, but must be understood as a conditional one.

These patterns are very difficult to identify; indeed, it may be that each case necessitates a new *ad hoc* method, since a fine-grained analysis is required (Condamines and Rebeyrolle, 2001). Thus, at present, these patterns cannot be identified automatically.

2.2.3 Dependency in terms of text genre

Sometimes regularities of expression are not specific to a corpus from a single source but, rather, to a corpus constituted of texts with similar extra-linguistic and linguistic characteristics, (i.e., texts belonging to the same genre). In the rest of this paper I will examine specifically this kind of dependency.

Bahtia (1993: 13) states that:

> Taking genre, after Swales […] it is a recognizable communicative event characterised by a set of communicative purpose(s) identified and mutually understood by the members of the professional or academic community in which it regularly occurs. Most often it is highly structured and conventionalised with constraints on allowable contributions in terms of their intent, positioning, form and function value.

In other words, text genre takes into account both linguistic and extra-linguistic elements in order to explain how members with the same communicative purpose may communicate. The use of such a notion supposes that there is a co-variance of both linguistic and extra-linguistic elements. This co-variance would be learnt during language acquisition and genre would constitute a kind of prescriptive frame for a reader/listener (Todorov, 1984). Accordingly, a genre would be identified (newspaper article, scientific article, administrative letter, advertisements, *etc*.) and then the reader/listener would expect to observe certain linguistic features. But, from a descriptive perspective, the identification of all relevant linguistic features may be very difficult: more than one linguistic level may be concerned (lexical, syntactic, morphological) and speakers are not always aware of them. From a semantic point of view, some regularities concerning lexical variations are easy to see; for example, in the domain of medicine, nouns used in specialised handbooks are not the same as the ones used in popular journals (e.g., *oncology* versus *cancer research*), (see, for example, Rogers, 2000). However, such regularities are not as clear when, for instance, semantic variation concerns prepositions. Consequently, describing such variations requires the analysis of large numbers of occurrences from different corpora, as in this paper.

Numerous researchers have studied this notion of genre, such as Bakhtine in the Russian movement (Todorov, 1984), or Bahtia (1993), Biber *et al*. (1999), Firth (1969), Halliday and Hassan (1985) and Swales (1990) in the English-speaking tradition. Genre may be a means of explaining variation, and using it can be useful in evaluating and explaining variation even if certain problems may be encountered.

First, genre involves the variation of both linguistic and extra-linguistic features. Biber (1988) uses two words in order to distinguish these two kinds of features:

> I use the term 'genre' to refer to text categorizations made on the basis of external criteria relating to author/speaker purpose (p. 68)
>
> I use the term 'text type' on the other hand, to refer to groupings of texts that are similar with respect to their linguistic form, irrespective of genre categories (p. 70)

Generally, corpora are first constituted according to extra-linguistic criteria and then their linguistic features are observed. This method may be regarded as problematic when corpora are constituted on the basis of an intuitive categorisation of genre and then examined from a linguistic point of view. But it is very difficult to conceive of another method, specifically when the features that are examined belong to semantics. In fact, typically, the hypothesis concerning the link between linguistic and extra-linguistic features is refined during the study, and the initial hypothesis concerning the relevance of some genres may be re-evaluated. The same method is used in the studies presented in this paper.

Secondly, the point of view adopted may entail differences in the method of classifying a text into one genre or another. For example, if you want to study a particular preposition, you may identify a type of genre organisation relevant to your point of view, but it might not be relevant to another study. This is the true of the patterns studied below.

Thirdly, a text may belong to different genres. For example, a letter may be quoted within an article, thus raising the problem of embedded genres.

In the studies I present here, another difficulty has appeared, as signalled in the introduction: it is sometimes difficult to define the relevant level of categorisation for the pattern identified. Only fine analysis may allow us to say if taking domain into account is sufficient for explaining the meaning of patterns or if it is necessary to delineate in more detail the features of the situation (see below). In spite of these difficulties, the notion of genre may be very useful in describing linguistic phenomena and this is particularly true with certain conceptual relation patterns. The effect of genre on patterns can be examined from two points of view: a quantitative and a qualitative one.

The quantitative point of view is generally the most commonly used in corpus linguistics. The reason for this is rather obvious: it is easy to

compare the distribution of forms between various parts of a corpus or between various corpora. Furthermore, this approach has recently benefitted from natural language processing, see, for example, Biber's studies (e.g., Biber *et al*., 1999). On the other hand, the qualitative perspective is more difficult to envisage because it is necessary to take not only the forms into account, but also the meaning of these forms (McEnery and Wilson, 2004).

Concerning relational patterns, what is important is not simply that such-and-such a pattern occurs, but that it occurs with a certain meaning, (i.e., it can be used to detect a specific relation). For example, *comme* will be treated as a productive pattern for the hypernymic relation in connection with a given text genre only if it is possible to prove that, in texts of this genre, the pattern is used much more frequently in the hypernymic relation than in other kinds of roles. In other words, what is studied is the relation/pattern/genre triplet and not only the pattern/genre pair. Thus, in order to study the role of genre in the relation/pattern pair, it is necessary to examine all the occurrences of a putative pattern within a corpus and to interpret them in terms of whether they belong to a certain relation or not. And even when the study is completed, it is obvious that some cases must be examined in more detail and that certain variations occur according to the genre. Thus, three kinds of situations may be identified concerning the role of genre in the relation/pattern pair:

- in some cases, genre influences the relation/pattern pair only in terms of quantitative results;
- in other cases, it influences the pair in terms of qualitative results; and,
- in still other cases, it influences the pair in terms of both quantitative and qualitative results.

In the next section I will explain and exemplify these three situations.

