Organizing Linguistic Data: Thematic Introducers as an Example*

Sylvie Porhiel

Laboratoire Langues, Textes, Traitements informatiques, Cognition and Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur

1. Introduction

Locating thematic structures in texts is interesting from two points of view: first, they help organize information, and second they help trace textual thematic coherence, which should improve the results of automatic summarization. To identify them, I rely on context-independent thematic introducers, i.e. lexical items such as *au sujet de X, en ce qui concerne X, quant à*, 'about, as far as X is concerned, as for', etc. (See Appendix 1).

These linguistic markers are modeled in the ContextO platform (Université de Paris IV). The software uses a linguistic database in order to trace the relevant semantic information the user is looking for (e.g. thematic coherence). These data provide indices used to write contextual rules matching a discourse category. Papers about the ContextO software have so far described how the platform works and how to declare linguistic data, although the distribution of the markers in the database has not been well documented. While browsing the available data I noticed a discrepancy between a conceptual representation of the linguistic markers and their actual representation in the database leading the linguist to work with two different representations. This is why my research aims at creating a database that matches the linguistic representation in order to create a linguist-friendly tool. The modeling of thematic introducers raises several questions that are worth considering: how is the linguistic data going to be organized in the computer application? Is an organization based on linguistic criteria suitable for a computer application? And, when designing the computer application, what is the cost of implementing each linguistic property and each marker?

In this paper I explain how I proceeded in order to propose a customized distribution of the thematic introducers. First, I state how thematic introducers can be organized in linguistics; second, I explain how the linguistic criteria fit into the computer application; third, I propose a classification of the thematic introducers within the ContextO software; and lastly, I comment on the conclusions of my analysis.

2. How Linguistic Analyses Organize the Thematic Introducers

This section considers how a discourse and a syntactic perspective distribute the thematic introducers. For each of the four approaches, a general overview of the authors' classification is given in appendix 2. I list the criteria used by the authors and indicate where the introducers belong within these classifications. I then comment on these distributions and list the linguistic properties of the thematic introducers.

2.1. Discourse distributions

I consider two discourse distributions: one by Dalcq et al. 1989 and one by Charolles 1997.

* Thanks to J-L. Minel who reread this paper. This research has been written in the framework of the project 'Cognitique 2000', under the direction of G. Sabah.

Coyote Papers 12, 42-61
Language in Cognitive Science

Dalcq et al. (1989) propose to distribute logical linking sentences, i.e. " the set of semantico-logical operations of the thought " (Dalcq et al. 1989:83), to help students interpret and understand technical texts. Logical linking phrases break down into two main groups: the linking phrases relating to the objects of the worlds and the linking phrases inherent in discourse, which are in turn subdivided into two. Within each subdivision, auxiliary notions have been distinguished (appendix 2a). Most thematic introducers are grouped under the auxiliary notions of 'ways of doing' and 'point of view', the latter one being split into 'general', 'particular' and 'universal'.

Table 1: Distribution according to Dalcq et al. 1989

Linking phrases		
inherent in discourse	*relating to the objects of the world*	*Auxiliary notions*
	à propos de 'about', à l'égard de 'with regard to', au sujet de 'about', concernant 'concerning', de ce point de vue 'from this point of view', etc.	point of view (particular)
en ce qui me concerne 'as far as I am concerned', quant à moi 'as for me'		modalization (speaker's point of view)
	au niveau de 'at the level', du côté de 'on the side of'	position

Table 1 shows first that most lexical items (whether prepositions or adverbs) have been classified in the group 'linking phrases relating to the objects of the world'; second, that *quant à moi* 'as for me' and *en ce qui me concerne* 'as far as I am concerned' are classified in the group 'linking phrases inherent in discourse'; third, that these two groups belong respectively to the subdivisions 'linking phrases relating to facts' and 'speaker's intervention'; fourth, that *au niveau de* 'as regards' and *du côté de* 'on the X side; as far as X is/are concerned', as thematic introducers, are considered incorrect[1], which explains why these lexical items are found under 'linking phrases relating to a spatio-temporal frame' and 'place' and can then respectively be translated as 'on/at the level' and 'on the side of'.

In his essay *L'encadrement du discours* [discourse frames], Charolles proposes a theory of frames. In that textual perspective *frames* are used to refer to circumstances in which a certain state or a series of events has to be considered. The general category of frame can be subdivided into four other categories; one of these categories is the thematic frame introducers. These thematic introducers specify the theme of the following proposition(s) as opposed, for example, to the verifunctional ones that specify in which circumstances one or more propositions are valid. The author lists seven properties pertaining to that group: 1) thematic introducers are placed initially and are syntactically loose, 2) the nominal phrase after the introducer is usually taken up in the following proposition by a pronoun, 3) the detachment is referentially constrained since, to be detached and extracted, the nominal phrase must either have been explicitly introduced in the discourse or be implicitly available, 4) by using thematic

[1] *Au niveau de, du côté de* as well as *au plan de* are considered prescriptively incorrect (Dupré 1972, Hanse 1994). I consider them to be thematic introducers, first because speakers use them to introduce thematics and second because they can be translated by English thematic introducers.

introducers, the speaker signals that what s/he is about to say concerns X and not Y, 5) on the interlocutor's side, it is then assumed that s/he establishes a link with what has been previously stated, 6) thematic introducers do not question the truth of the following propositions (the proof being that they can prefix questions: *Concernant, X, je me demande si* 'Concerning X, I wonder if'), and 7) they partition information.

2.2. Syntactic distributions

From a syntactic point of view, I consider how the thematic introducers are distributed in the lexicon-grammar theory and in a combinatory distribution.