3. Case studies

I shall describe four examples of patterns associated with a relation and variations according to genre: three French prepositions (*avec*, *comme* and *chez*) and a discourse structure (the demonstrative nominal anaphor). For each pattern, a contrastive corpus has been constructed and each occurrence of the pattern identified, quantified and analysed. It is important to note that, at this first stage of the study, the component texts of the corpora are chosen without a very precise initial hypothesis (and also according to the availability of texts in our bank of texts base or on the web). The real hypothesis concerning the way patterns are linked both to a relation and to a text (belonging to a genre) may only be built when the first results are obtained. In such semantic studies, the aim of the first study is to

experiment with an imprecise semantic hypothesis in order to build a solid hypothesis.

3.1 Quantitative differences in terms of genre: the case of *chez* and *comme*

3.1.1 The case of *chez* and the meronymic relation

In some cases, the preposition *chez* occurs in sentences where a meronymic relation can be identified:

(6) *Chez les colobinés*, *le nez fait saillie sur la lèvre supérieure.*

[Among the colobines, the nose juts out over the upper lip.]
(there is a meronymic relation here between 'nose' and 'colobines')

I have tried to identify whether some kinds of texts (i.e., texts belonging to a certain genre) are more susceptible to this interpretation (Condamines, 2000). It seems that this is the case with texts of didactic origin dealing with natural science.

A corpus was constructed with four categories of texts:

- A set of articles from the *Encyclopaedia Universalis*[2] dealing with natural science, more precisely, Bacteria, Algae, Fruit flies and Insects (selected at random from the classification proposed by the EU). This corpus contains 24,770 words and is called EUbio.
- Two biology textbooks (written by two different authors), selected from the French text database Frantext.[3] This corpus contains 56,880 words and is called Manuelsbio.
- Sixty-five articles from the French newspaper *Le Monde* (CD-Rom 1997–8) dealing with natural science, more specifically Insects, Animal illness and Fish (selected at random from the classification proposed by *Le Monde*). This corpus contains 49,440 words and is called LMbio.
- A set of articles from the *Encyclopaedia Universalis* in the field of culture, covering the following subject areas: Positivism, Expressionism, Phenomenology, Psychoanalysis, Psychiatry, Cubism, Romanticism (–natural science and +didactic). This corpus contains 79,700 words and is called EUculture.

The features of this corpus can be seen under Table 1.

[2] See: http://www.universalis-edu.com
[3] See: http://atilf.atilf.fr/frantext.hym

=Insert Table 1 about here=
=Insert Table 2 about here=
=Insert Figure 1 about here=

The 456 occurrences of *chez* have been examined and divided into meronymic versus non-meronymic relations. The results of this quantitative study are presented in Table 2 and Figure 1. This table and diagram show that there is a clear difference between EUbio and Manuelsbio on the one hand, and EUculture and LMbio on the other. For EUbio and Manuelsbio, the meronymic relation is relevant for around half of the occurrences, while in the case of EUculture and LMbio, only a few occurrences may be interpreted as involving a meronymic relation. I have also analysed the results of Table 2 from a statistical point of view by calculating chi square in order to examine whether the factor corpus has an influence on the factor type of *chez* (meronymic versus non meronymic). It reveals that corpus influences significantly the use of meronymic *chez*: $\chi^2 = 62.74$, $p < .005$). These results confirm the initial hypothesis: the association *chez*/meronymy depends on the nature of the text in which *chez* appears; more precisely, in texts belonging to the natural science and didactic genres, *chez* can be used as a means of identifying a meronymic relation. But it is important to understand how this 'pattern' is indicative of such a relation. As a matter of fact, it is not true to say that *chez* elicits a meronymic interpretation: nobody spontaneously produces this preposition in a meronymic pattern, and its etymology (from the Latin *casa*, 'house') is in no way linked to such an interpretation. A detailed analysis shows that this preposition occurs in structures in topic position (at the beginning, the middle or the end of the sentence), that is, in structures that introduce a new referent into the discourse.[4] In didactic natural science texts, what is often said about these new referents (animals or plants) has to do with their anatomy or composition. It is interesting to note that this meronymic information frequently appears at the beginning of the articles in the *Encyclopaedia Universalis* and that all of them follow the same model: they first present the anatomy and then the life cycle (living conditions, feeding, *etc*.) as in:

(7) *Chez les Algues unicellulaires mobiles*, *la motilité est due à un appareil cinétique comportant des flagelles.*

[In mobile unicellular algae, motivity is due to a kinetic system comprising flagellae.]

Quantitative results show that, within didactic natural science texts, anatomy-linked information is found more often than other kinds of information.

[4] Note that for the same topic position, it is not *chez* but *dans* that is used for minerals; *chez* is exclusively used for living entities.

In the other texts, things are of course very different. As in the natural science corpus, *chez* is used to focus on an object, but the rest of the sentence never involves meronymy:

(8) *Chez Toulouse-Lautrec se profilent toutes les possibilités expressives de la ligne.*

[With Toulouse-Lautrec, all the expressive possibilities of the line can be detected.]

So, in discourse, a meronymic relation may appear in certain texts but, from a linguistic point of view, it is not really true that *chez* is a diagnostic marker for this relation. On the other hand, from a computational point of view, *chez* may be used to identify structures where a meronymic relation occurs in didactic natural science texts. In about half of the occurrences, there is a meronymic relation because this kind of information is essential within natural science. Consequently, *chez* can be used as a clue to detect a meronymic relation, but it does not necessarily provide an interpretation for this relation. This case is very interesting because it shows that linguistic and computational approaches are not always equivalent.