In the lexicon-grammar theory (M. Gross 1986), the perspective is purely syntactic (See appendix 2b). The thematic introducers are mentioned in an analysis concerning adverbs (in a broad sense) and are distributed into five tables with specific structures as follows:

Table 2: Distribution according to M. Gross 1986

Structures[2]	*Examples*	*Thematic introducers*
Prép C de N (PCDN)	au moyen de N 'by means of'	à l'égard de 'regarding', à propos de 'about', au plan de 'as regards', au sujet de 'about', en fait de 'as regards', en matière de 'as far as X is/are concerned'
Prép C prép N (PCPN)	par rapport à N 'in comparison with'	côté 'on the X side', quant à 'as for', relativement à 'as far as X is/are concerned'
Prép Dét C (PDETC)	contre toute attente 'against all expectations'	à ce propos 'in this connection', à ce sujet 'on that subject', en cette matière 'on that matter'
P (frozen sentence) (PF)	Dieu seul le sait 'goodness only knows'	concernant 'concerning', en ce qui concerne 'as far as X is/are concerned'
Prép V W (PV)	à dire vrai 'to tell the truth'	concernant 'concerning', s'agissant 'as regards'

This distribution is based on the internal structure of the adverbs, i.e. the lexical combinations that constitute them. For example the PCDN (structure: Prép C *de* N) class gathers adverbs whose second preposition is *de*, whereas PCPN (structure: Prép C prép N) gathers adverbs whose second preposition is not *de.*

Porhiel's 1999 study relies on the notion of *frame* (in the sense of Charolles 1997). Accordingly, only detached and syntactically unbound phrases are under consideration. The analysis aims at finding syntactic criteria to back the intuitive distribution between the thematic introducers (which specify the themes of the following proposition(s)) and the verifunctional introducers (which specify the circumstances in which one or more propositions are valid), in particular the enunciative introducers and the knowledge ones. It starts from the hypothesis that the category thematic introducers should possess combinatory properties that other frame introducers do not have and then hypothesizes that this category can be further subdivided. For example, a distinctive feature is that thematic introducers are not followed by an adjectival phrase, whereas verifunctional introducers can be; in the same way, thematic introducers can be followed by adverbial phrases (*en ce qui concerne ici* 'as far here is concerned'), whereas verifunctional ones cannot. According to their combinatory criteria (external structure), i.e. whether they are or are not followed by proper nominal phrases, common nominal phrases or verbal phrases, the thematic introducers break into two main groups, T_1 and T_2. T_1 consists of introducers which do not combine with proper nominal phrases unless they are preceded by a determiner: *au chapitre de la France* 'as regards France', * *au chapitre de France*. T_2 consists of introducers combining with proper nominal phrases that can

[2] W: complement, P: letter prefixing each class of adverbs, P/prép: preposition, N: noun, Dét: determiner, P: frozen phrase, V: verb, C: frozen complement. Classes are in brackets.

be preceded by a determiner: *à propos de Pierre* 'concerning Pierre'*, à propos de l'Angleterre* 'concerning England'. These two groups can be further broken into five homogeneous sub-groups as shown in Figure 1 below:

Figure 1: Distribution according to Porhiel 1999

Thematic Introducers
- T_1
  - • no determiner before the nominal phrase (*au point de vue professeurs* 'Concerning teachers')
  - • definite article before the common nominal phrase (*au chapitre du chômage* 'in the matter of unemployment')
- T_2
  - • no determiner possible before the nominal phrase (*à propos d'animaux/à propos des animaux* 'about animals/about the animals')
  - • determiner compulsory before the nominal phrase (*pour ce qui est de ce livre* 'concerning this book)

This study, as well as Porhiel 1998, mentions that the introducers can vary morphologically, paradigmatically and can be more or less frozen (G. Gross 1990, 1996). As for their lexico-semantic compatibilities, thematic introducers do not impose any particular constraints.

2.3. Synthesis

The four presented analyses are not strictly dedicated to the study of thematic introducers. They put forward ways of organizing them with regards to other lexical items. These studies underline that the classifications are dependent on three factors: the theoretical frame, the criteria considered and the aim of the researcher. They first depend on a particular theoretical framework: they can be discourse-oriented or sentence-oriented. In two studies, Dalcq et al. (1989) and M. Gross (1986), the placement criterion is not mentioned and the authors give examples with the two positions. For Charolles (1997), the placement criterion is differential; although it is also a distinctive feature with the combinatory distribution, it does not show on Figure 1. The classifications also depend on the criteria considered, which explains the variations between the two syntactic distributions: one is centered on the internal structure of the lexical items, the other on their external structure. Lastly, they rely on the aim that the researcher strives for. For example, the study by Dalcq et al. has a practical aim and provides the students with a list of linguistic items that can be used to express specific discourse relations. Charolles' study is more theoretical and does not provide a list of thematic introducers. Its aim is to account for the cognitive implications and the mental operations that frames instantiate in discourse.

These analyses, though conducted on different levels, highlight the linguistic properties of thematic introducers. Concerning their internal structure, they can vary morphologically (*au(x) chapitre(s)* 'as regards') and in tense (*pour ce qui est/était de* 'as far as X is/was concerned'); they can have paradigms (*au/sur le sujet de* 'about') and they can accept insertions (*pour ce qui est* notamment *de* 'as far as X is particularly concerned'). Concerning their external structure, they can combine or not combine with determiners, they can introduce or not introduce a complement with a preposition, *à* or *de*, they impose no semantico-lexical constraints on the complement they introduce. As for their place in the sentence, they are detached and not syntactically bound to a verb, a noun or an adjective.

On the cognitive side, thematic introducers organize and partition information. Indeed frames are markers of coherence that surface in texts and reproduce the writer's strategy by splitting information into homogeneous categories with regard to certain characteristics. They are essential to the segmentation of a text into frames in an automatic textual analysis since a discourse segment[3] (Fraser 1999) can be integrated in a frame, although no

[3] I will use 'discourse segment' as a cover term to refer to 'proposition', 'sentence', 'utterance' and 'message' unless more specificity is required. (Fraser 1999: 938)

linguistic indicator introduces that discourse segment. As I want to locate thematic structures in texts, Charolles' frame theory seems to be the appropriate theoretical framework. In that discourse framework, thematic introducers form a homogeneous group.

3. A Computer Application in ContextO

This section deals with the ContextO software and the indices, i.e. specific markers, that this software uses to locate thematic structures.

3.1. The software and the theoretical framework

To capture the thematic introducers, I use the ContextO platform[4] (Ben Hazez et al. 2001; Minel et al. 2000, 2001). This software identifies specific semantic information in texts and extracts relevant sentences meeting applied criteria. To achieve this, the system uses linguistic data declared in a database and declarative contextual rules written according to the contextual exploration method (Desclés 1997). This theoretical framework is textual: a unit is not necessarily analysed in a set of adjacent sentences; a linguistic item is given a semantic or a discourse tag once heuristic rules have analysed the context; indices provide the inferential clues used by the system.

As a textual theory, the contextual exploration method is in line with Charolles's frame theory and also corresponds to my aim, i.e. locating thematic structures in texts. The question is however, is it possible within that software and within that theoretical framework to reflect the conceptual textual organization of the discourse frames? Both theories underline the importance of incremental processes. This cognitive operation has not yet been implemented in computer science and only surface items (indices in the contextual exploration method) provide implemental clues. Consequently, what is mentally comprehended needs a tangible counterpart for the computer to process it and the data distribution in a computer program is unlikely to map a textual one.

3.2. The indices

The tangible elements consist of two kinds of indices, the triggering indicators and the complementary indices. The triggering indicators, i.e. the thematic introducers, are indices that instantiate contextual rules and from which a search space is defined. Concerning their forms, all (and only the relevant forms) are captured: upper/lower case markers, plural/singular and tense variations. For instance, *s/Sur le sujet de* 'o/On the subject of' et *a/Au sujet de* 'a/About' and their plural form when it exists: *sur les sujet de*, **aux sujets de*. As for insertions, all complex indicators, except for *quant à* 'as for' and *relativement à* 'as far as X is/are concerned', accept them after the word-base. To take this syntactic configuration into account, without generating too much noise, requires the preposition introducing the complement to be declared as an index (or mark) and the indicators not to be declared beyond their word-base: *à propos de* 'about' is declared *à propos* (See appendix 3). Lastly, some indicators present paradigmatic variations. *Au, sur, sur le, sur les* can precede the word-base *sujet* but only *sur le sujet de* 'on the subject of' accepts insertions before the word-base. As a result, it is not relevant to use a set of prepositions to jointly declare the two lexical items: on the one hand the declared form for *au sujet de* is *au sujet* and on the other hand the one for *sur le sujet de* is *sujet*, *sur le* and *de* being indices, which allows for insertions. For other lexical items, a set that groups paradigmatic variations is appropriate: *en ce qui concerne, pour ce qui concerne, pour tout ce qui concerne* 'as far as X is/are concerned'. To conclude, morphological, aspectual and paradigmatic variations are relevant information, which do not influence the distribution of the thematic introducers. Conversely, considering the insertions breaks down the indicators into two major groups: those not followed by a preposition and those followed by a preposition, the latter subdividing into those followed by *à* or by *de* (See appendix 4a). The data collected by the morpho-syntactic analyses are then of help, if they lead to the surface identification of the lexical

[4] The ContextO software is composed of three subsystems: a linguistic knowledge database, a contextual exploration engine, and specialized agents. This architecture makes it possible to update the software.

items and if they meet the constraints of the system.

Complementary indices favor or inhibit the recognition of triggering indicators. They are general indices that apply to the whole category. The sentence-initial position is an example of such an index and is used to distinguish between morpho-syntactic bound or unbound lexical items. Though they are differential in extracting the proper lexical items, they do not interfere in the data organization. Complementary indices can also be specific indices that only apply to some indicators. For instance, there must be no comma between *à propos* and *de*; *au chapitre* must not be followed by a number or an adjective. As the elements combining with different indices are not grouped together, specific indices set apart some thematic introducers and create as many classes of introducers as there are thematic introducers requiring specific indices. The consequences of this principle are that the groups previously mentioned can be further subdivided into two: the introducers combining with general indices and those combining with specific ones.

4. Classification of the Thematic Introducers within the ContextO Software

Section four is about the distribution of the thematic introducers in ContextO. It also accounts for the calculated choices made to enhance the location of the thematic introducers.

4.1. The distribution

The linguistic analyses provide criteria for all the thematic introducers on the syntactical level (e.g. the sentence-initial position) or, on the textual level (e.g. the partitioning of information). Finer morpho-syntactic distinctions such as in M. Gross (1986) and Porhiel (1999) split the category into subgroups. However, in the ContextO software, such a degree of specificity is not appropriate as the first priority is the surface recognition of specific lexical items, i.e. how they are likely to decline in texts. This goes hand in hand with the constraints imposed by the system. As a result, the thematic introducers are distributed according to the formal, syntactic and lexical indices they combine with (See appendix 4b).

Thematic introducers are first divided according to whether or not they are followed by a preposition. If so, they can be broken down into a further two: those followed by the preposition *à* and those followed by the preposition *de*. The three groups are then sub-categorized according to the locating indices favoring the surface recognition of the thematic introducers, i.e. whether or not they accept insertions either before or after the word-base, or before and after the word-base. Following these general morphological and syntactic specifications, this initial linguistic schematic distribution can be improved. Some thematic introducers combine with general indices and others with specific indices, which breaks down the thematic introducers into fourteen groups. Furthermore, those thematic introducers combining with specific lexical and semantic indices subdivide further as shown on the distribution tree in appendix 4b. Some groups are void as this distribution is subjected to change.

4.2. Relevance testing

The distribution proposed in appendix 4b has undergone relevance tests to grade the cost of considering some linguistic criteria, e.g. insertions, and the choice of some lexical items.

The notion of search space, i.e. a textual segment delimited from the indicator, allows us to consider the type of insertion put forward by the linguistic analysis and to limit the number of words between the word-base and the preposition. Porhiel 1998 locates introducers of interest[5] in the newspaper *Le Monde* (1993) with finite state transducers within INTEX (an integrated corpus processor based on the use of large covering lexicons; Silberztein 1993). She estimates the number of words constituting an insertion at one to four words before the word-base and at one word to a proposition after the word-base. In the corpus of *Le Monde Diplomatique*, the search space was limited to three words before and after the word-base. Generally speaking, the number of occurrences increases

[5] That notion includes thematic introducers.