Finally, it should be noted that the identification of the specialised domain (in this case, a precise domain: natural science) is not sufficient to characterise relevant texts. It is also necessary to take into account the didactic perspective, namely, that texts are written for non-specialist readers and with the aim of giving explanations about the domain.

3.1.2 The case of *comme* as marker of a relational pattern

Comme is sometimes proposed as a marker of a potential hypernymic pattern, as in:

(9) *Une fleur comme la rose se vend particulièrement bien*.

[A flower such as the rose sells very well.]
('flower' is a hypernym of 'rose')

Our study of *comme* is still in progress, but I can present the first results. First of all, two important points should be noted concerning the way the hypernymy operates:

(*i*) In some cases, the direction of the relation is inverted: it is not the first noun which is hypernymic but the second one:

(10) *La rose comme fleur de décoration est très appréciée.*

[The rose, as a decorative flower, is very much appreciated.]
('decorative flower' is a hypernym of 'rose')

(*ii*) In some other cases, the relation between the two nouns is not hypernymy but rather co-hyponymy as in:

(11) *La rose comme l'orchidée sont très appréciées des clients.*

[*The rose as well as the orchid are very much appreciated by customers.*]
('rose' and 'orchid' are both hyponyms of flower, so they are co-hyponyms)

In the study, I have not tried to take into account the precise nature of the relation but to distinguish sentences in which *comme* may be considered as marking a relational pattern (cases of hypernymy or co-hyponymy, these two cases being both interesting for us) from the ones in which it plays another role. These other potential roles are well known and well described (Fuchs and Le Goffic, 2005): *comme* may introduce an exclamation (*comme elle est belle!* [*how beautiful she is!*]), or a subordinate clause (*elle entra comme le rideau se levait* [*she came in as the curtain was rising*]) or an adverb (*comme souvent*, *Paul est en retard*, [*as often*, *Paul is late*]) *etc*. When *comme* plays a relational role, the presence of two nouns is necessary, one in the context to the right and one to the left. However, in this study, what I wish to analyse is how the nature of texts intervenes in the interpretation of the preposition, even in precedence to the nature of the syntactic context in which it appears. Several factors justify this perspective.

From a Natural Language Processing point of view, it is sometimes very difficult to identify automatically the correct nouns linked by a preposition: the two closest nouns are not always the ones concerned. Furthermore, what the study shows is that in some cases, *comme* has to be preceded by a verb in order to mark a relational pattern:

(12) *Le faisceau se comporte comme un conducteur unique de rayon équivalent*.

[The cluster behaves like a single conductor of equivalent radius.]

Like many linguists (e.g., Biber, 1988; Todorov, 1984), I think that text genre constitutes a way of comprehending the world. Thus, some particular situations may contribute to preparing speakers to select particular meanings even before taking into account the syntactic context. For example, we can easily imagine that *comme* as an introducer of an exclamation would not be very frequent in technical documents. But it is much more difficult to decide *a priori* if the preposition will be frequent or not in a relational pattern.

The hypothesis with *comme* is that it marks a relational pattern (hypernymic or co-hyponymic) more frequently in specialised texts than in other genres (here, journalistic ones).

As for the other studies, a corpus was constituted with two categories of texts:

(*i*) specialised texts

- a mechanics' textbook, (Mécanique): 105,700 words;
- a technical handbook, *Guide de Planification*, issued by EDF, (GDP): 165,700 words; and,
- an atlas – *Atlas de la France scolaire*, (Atlas): 60,000 words.

(*ii*) non-specialised texts

- a journalistic corpus: some articles from *Le Monde Diplomatique*, selected at random, from 1989 (Monde Diplo.): 170,200 words.

All 573 occurrences of *comme* have been studied and their potential for use in a relational pattern has been examined.

As suggested by several authors, including Fuchs and Le Goffic (2005), the main meaning of *comme* is a comparative one. Thus, when *comme* is used for linking two nouns, we can suppose that these two nouns are compared in one way or another. The notion of comparison may perhaps explain how *comme* can play the role either of a hyponymic pattern or a co-hyponymic one. Indeed, one might say that there is a kind of analogy between a hypernym and its hyponyms on the one hand, and between a hyponym and its co-hyponyms on the other, since it is generally considered that, in the two cases, there is just a single feature which distinguish the nouns. In the case of hypernymy, a feature is added to the hyponym and in case of co-hyponymy, there is one feature which distinguishes it from all the siblings (co-hyponyms).

However, where a difficulty can arise is in differentiating cases in which *comme* expresses merely a comparison, that is, a point of view of the author, as in:

(13) *D'autres machine pneumatiques*, *tenues en main comme un gros crayon.* (Mécanique)

[Other pneumatic machines, held in the hand like a large pencil.]
('pneumatic machines' is compared to 'large pencil')

from the ones where the comparison corresponds to a hypernym or a co-hyponym (as in (14)), that is, a piece of well-established knowledge:

(14) *On doit veiller à ce que le métal ne soit pas recouvert par une couche d'oxydation agissant comme un isolant.* (Mécanique)

[Care should be taken that the metal is not covered by an oxidised layer acting as an insulant.]
('insulant' is a hypernym of 'oxidised layer')

Table 3 and Figure 2 show the first results of the study. *Comme* appears more frequently within specialised corpora than within journalistic ones, and relational *comme* is more frequent within specialised corpora than within journalistic ones. From a statistical point of view, χ^2 built from Table 3 shows that the influence of the factor corpus on the factor type of *comme* (relational versus non relational) was on the threshold of significance in the chi square analysis ($\chi^2 = 6.64$, $.10 < p < .05$).