significantly and in some cases even doubles. However, a careful analysis of the occurrences shows that there are few insertions, which confirms what has already been observed by Porhiel 1998. The author states that the percentage ratio of insertions, regardless of their place in the sentence, for the lexical items *en/pour ce qui concerne* 'as far as X is/are concerned', *s'agissant de* 'as regards', *en fait de* 'as regards' is respectively 1.01%, 6.28% and 7.33%. It also indicates that using a search space generates noise that needs to be reduced. On the one hand, studying the possible insertions before the word-base, I came to the following conclusions about the thematic introducers. *Au chapitre (de)* 'as regards', *au niveau (de)* 'as regards', and *au plan (de)* 'as regards' do not accept insertions before the noun-base (*Au chapitre de*: Au **troisième** chapitre, l'auteur (...)'In the third chapter, the author'; a*u niveau de*: Au niveau **modeste** des collectivités locales, (...)'At the modest level of the local authority, (...)). E*n/pour ce qui concerne* accepts personal pronouns between *en/pour ce qui* and *concerne*; *en/pour ce qui concerne* (*En ce qui concerne*: En ce qui **les** concerne, je[6] (...)'As far as they are concerned (...)'). *En/pour ce qui a trait, en/pour ce qui regarde* 'as far as X is /are concerned' only accept *tout* between *en/pour* and *ce qui concerne/a trait/regarde* (pour **tout** ce qui a trait à). *Sur ces chapitres* 'on these subjects' and *sur les plans* 'from these points of view' also only accept *tous* as an insertion (*Sur **tous** ces chapitres*, (...) 'On all these subjects, (...)'). On the other hand, the insertions after the base are either adverbs (*S'agissant **le plus souvent** d'investissements*, (...) 'as regards: As regards most of the time investments (...)') or adjectives (*Sur le sujet **particulier** de X*, (...) 'On the particular subject of X (...)'). As a consequence, it is too costly to have a search space of three words. As the aim of the application is to process large amounts of texts, it is proportionally not worth it considering insertions unless applying these suggestions: 1) *tout* and *tous* being the only possible insertions with some introducers, one should declare two forms for the markers in the database: without *tout/tous* and with *tout/tous*; 2) as the accepted insertions can be grouped into definite sets of items (the pronouns, the adverbs and the appreciative adjectives), only words declared in these sets should be allowed as insertions.

The second set of tests was done using simple prepositions and resumptive pronouns (Auricchio et al. 1995). Although the prepositions *pour* 'for' and *sur* 'on' may instantiate thematic frames, in a newspaper corpus such as *Le Monde Diplomatique*, they generate far too much noise and should not be used as triggering indicators. *À ce niveau* 'at this level', an indicator with a resumptive pronoun, as opposed to *sur ce plan* 'from this point of view' and *sur ce chapitre* 'on this subject', is not a thematic introducer in any of its usage and, consequently, should not be selected.

A thorough analysis of the insertions highlighted those that might be interesting from a linguistic point of view, but that might cost too much from a practical one, as shown by two analyses carried out in different newspaper corpora. This outcome influenced and modified the distribution of the thematic introducers as well as their declaration in the database.

4.3. Comments

The following should be noted when considering the distribution of the thematic introducers in the ContextO software. First, the general and specific indices that are differential in locating thematic introducers, and the possibilities and constraints of the operating system, have implications as to the distribution of the lexical items. The linguistic analysis is important and cannot be bypassed. But its results have to comply with the requirements of the ContextO software. Compared to the linguistic analysis, the thematic introducers are divided into fourteen groups (some with sub-groups) and the linguistic properties only help to locate the form of the lexical items. What is relevant in linguistics does not have the same relevance in the computer application. It is also noticeable that the linguistic generalizations are erased and this is particularly true for the textual distributions. The distribution tree is an organized representation of the formal elements needed by the system to locate specific surface forms, which by no means assures that the proper forms have been identified. It is also the reflection of the calculated choices (e.g. insertions) to locate thematic introducers as cost effectively as possible.

Locating indices subdivide the discourse category thematic introducers. Although these markers are

[6] In some instances these forms would be classified as enunciative introducers.

important in discourse, neither textual criteria organize them, nor does the important sentence-position criterion appear in the distribution tree. These are modeled as textual and structural indices as they belong to another level of description.

5. Conclusions and Prospective Research

Now that the distribution of the thematic introducers has been commented on, I can answer the other questions posed in the introduction. The linguistic distribution of the thematic introducers is not reflected in the computer distribution. This is particularly true of the discourse approach (where they form, for the time being, one single homogeneous group) as the lexical items have been distributed into fourteen groups, and yet this is also the approach favored to locate the thematic structures in texts. However, the distribution is only computer-logical. The system needs linguistic information that we do not need as human beings to adjust itself appropriately. These are, for example: case and morphological variations, insertions that have to be anticipated, without which the system cannot operate. These formal requirements as well as the constraints of the system are seen in the distribution of the thematic introducers. They are an essential base towards a more targeted location of the thematic introducers without dictating their distribution.

This level, though essential, is not satisfactory in a discourse or in a textual approach. The properties that unify the thematic introducers as a textual category do not appear explicitly on the distribution tree. No indication allows the system to identify the correct lexical items by means of general positional indices. Lastly, it is not linguist-friendly enough. The distribution tree has very little to do with the discourse oriented conceptual linguistic view. Although similar, it is not going to be identical to the one obtained with the locating indices. For instance: *quant à* 'as for' usually appears structurally at the end of an enumeration; and frames can integrate other frames (1). However, there is no real integration in (2):

(1) En ce qui concerne les inférences, du point de vue sémantique (...).
'As far as inferences are concerned, from a semantic point of view (...)'

(2) Au sujet de Paul, en ce qui concerne la voiture (...).
(...) 'About Paul, as far as the car is concerned'

Most of this analysis has yet to be done.

At this stage of my analysis, I think that the representation of the thematic introducers should be a multi-level one in order to be more linguist-friendly. There should be a first level distribution with formal locating indices (indicators, general and specific indices). The second level should be a conceptual view matching the linguistic organization of the discourse frames. This level, organized with structural and textual indices, would then partially meet the linguistic specifications. Still, for the time being there is no way to state the general properties of the different categories of frames. The next step of my research will consist of finding links to switch from one level of representation to another.

References

Auricchio, Agnès, Caroline Masseron and Claude Perrin (1995). L'anaphore démonstrative à fonction résomptive. *Pratiques* 85: 27-52.

Ben Hazez Slim, Jean-Pierre Desclés and Jean-Luc Minel (2001). Modèle d'exploration contextuelle pour l'analyse sémantique de textes. TALN 2001, Tours 2-5 July 2001.

Charolles, Michel (1997). L'encadrement du discours – Univers, champs, domaines et espaces. *Cahier de recherche linguistique 6.*

Dalcq, Anne-Elizabeth, Dan Van Raemdonck and Bernadette Wilmet (1989). *Le français et les sciences – Méthode de français scientifique avec lexique, index, exercices et corrigés.* Paris: Editions Duculot.

Desclés, Jean-Pierre (1997). Système d'exploration contextuelle. Co-texte et calcul du sens. In Jean-Claude Guimier (dir.) Caen, Presses Universitaires de Caen, pp. 215-232.

Dupré, P. (1972). *Encyclopédie du bon français dans l'usage contemporain - Difficultés, subtilités, complexités, singularités*. Paris, Éditions de Trévisse. 3 tomes.