=Insert Table 3 about here=
=Insert Figure 2 about here=

Thus, the initial hypothesis seems to be confirmed: *comme* is more often used in a relational pattern in specialised corpora than in journalistic ones, even if the hypothesis, from a statistical point of view, is not very strongly confirmed. Two points should be noted. First, there is a strong case for analysing whether or not specialised texts have other common features: there are two technical manuals and an atlas – that is, three types of texts with a high degree of didacticity. It is possible that this feature plays a central role in the meaning of *comme*. Secondly, it could be considered that *Le Monde Diplomatique* belongs to a specialised genre: that of geopolitics. So the problem of delimiting domains appears as a difficulty when considering domain as a sufficient feature for characterising relevant texts.

3.2 Qualitative differences according to genre: the case of nominal anaphora with the hypernymic relation

It is well known that, in some cases, there is a hypernymic relation between a head noun within a nominal anaphor and its antecedent (Cornish, 1986):

(15) *A bomb exploded yesterday in* a car *belonging to a well-known magistrate*. *No-one was hurt*, *but* the vehicle *was completely destroyed.*
(example (24b) in Cornish, (1986: 20))

However, the relation may also be of a different nature. For example, the anaphoric noun may be the nominalised form of a verb (Cornish, 1986: 27–8):

(16) *Ce guide* propose *des plans types pour la documentation de la réalisation des logiciels scientifiques et techniques*. *Il appartient ensuite à chaque projet d'adapter cette* proposition […] .

[This guide proposes models for documenting the development of scientific or technical software. Each project can then adapt *this proposal* (…) .]

At the beginning of my study of nominal anaphora, my hypothesis was that this hypernymic relation would be very frequent in specialised handbooks (Condamines, 2005). Therefore, as with the other studies, I have constructed a corpus of texts belonging to five genres:[5]

- A scientific handbook: *Précis de Géomorphologie* (Geomorphology Handbook), written by various authors (GEO): 206,700 words;
- A technical handbook: *Guide de Planification* (Handbook for Planning), written by various authors (GDP): 148,100 words;
- Another technical handbook: *Méthodes et Outils de Génie Logiciel pour l'Informatique Scientifique* (Software Engineering Methods and Tools for Scientific Computing), written by various authors (Mouglis): 45,100 words;
- Some articles from the newspaper *Le Monde Diplomatique*, from 1989 (LMD): 110,700 words; and,
- A novel by Maupassant, *Bel Ami*. (Bel A.): 170,200 words.

Only demonstrative anaphoric nouns were examined (although the most frequent determiner was the definite article); this was a means of reducing the number of sentences to be analysed (Apothéloz and Reichler-Béguelin, 1999).

First, it should be noted that it is not easy to identify what kind of relation is involved between an anaphoric noun and its antecedent, especially within specialised corpora. Secondly, and this is a very important point, for many anaphora, it is impossible to identify an antecedent noun phrase. This is what I have called 'cases of substitution':

(17) *La pression atmosphérique avait été évaluée il y a une trentaine d'années à 1.12 de celle de l'atmosphère terrestre*. *On a réduit* cette appréciation *car Mariner 4 a trouvé qu'elle équivalait à 6 millibars* […] . (GEO)

[Around thirty years ago the atmospheric pressure was evaluated at 1.12 of that of the Earth. This estimate has been reduced as Mariner 4 found it to be 6 millibars (…) .]

[5] Note that my aim is not to study, from an NLP perspective, how anaphora may be resolved automatically, but to analyse whether genre plays a role in the kind of anaphora which are present in the text.

It is not easy to identify an antecedent noun phrase for *estimate*: the antecedent may be the first sentence taken as a whole. Table 4 and Figure 3 present the results concerning the distribution of the 1,339 demonstrative anaphora between hypernymic and non-hypernymic relations.

=Insert Table 4 about here=
=Insert Figure 3 about here=

The results are very clear: the frequency of hypernymic relations within handbooks compared with other genres is not very high, with the exception of Mouglis (at 60 percent). The statistical analysis shows that the factor corpus has a significant influence on the factor type of anaphora (hypernymic versus non-hypernymic) ($\chi^2 = 63.025$, $p < .005$). From this it may be deduced that the nature of corpora plays a role in the type of anaphora observed (specialised texts favour hypernymic anaphora), but it is difficult to conclude that anaphora is a useful pattern for identifying hypernymic relations since hypernymic anaphora are not very frequent whatever the text genre may be.

In the case of the Mouglis data, a more detailed analysis would be required. Mouglis seems to comprise the more technical texts among the five constitutive genres. But it is not certain that this is the only feature that would explain this result. It will be necessary to continue the analysis by studying other technical texts and to compare the results. This is one of our perspectives.

Nevertheless, it is necessary to continue the analysis, particularly by examining instances of substitution. In these occurrences, head nouns can be interpreted either as classifiers (i.e., they may constitute a head for a taxonomy) or non-classifiers. Several characteristics allow us to suggest that these head nouns are more often classifiers in handbooks (technical handbooks at least) than in texts of a different genre.

First of all, certain nouns which look like non-classifiers – when taken out of context – must be interpreted as classifiers within the discourse, as in:

(18) *Quatre types de* responsabilités *sont associés au Plan de Gestion de Configuration*:
La rédaction et l'évolution du PGC.
Cette responsabilité *incombe au Responsable Assurance Qualité de niveau 1 en collaboration avec le Chef de projet*. (Mouglis)

[Four kinds of *responsibilities* are associated with the Configuration Management Plan: The writing and evolution of the CMP.
This responsibility falls to the Level 1 Quality Insurance Manager in collaboration with the Project Leader.]

Taken out of context, *responsibility* looks like a non-classifier, but within this particular example and, more generally within this corpus, it is a classifier.

Secondly, in handbooks, several nouns are used sometimes as substitutes and sometimes as hypernyms. For example, *activity* is a hypernym in Example (19), where 'software test' is the hyponym, and a substitute in Example (20), where no noun can be regarded as a hypernym of *activity*).

(19) Test du logiciel. Cette activité *consiste à exécuter les procédures de tests spécifiées pour qualifier le logiciel*. (Mouglis)

[Software test. *This activity* consists in carrying out the test procedures specified in order to qualify the software.]

(20) *Les unités de configuration à modifier doivent être identifiées afin de pouvoir séquentialiser les modifications à apporter à un même composant*. *Le résultat de* cette activité *est la constitution du dossier de modifications*. (Mouglis)

[The configuration units to be modified must be identified in order to sequentialise the modifications to be applied to a single component. The result of *this activity* is the drawing up of the modifications file.]

There is no example of this phenomenon in the other corpora under investigation.

Thirdly, most of the head nouns which are either hypernyms or substitutes in technical handbooks occur as heads of several compound nouns. This is the case with *activity*:

Activité	*de conception* (*design activity*)
	de codage (*coding activity*)
	de vérification (*verification activity*)
	…

There are, therefore, several factors in favour of considering substitute nouns in technical handbooks (perhaps in handbooks in general) as hypernyms for the field covered by the corpus. This means that, even if the nominal anaphor cannot be considered to involve a hypernymic pattern, it can be considered as a hypernym marker: it is not the relation itself which is identified, but only one element of this relation – hence the hypernym. Consequently, other patterns have to be used in order to identify hyponyms as being related to hypernyms, which are highlighted by the nominal anaphora pattern.

3.3 Quantitative and qualitative differences according to genre: the case of *avec* and the meronymic relation

All the authors who have analysed the functioning of *avec* (Cadiot, 1997; Mari, 2003) have noticed that, in certain sentences, the preposition may be associated with a meronymic relation: *une robe avec des dentelles* (*a dress with lace*). I have tried to evaluate the influence of certain text genres on such an interpretation.

The corpus was constructed using texts belonging to five genres:

- A Zola novel, *Germinal*, (Germinal): 210,000 words;
- A scientific handbook, *Manuel de Géomorphologie* ('Geomorphology handbook'), (GEO): 206,700 words;
- A toy catalogue, *Catalogue de jouets Leclerc*, (Toy catalogue): 93,000 words;
- A set of property adverts collected from three different websites, (Small-ads): 22,600 words; and,
- A set of itinerary descriptions – a corpus constituted for the purpose of a psycholinguistic study, (Itineraries): 48,000 words.

As for the previous studies, all occurrences of *avec* were examined in order to divide them into meronymic and non-meronymic occurrences. Here are some examples of meronymic occurrences:

(21) *Porteur évolutif avec canne*, *repose-pieds et ceinture de sécurité amovible.* (Catalogue)

[Adaptable carrier with cane, footrest and removable safety belt]

(22) *A l'étage*: *3 chambres avec placards.* (Petites annonces)

[First floor: 3 bedrooms with wardrobes.]

(23) *Vous la repérerez grâce à une église avec une coupole.* (Itinéraires)

[You will find it thanks to a church with a dome.]

And an example of a non-meronymic occurrence:

(24) *Maison rénovée avec soin*. (Petites annonces)

[House renovated with care.]

Table 5 details the results of this quantitative study.

=Insert Table 5 about here=

From a statistical point of view, results obtained by calculating chi square show that the factor corpus has a significant influence on the factor type of *avec*, meronymic versus non-meronymic, ($\chi^2 = 100.87$, $p < .005$). On the basis of frequencies alone, two groups of corpora can be identified. In the first one (Germinal and GEO) the number of occurrences of meronymic *avec* is very low. On the other hand, in the second (toy catalogue, property adverts, itineraries), the number of occurrences of meronymic *avec* is very high. This observation warrants the claim that *avec* (or more exactly, 'déterminant nom avec déterminant nom' [determiner Noun with determiner Noun]) may be considered to involve a meronymic pattern within these corpora.

It is important to analyse the nature of this meronymic relation in greater detail. In all the corpora studied, one particular phenomenon stands out. What is marked by many occurrences of the meronymic *avec* pattern in these corpora is more than merely information about a part of the whole. More pertinently, the part mentioned is used in order to indicate a particularly salient feature – a salient commercial feature in the adverts and the catalogue, and a perceivable salient feature in the itineraries. For example, when an advert indicates:

(25) *Salon avec cheminée.* (Petites annonces)

[Living-room with fireplace.]

what is understood is that this fireplace constitutes a special feature of the accommodation, and is, thus, a way of justifying the price.

In the case of itineraries, in Example (23), what is important is not (or not only) that the dome is a part of the church, but that it is a distinctive, unusual part of it. In view of its peculiarity, this part may be an element used to locate the main object (in fact, the holonym).

In the toy catalogue corpus the problem is different. Indeed, some occurrences of *avec* do not correspond at all to standard descriptions of this preposition. Traditionally, *avec* introduces a non-essential part and this is the case in all the examples quoted above. However, some occurrences of examples from the toy catalogue seem to be exceptions to this rule: they present expected parts of the object usually intended for adults. What is important here is that the object intended for children looks like the one intended for adults. The identification of the object is reinforced by the use of words such as *réel* ('real'), *vraiment* ('truly'):

(26) *Cuisine avec plaque de cuisson qui rougit vraiment.* (Petites annonces)

[Kitchen with hotplate that really glows red.]

From these observations, we can conclude that when *avec* is frequently used in a corpus, it is likely that the main interpretation of the occurrences

concerns the salient nature of the referent introduced by *avec*. But that which supports this salient element is definitely a part of the main object. So it is true to say that [det N1 *avec* det N2] may be used as a pattern of meronymy in certain corpora.