Fraser, Bruce (1999). What are discourse markers? *Journal of Pragmatics* 31: 931-952.

Gross, Gaston (1990). Les mots composés. *Modèles linguistiques* 12(1): 47-63.

Gross, Gaston (1996). *Les expressions figées du français.* Paris: Ophrys.

Gross, Maurice (1986). *Grammaire transformationnelle du français 3 – Syntaxe de l'adverbe.* Paris: Asstril.

Hanse, Joseph (1994). *Nouveau dictionnaire des difficultés du français moderne*. 3ème édition. Louvain, Éditions de Boeck-Duculot.

Knowles Gerry (1996). Corpora, databases and the organization of linguistic data. In Jenny Thomas and Mick Shord (eds), *Using Corpora for Language Research*. New York: Longman, pp. 36-53.

Minel, Jean-Luc and Jean-Pierre Desclés (2000). Résumé automatique et filtrage des textes. In Jean-Marie Pierrel (dir.), *Ingénierie des langues*, Paris, Editions Hermès, pp. 253-270.

Minel, Jean-Luc, Jean-Pierre Desclés, Emmanuel Cartier, Gustavo Crispino, Slim Ben Hazez and Jackiewicz Agata (2001). Résumé automatique par filtrage sémantique d'informations dans des textes. Présentation de la plate-forme FilText. *Revue Technique et Science informatiques* 3.

Porhiel, Sylvie (1998). *Les indicateurs d'intérêt*. Doctoral dissertation, Université de Paris XIII, Villetaneuse. Distributed by Les presses du Septentrion.

Porhiel, Sylvie (1999). Proposition de classification des indicateurs d'intérêt. Unpublished Workpaper.

Silberztein, Max (1993). *Dictionnaires électroniques et analyse automatique de textes: le système INTEX*. Paris: Masson.

8, rue de Quimper,
29190 Pleyben, France

sylvieporhiel@hotmail.com

Appendix 1
Possible English equivalents of some thematic introducers
(According to bilingual dictionaries)

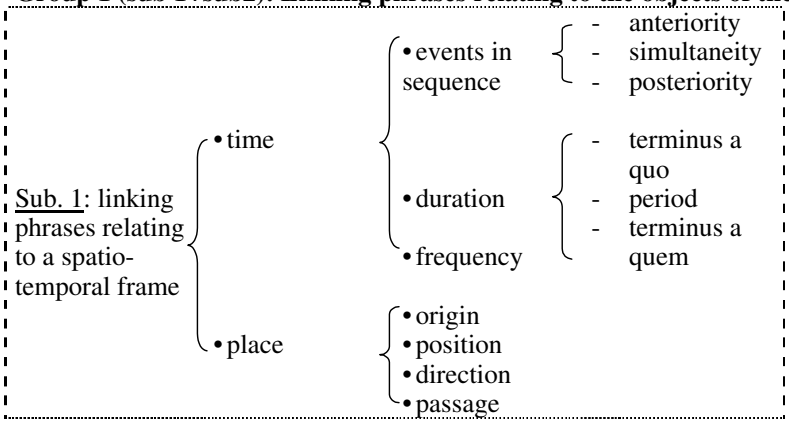
Thematic introducers	*Possible English equivalents*
à ce propos	in this connection
à ce sujet	on that subject, about that
à l'égard de	with regard to, regarding
à propos de	about, regarding
au chapitre de	as regards, in the matter of
au niveau de	as regards
au plan de	as regards, as far as X is concerned
au sujet de	about, on the subject of
concernant	as regards, with regard to, as far as X is/are concerned
côté	~wise, on the X side, as far as X is/are concerned
dans ce cas	in that case
dans le cas de	in the case of
du côté de	as for
du point de vue de	as far as X is/are concerned
en/pour ce qui a trait	as far as X is/are concerned
en ce qui concerne, pour ce qui concerne	as regards, as far as X is/are concerned, in matters of, with regard to
en fait de	as regards
en matière de	as far as X is/are concerned
pour tout ce qui concerne	in all matters of
pour tout ce qui touche à	in all matters of
quant à	as for
relativement à	as far as X is/are concerned
s'agissant de	as regards
sur ce chapitre	on this subject
sur le chapitre de	on the subject of, as regards, in the matter of
sur le plan de	from the point of view of, in terms of

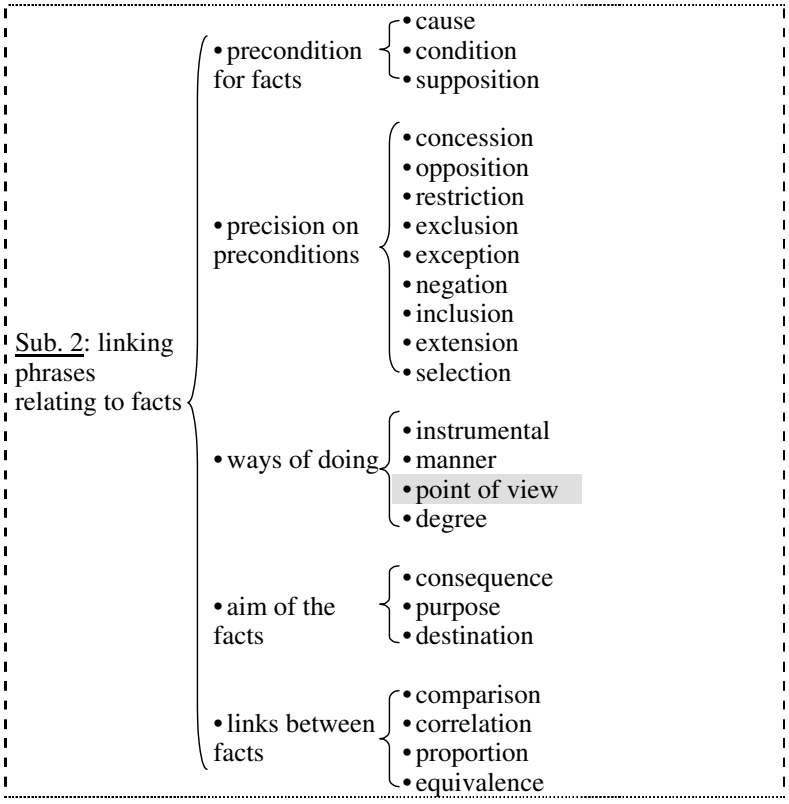
Key:
X : stands for 'complement' (a noun, a verb, and adverb)