Concerning the characterisation of textual features, the case of *avec* is quite interesting since domain does not constitute a relevant feature. For example, meronymic *avec* is not very frequent in the geomorphology handbook and, by contrast, very frequent in an oral corpus, such as the itinerary descriptions. Toy catalogues may be seen to belong to the toy domain, but it is not this feature that determines the frequent presence of meronymic *avec*. Hence, in this case, relevant features concern mainly communicative objectives, either persuasive or 'informative'.

The study of these four patterns shows that their behaviour is quite different. The link with text genre is obvious from a quantitative point of view. However, a fine-grained (qualitative) analysis of text genre role – even though it is time consuming – is also necessary.

4. Conclusion

Corpus linguistics is extremely useful in specifying the description of conceptual relation patterns. We have seen that genre can influence the pattern/relation pair from a quantitative point of view or from a qualitative one, or from both.

Even when it is mainly quantitative, an analysis requires that all the occurrences be examined, since what is sought is the frequent association of a pattern with a relation in certain specific texts. A suitably high frequency (more than half of the occurrences) allows us to claim that there is some relevance in using these patterns to identify relations from a computational viewpoint: when analysing all occurrences which correspond to those patterns, the 'noise' will not be too high. From this point of view, only the three prepositions may be considered to mark relational patterns. But, from the statistical point of view, we can say that for all the patterns studied, the factor corpus has (or is on the threshold of reaching) a significant influence on the factor type of patterns. Another major contribution of such a fine-grained analysis is that it often makes it possible to understand why there is such a frequent link between a pattern and a relation in texts belonging to a certain genre.

The main purpose of the study is a crucial determinant in categorising text genres. If the purpose is to identify relations, then that may result in a certain type of text classification. For example, concerning the use of the meronymic *avec*, property adverts are closer to toy catalogues than to car adverts, because there is no use of the meronymic *avec* in car

adverts.[6] One of the issues addressed by this paper is the role of domain in the potential for a pattern to be used to identify a conceptual relation. For some patterns, the fact that a text belongs to a specialised domain is a sufficient reason to say that this pattern may be used as a relational marker (as for *comme*). But it is not always the case: some patterns (such as nominal anaphora) cannot be treated as relational markers even in specialised texts; in some other cases, it is necessary to determine relevant features more precisely (as in the case of *chez*, for which just one domain (natural science, with a didactic perspective) is relevant).

These observations are important in determining which text features (and, specifically, extra-linguistic text features) are relevant to establishing that a given pattern is a relational marker. A very important issue underlying this study concerns what genre is – specifically the link between genre and domain. Generally, in terminology, domain plays a very important role; but what this study shows is that in the case of relational markers, domain is not always crucial for explaining why some patterns may have the role of relational markers. In some cases, it is not at all the domain that comes into play, but the communicative objectives (see for example the case of *avec*). So, explaining variation in terminology requires us to take both domain and communicative objectives into account. In other words, it would be necessary to ponder the definition of genre and, more specifically, its link with the notion of domain. Two kinds of perspectives open up as possibilities for systematising the study of the link between patterns and text genre and, subsequently, making the hypothesis relevant for Natural Language Processing and ontology learning. The first one is to verify on other texts the hypothesis proposed in this paper concerning the semantics of the four patterns and the way they can express conceptual relations through text features, especially hypernymic anaphora in technical texts. It would be necessary to analyse the patterns in texts having the same extra-linguistic features. For instance, the salience value associated with meronymic *avec* can be used to select other corpus genres in which the pattern/relation pair (*avec*/meronymy) will be frequent. This could be the case, for example, with travel guides. For *comme*, it would be necessary to select texts belonging to other specialised domains than the ones examined.

The second perspective concerns the possibility of characterising linguistic features of relevant texts (which arises from the potential for a linguistic pattern to be used in order to identify a semantic relation). This could be achieved by Natural Language Processing methods such as the ones proposed by Biber (1988). Such methods aim to identify automatically the most similar features among different texts in an annotated corpus. But the annotation concerns only syntax. On this basis it

[6] This may appear strange, but in car adverts, parts of the vehicle are just listed without a marker, for example *BMW 525 TDS pack*, *clim*, *ABS*, *cuir*, *an 94*.

might be possible to combine the two kinds of approaches: the ones based on extra-linguistic features and the ones based on linguistic features.

Finally, I hope that this paper makes clear that elaborating a hypothesis about the role of genre in determining the meaning of a potential relation pattern necessitates preliminary studies that may be very lengthy but are, nonetheless, indispensable.

References

Apothéloz, D. and M.-J. Reichler-Béguelin. 1999. 'Interpretations and functions of demonstrative NPs in indirect anaphora', Journal of Pragmatics 31 (3), pp. 363–97.

Aussenac-Gilles, N., B. Biébow and S. Szulman. 2000. 'Revisiting ontology design: a methodology based on corpus analysis' in R. Dieng and O. Corby (eds) Lecture Notes in Artificial Intelligence, Volume 1,937, pp. 172–88. Berlin: Springer-Verlag.

Bahtia, V.K. 1993. Analysing Genre: Language Use in Professional Settings. London: Longman.

Biber D. 1988. Variation Across Speech and Writing. Cambridge: Cambridge University Press.

Biber, D., S. Johansson, G. Leech, S. Conrad and E. Finegan. 1999. The Longman Grammar of Spoken and Written English. London: Longman.

Cabré, M.-T., J. Morel and C. Tebé. 1997. 'Las relaciones conceptuales de tipo causal: un caso practico', Proceedings of V Simposio de Terminologia, Mexico: RITERM. Available online at: http://www.riterm.net/actes/5simposio/cabre6.htm

Cadiot, P. 1997. '*Avec*, ou le déploiement de l'éventail' in C. Guimier (ed.) Co-Texte et Calcul du Sens, pp. 135–55. Caen: Presses Universitaires de Caen.

Cimiano, P., A. Pivk, L. Schmidt-Thieme and S. Staab. 2004. 'Learning taxonomic relations from heterogeneous evidence', Proceedings of the 16th European Conference on Artificial Intelligence (ECAI 2004), Workshop W18 on ontology learning and population, pp. 25–30. Valencia, Spain.

Condamines, A. 2000. '*Chez* dans un corpus de sciences naturelles: un marqueur de méronymie?', Cahiers de Lexicologie 77, pp. 165–87.

Condamines, A. 2002. 'Corpus analysis and conceptual relation patterns', Terminology 8 (1), pp. 141–62.

Condamines, A. 2005. 'Anaphore nominale infidèle et hyperonymie: le rôle du genre textuel', Revue de Sémantique et Pragmatique 18, pp. 23–42.

Condamines, A. and J. Rebeyrolle. 2001. 'Searching for and identifying conceptual relationships via a corpus-based approach to a terminological knowledge base (CTKB): method and results' in D. Bourigault, M.C. L'Homme and C. Jacquemin (eds) Recent Advances in Computational Terminology, pp. 127–48. Amsterdam/ Philadelphia: John Benjamins.

Cornish, F. 1986. Anaphoric Relations in English and French: A Discourse Perspective. London: Croom Helm.

Cruse, D.A. 1986. Lexical Semantics. Cambridge: Cambridge University Press.

Firth, J.R. 1969. Papers in Linguistics 1934–1951. (Fifth edition, 1957.) Oxford: Oxford University Press.

Fuchs, C. and P. Le Goffic. 2005. 'La polysémie de *comme*' in O. Soutet (ed.) La Polysémie, pp. 267–92. Paris: Presses de L'Université Paris-Sorbonne.

Green, R., C.A. Bean and S.H. Myaeng. 2002. The Semantics of Relationships: An Interdisciplinary Perspective. Dordrecht/ Boston/London: Kluwer Academic Publishers.

Halliday, M.A.K. and R. Hasan. 1985. Language, Context and Text: Aspects of Language in a Social-Semiotic Perspective. Victoria: Deakin University Press.

Hearst, M.A. 1992. 'Automatic acquisition of hyponyms from large text corpora', Proceedings of the 14th International Conference on Computational Linguistics, pp. 539–45. Nantes, France.

L'Homme, M.-C. 2004. 'A lexico-semantic approach to the structuring of terminology', Proceedings of Computerm'2004, Computational Linguistics 2004, pp. 7–14. University of Geneva, Switzerland.

Lyons, J. 1977. Semantics. (Two Volumes.) Cambridge: Cambridge University Press.

Maedche, A. and S. Staab. 2001. 'Ontology learning' in S. Staab and R. Studer (eds) Handbook on Ontologies, pp. 173–89. Berlin: Springer-Verlag.

McEnery, T. and A. Wilson. 2004 [1996]. Corpus Linguistics. Edinburgh: Edinburgh University Press.

Mari, A. 2003. Principes d'Identification et de Catégorisation du Sens: Le Cas de *avec* ou l'Association par les Canaux. Paris: L'Harmattan.

Marshman, E., T. Morgan and I. Meyer. 2002. 'French patterns for expressing concept relations', Terminology 8 (1), pp. 1–30.

Meyer, I. 2001. ‘Extracting knowledge-rich contexts for terminography: a conceptual and methodological framework’ in D. Bourigault, M.C. L’Homme and C. Jacquemin (eds) Recent Advances in Computational Terminology, pp. 279–302. Amsterdam/Philadelphia: John Benjamins.

Neches, R., R. Fikes, T. Finin, T. Gruber, R. Patil, T. Senator and W.R. Swartout. 1991. ‘Enabling technology for knowledge sharing’, AI Magazine 12 (3), pp. 36–56.

Pearson, J. 1998. Terms in Context. Amsterdam/Philadelphia: John Benjamins.

Rogers, M. 2000. ‘Genre and terminology’ in A. Trosborg (ed.) Analysing Professional Genres, pp. 3–19. Amsterdam/Philadelphia: John Benjamins.

Swales, J.-M. 1990. Genre Analysis: English in Academic and Research Settings. Cambridge: Cambridge University Press.

Todorov, T. 1984. Mikhail Bakhtin: The Dialogical Principle. (Translated by Wlad Godzich.) Minneapolis: University of Minnesota Press.

	Natural science	Didactic
EUbio	+	+
Manuelsbio	+	+
LMbio	+	–
EUculture	–	+

Table 1: Detail of the corpus constructed for the study of *chez*

	EUbio	Manuelsbio	LMbio	EUculture
Number of words	24,770	82,000	49,440	79,700
Chez	155 (0.63 %)	91 (0.11 %)	107 (0.22 %)	103 (0.13 %)
Meronymic *Chez*	83 (53.6 %)	48 (53 %)	27 (25.2 %)	0