Appendix 2
Linguistic distributions

a) According to Dalcq et al. 1989

Group 1 (sub 1+sub2): Linking phrases relating to the objects of the world





Group 2 (sub 3+sub 4): Linking phrases inherent in discourse

Sub. 3: Discourse organization	• beginning • continuation • end • parenthesis • implication • conclusion- • summary
Sub. 4: Speaker's intervention	• emphasis • objection • dialectic • explanation • justification • parenthesis • modalization: • certain, • probable, • possible, • point of view

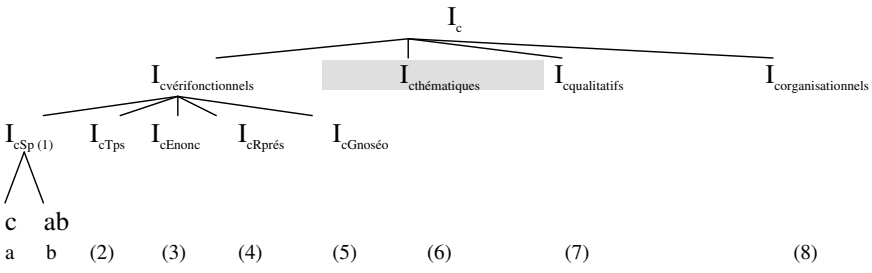
Key:
Sub.: Subdivision
: highlights where the thematic introducers are found in the classification

In Sub. 2 these are: à propos de, à l'égard de, au sujet de, concernant, dans le cas de, de ce point de vue, du point de vue, du point de vue de, en ce qui concerne, en cette matière, en fait de, en matière de, pour ce qui est de, quant à, relativement à, sur ce plan, sur le plan de.

In Sub. 3 these are: en ce qui me concerne, quant à moi.

b) According to Charolles 1997 (completed)

Schematic diagram representing the distribution of the frame introducers

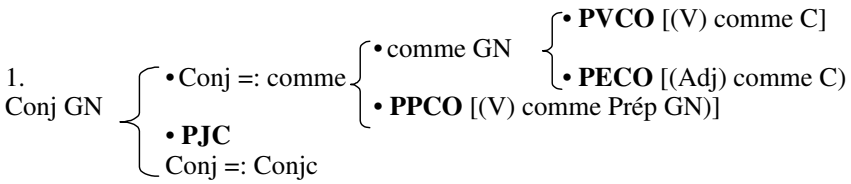


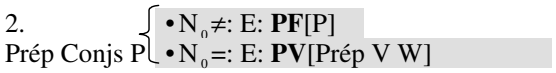
Key:
I_c: frame introducers
$I_{cvérifonctionnels:}$ verifunctional introducers
$I_{cthématiques:}$ thematic introducers
$I_{corganisationnels:}$ organizational introducers
$I_{cSp:}$ (Introducteurs spatiaux) spatial introducers
ab: abstract introducters
c: concrete introducers
$I_{cTp:}$ (Introducteurs temporels) temporal introducers
$I_{cEnonc:}$ (Introducteurs énonciatifs) enunciative introducers
$I_{cRprés:}$ (Introducteurs représentatifs) representative introducers
$I_{cGnoséo:}$ (Introducteurs gnoséologiques) knowledge introducers
: highlights where the thematic introducers are found in the classification

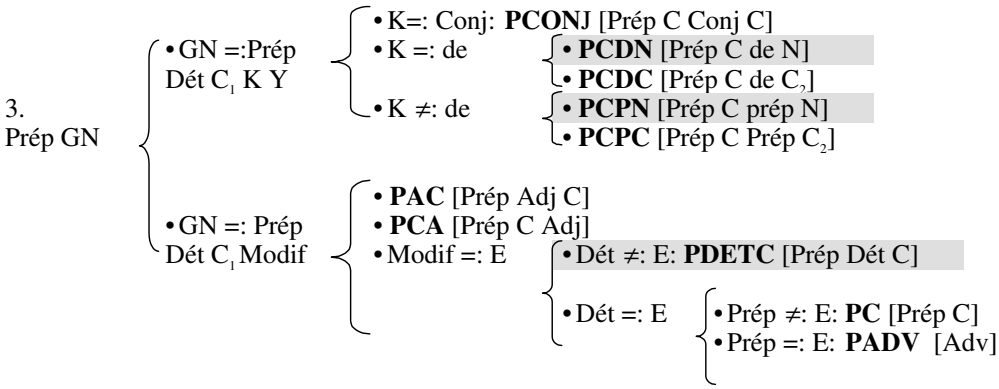
Examples for each category:
(1) **En** Corée du Sud, p. '**In** South Korea, p.'
(1a) **Du côté de** la plage, p. '**In the direction of** the beach, p'.
(1b) **Sur le front des** dépenses sociales, p '**On the front of** the social expenditure, p.'
(2) **Pendant** l'été, p. ' **During** Summer, p.'
(3) **Selon** le Président des États-Unis p. '**According to** the President of the United States, p.'
(4) **Dans le** dernier **film de** Lelouch, p. '**In** Lelouch's latest **film**, p.'
(5) **Dans la perspective de** la physique quantique, p. '**In the** quantum physics **perspective**, p.'
(6) **En ce qui concerne** la Corée du Sud, p. '**As far as** South **Korea is concerned**, p.'
(7) **Par chance**, p '**Fortunately**, p.'
(8) **En bref**, p '**In short**, p.'

c) According to M. Gross 1986

Classes are in upper cases and are preceded by the letter P. Their structure is square brackets. The names of the different classes are in bold.







Key:
C, C1 and C2 are frozen complements
Conj: conjunction
Conjc: coordinating conjunction
Conjs: subordination conjunction
K: connector
E: "null"
N are nominal variable nominal elements
N_0: subject
Modif (modifier): attributive adjectives, possessive phrases, relative clauses
Y: variable; here either N (variable nominal phrase) or C_2 (frozen part)
=: equals
≠: does not equal
▒ : highlights where the thematic introducers are found in the classification

d) According to Porhiel (1999)

The introducers are subdivided into: a) T1 (these do not combine with proper nominal phrases unless preceded by a determiner) and b) T2 (these combine with proper nominal phrases that can be preceded by a determiner).