Table 2: Quantitative Results for *chez*

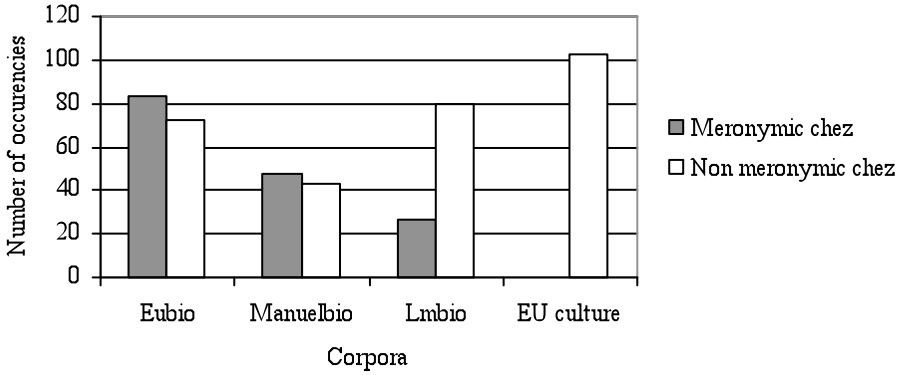

Figure 1: Proportion of meronymic *chez* relative to non-meronymic *chez* according to their corpus origin

	Specialised corpus			Journalistic corpus
	Mécanique	GDP	Atlas	Monde Diplo.
	105,700	165,700	60,000	170,200
Comme	98 (0.09 %)	87 (0.05 %)	46 (0.08 %)	342 (0.2 %)
Relational *comme*	53 (54.08 %)	44 (50.57 %)	36 (78.26 %)	101 (29.53 %)

Table 3: Distribution of *comme* in each corpus

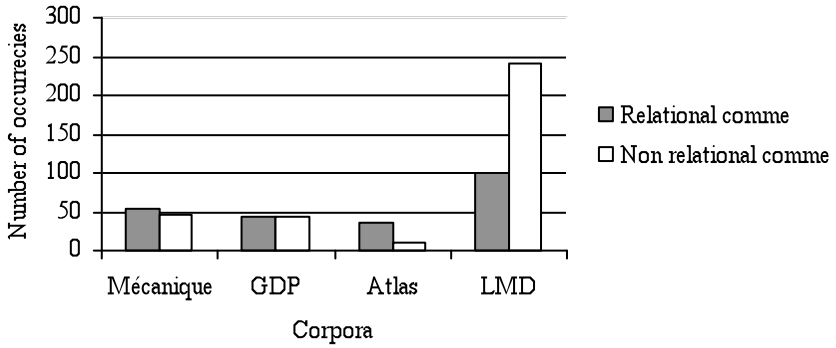

Figure 2: Proportion of relational *comme* relative to non relational *comme* according to their corpus origin

	GEO	GDP	Mouglis	LMD	*Bel Ami*
Number of words	206,000	148,000	45,100	110,700	170,200
Demonstrative anaphora	266	246	122	415	305
Hypernymic anaphora	69 (26 %)	79 (32 %)	79 (60 %)	79 (19 %)	47 (15.5 %)

Table 4: Quantitative results for anaphora

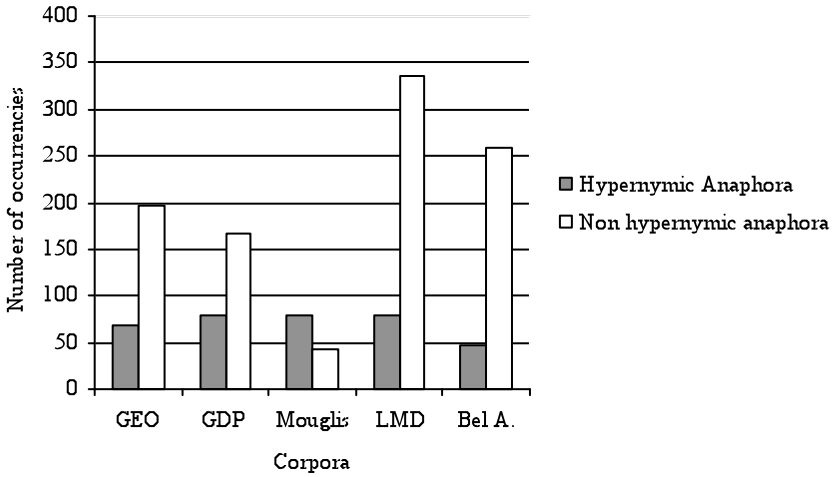

Figure 3: Proportion of hypernymic anaphora relative to non-anaphoric anaphora according to their corpus origin

	Germinal	GEO	Toy catalogue	Small-ads	Itineraries
Number of Words	206,700	230,000	93,000	22,600	48,000
Avec	667	432	236	185	114
Meronymic *avec*	43 (3%)	55 (12.7%)	161 (68.2 %)	141 (76.2 %)	75 (64.6 %)

Table 5: Quantitative Results for *avec*

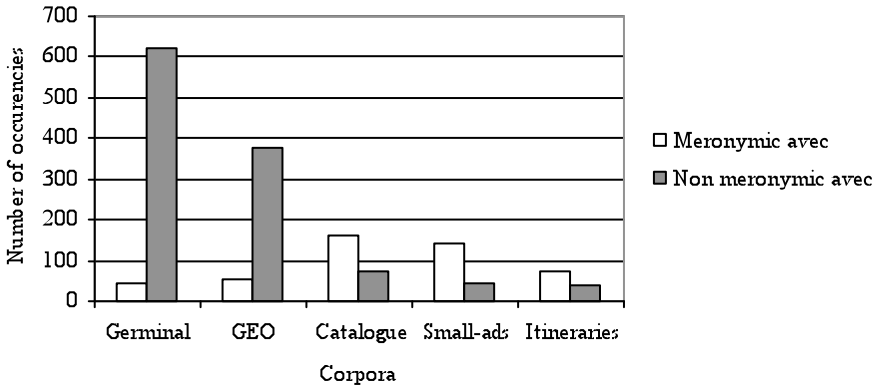

Figure 4: Proportion of meronymic *avec* relative to non-meronymic *avec* according to their corpus origin