Combinatories of the T1

	G.Npr	*G. Nco*				*G.V*
Introducers	***ØN***	***ØN***	***LeN***	***CeN***	***Son N***	***UnN***	***Vinf***
au chapitre	—	+	—	—	—	—	—
au niveau	—	+	—	—	—	—	—
au plan	—	+	—	—	—	—	—
au point de vue	—	+	—	—	—	—	—
chapitre	—	+	—	—	—	—	—
en fait de	—	+	—	—	—	—	+
côté	—	+	—	—	—	—	—
en matière de	—	+	—	—	—	—	—
niveau	—	+	—	—	—	—	—
point de vue	—	+	—	—	—	—	—
sur le chapitre	—	+	—	—	—	—	—
du côté de	—	—	+	—	—	—	—
au chapitre de	—	—	+	—	—	—	—
du point de vue de	—	—	+	—	—	—	—
sur le chapitre de	—	—	+	—	—	—	—
sur le sujet de	—	—	+	—	—	—	—

Combinatories of the T2

	G. Npr	*G. Nco*				*G.V.*
Introducers	***ØN***	***ØN***	***LeN***	***CeN***	***Son N***	***UnN***	***Vinf***
à propos de	+	+	+	+	+	—	+
au sujet de	+	+	+	+	+	—	+
s' agissant de	+	+	+	+	+	—	+
dans le cas de	+	+	+	+	+	—	—
pour ce qui est de	+	—	+	+	+	—	+
quant à	+	—	+	+	+	—	+
concernant	+	—	+	+	+	—	—
en ce qui concerne	+	—	+	+	+	—	—
en ce qui regarde	+	—	+	+	+	—	—
en ce qui touche	+	—	+	+	+	—	—
pour ce qui a trait à	+	—	+	+	+	—	—
pour ce qui concerne	+	—	+	+	+	—	—
pour ce qui regarde	+	—	+	+	+	—	—
pour ce qui touche	+	—	+	+	+	—	—
pour ce qui touche à	+	—	+	+	+	—	—
pour ce qui relève de	—	—	+	+	+	—	—

Key:

Le: le, l', la, les
Ce: ce, ces, cette, c'
Un: un, une, des
Son: son, sa ses
Ø: absence
Vinf: verb in the infinitive

'-': the property does not exist
'+': the property exists
G.Npr: proper noun phrase
G.Nco: common noun phrase
G. V: verbal phrase

Appendix 3

Sample of form declaration in ContextO

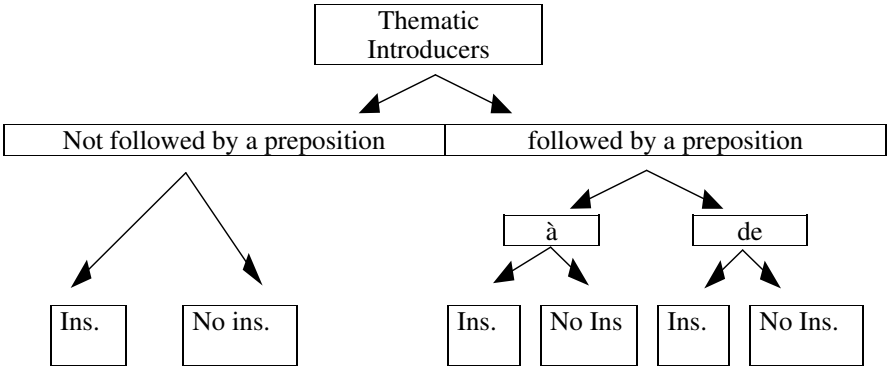
Thematic Introducers	Captured forms	Classes
à ce chapitre	à ce chapitre &virgules	&int_thématq_nonsuivis_prep_ins_surcesujet_min
à ce propos	à ce propos & virgules	&int_thématq_nonsuivis_prep_min
à ce sujet	à ce sujet &virgules	&int_thématq_nonsuivis_prep_min
à propos de	à propos	&int_thématq_suivis_prep_de_ins_min
au chapitre	&au_min &base_chapitre	&int_thématq_nonsuivis_prep_auchapitre_min
au chapitre de	&au_min &base_chapitre	&int_thématq_suivis_prep_de_ins_auchapitrede_min
au niveau de	&au_min &base_niveau	&int_thématq_suivis_prep_de_ins_auniveaude_min
au sujet de	au sujet	&int_thématq_suivis_prep_de_ins_min
en ce qui concerne pour (tout) ce qui concerne	&av_base_pdg_en_min ce qui &base_concerne	&int_thématq_nonsuivis_prep_min
quant à	quant &prep_à	&int_thématq_suivis_prep_à_min
sur le sujet de	&base_sujet	&int_thématq_suivis_prep_de__ins_surlesujetde_min
sur tous ces chapitres	sur tous ces chapitres	&int_thématq_nonsuivis_prep_min

Key:
'&' precedes each class of items
&au_min = {au, aux}
&av_base_pdg_en_min = {en ce qui, pour ce qui, pour tout ce qui}
&base_chapitre = {chapitre, chapitres}
&base_concerne = {concerne, concernera, concernait}
&base_niveau = {niveau, niveaux}
&base_sujet = {sujet, sujets}
&prep_à = {au, aux, à}

The classes can be traced on the distribution tree (Appendix 4b) as they are highlighted in gray.

Appendix 4
Distribution of the thematic introducers in the ContextO software (not exhaustive)

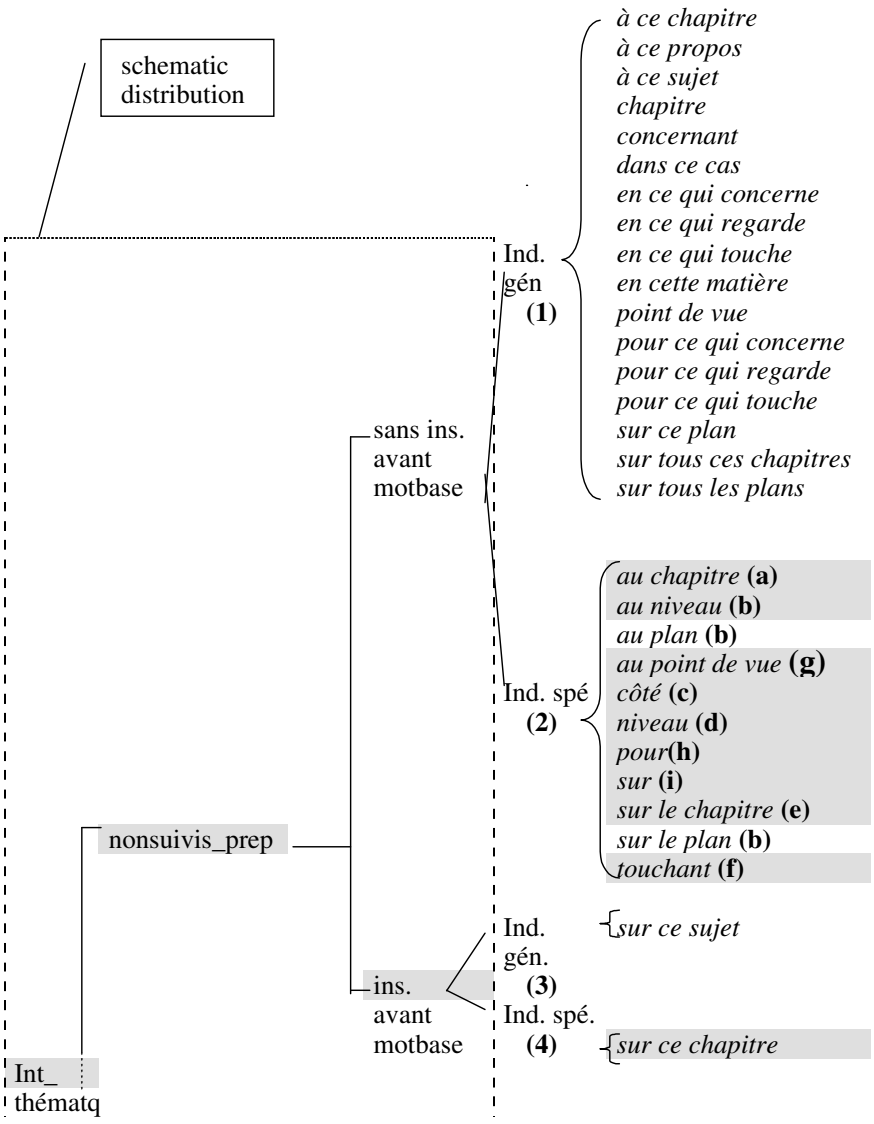
a) Schematic distribution



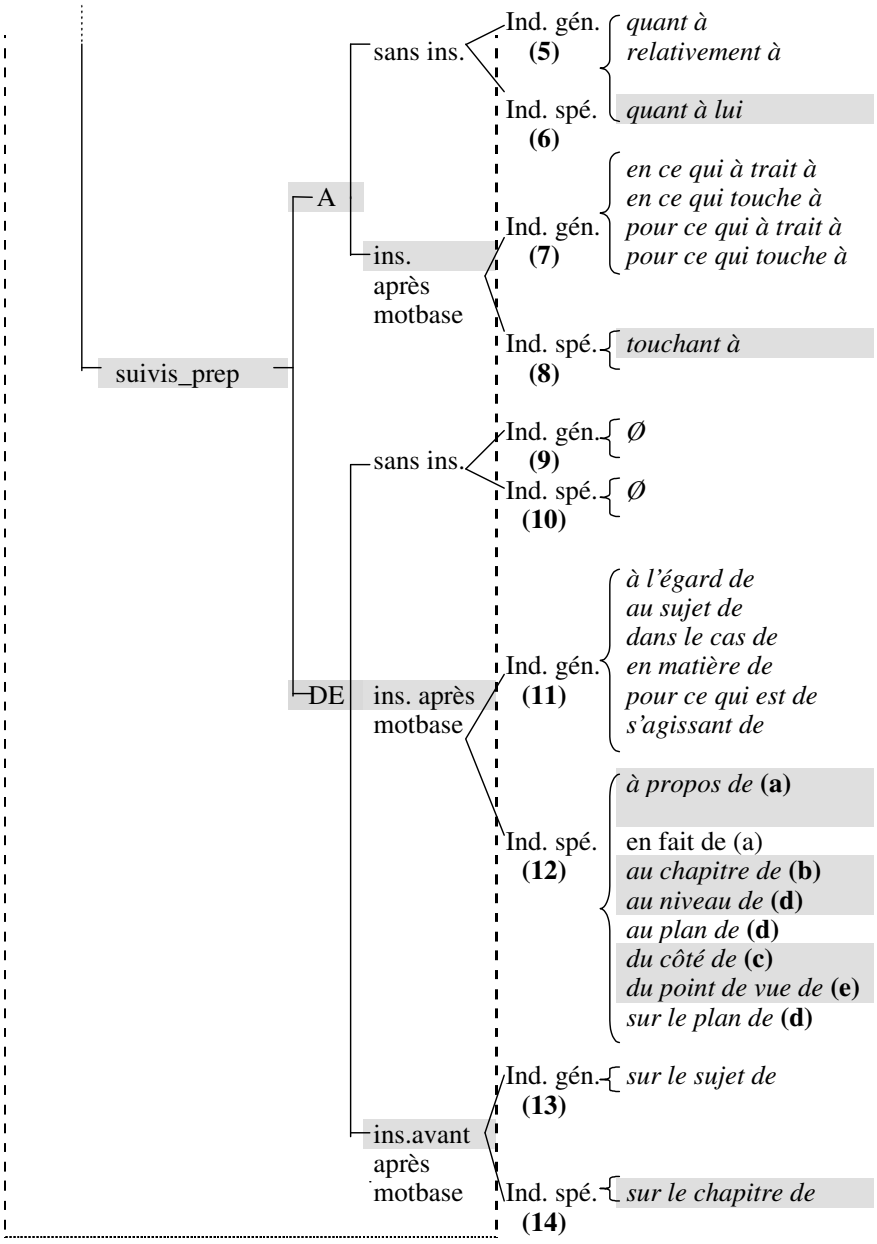
Key:
Ins.: insertion
No ins.: without insertion

b) Distribution tree

Figures from (1) to (14) correspond to subgroups within the category thematic introducers. The letters (a) to (g) indicate that indicators do not combine with the same specific indices and form a class of their own. For example, (1): int_thématq_nonsuivis_prep; (12a): int_thémtq_suivis_prep_de_ins_àproposde.



Thematic Introducers



Key:
Ind. gén.: general indices
Ind. spé.: specific indices
ins. avant motbase: insertion before the word-base

ins. avant/après motbase: insertion before and after the word-base
Int_thématq (introducteurs thématiques): thematic introducers
nonsuivis_prep: not followed by a preposition
sans ins. avant mot-base: without insertion before the word-base
suivis_prep: followed by a preposition
---- schematic distribution