Linguistic expressions as a tool to extract information[1]

Sylvie Porhiel – LaTTICe, LIMSI

When searching for multiple words in a corpus database, we all have encountered the problem of not coming up with 'the right thing'. Indeed, very often the results produced by the computer do not meet the analysis we would have produced as humans. In a text, in order to understand it correctly, we rely not only on the semantic content of the words but also on more grammatical means, which provide reading instructions. Those are the ones we will be dealing with here and we will use French prepositions (*pour ce qui est de, en ce qui concerne, à propos de, au sujet de*, etc.) which have the syntactic properties of being either detached and generally in an initial position or, of being dependent on another morphosyntactic constituent. This syntactic behaviour is of importance in the way in which information is processed and corresponds to two discoursal functions. When detached, these prepositions introduce a thematic (Charolles 1997) and mark an author's pragmatic intentions. When introducing a thematic they can integrate one or more propositions. When they depend on another constituent they simply focalise and do not have that integrative property. The ultimate aim of this ongoing analysis is to engineer a tool using linguistic expressions to extract thematic information.

The purpose of this paper is three fold: to detail how linguistic information concerning the thematic introducers is captured; to show how they help in segmenting texts; and to give a technical overview of the system currently being developed by the LaLIC team, utilising this linguistic information.

1. Capture of linguistic knowledge about thematic introducers

This section analyses the linguistic markers and captures the linguistic information related to them. The data collection goes through different steps.

1.1. Description of linguistic markers

The first step consists of describing the linguistic markers identified as potential thematic introducers (from now on T). Four criteria have to be taken into account:

- Morphological variations of the linguistic pattern which can be broken into:
  - number variations for the few that undertake them: *au chapitre (de), aux chapitres (de), sur le chapitre (de), sur les chapitres (de)* 'in the matter of';
  - aspectual variations: among those, only the variation used are declared, i.e.: *en ce qui concerne* 'as far as... is concerned', *en ce qui concernait, en ce qui concernera* but not **en ce qui a concerné.*[2]

- Case variation: all T can start a sentence and consequently start with an upper case but can also be after a word or a group of words, a comma or a semicolon, etc., in which case they start with a lower case.

- The unbroken or broken structure of the T: written as follows *à propos de*, the constituents of which are only separated by spaces is an unbroken linguistic pattern, whereas *à propos+de*, the constituents of which are linked by the metacharacter '+' is a broken linguistic pattern which matches: *à propos notamment de, à propos tout particulièrement de*, etc. The elements which fit between the basic constituents of the T are called 'insertions'. An analysis of the linguistic markers shows that there is before the base a single point of insertion in which case it is an adjective, and after the base there are two points of insertion, an adverb (*à propos de*) or an adjective and an adverb *(sur le sujet de, sur le chapitre de*). *Pour ce qui est de* accepts insertions such as: *pour ce qui est bien sûr de X* 'as far as of course X is concerned', *pour ce qui est d'abord de X* 'as far as first X is concerned', *pour ce qui est par exemple de X* 'as far as for example X is concerned'.

[1] This research takes place within the project 'Modèle d'exploration sémantique de textes guidé par les points de vue du lecteur' under the direction of G. Sabah. The LIMSI, the LATTICE, the CEA and the LaLIC of the CAMS contribute to it. I am indebted to J.L Minel and T. Samuels who reviewed this paper.

[2] Unlike the English *as far as sth is/are concerned, en ce qui concerne* does not undertake number variation.

- Paradigmatic variations, which concern the prepositions before and after the base and the determiners before the noun-base. Firstly, when *chapitre* is a constituent of a T it can be preceded by two prepositions *a* or *sur*: *au chapitre, sur le chapitre*; and *concerne* by either *en, pour* or *pour tout*: *en ce qui concerne, pour (tout) ce qui concerne*. Indeed, the more paradigmatic variations possible, the more linguistic patterns will have to be captured. *Pour ce qui est de* varying aspectually and paradigmatically, we currently have: pour/pour tout + ce qui + est/était/sera. Only a few T present preposition and determiner variations after the base: *en ce qui touche (à)* is an example of this. Secondly, determiners can vary and be definite, indefinite, and demonstrative. This raises two questions: a) should we consider all determiners for each T and b) should we mix the demonstratives with the others? Because of insertions, the definite and the indefinite article have three forms, respectively *le*[3]*, l', les* and *un, des, d': sur les sujets délicats de X et Y; sur des sujets aussi délicats que ceux de X et de Y, sur un sujet aussi délicat que celui de Y,* etc. As for demonstratives such as in *sur ce chapitre* and *sur ce chapitre du chômage*, they play a resumptive role in texts. This particular role is of importance in text segmentation, which explains why demonstratives are not part of the paradigm of the determiners.

1.2. Pointing towards the correct prepositional phrases

Software can be said to have located the proper prepositional phrase, i.e. a T, when the located phrase possesses the following characteristics: it is syntactically initially placed and prefixed; it sorts information and places it in 'boxes'; and it integrates one or more propositions, even a whole paragraph (Charolles 1997). This section deals with the recognition of the proper prepositional phrases and gives four scenarios which highlight the difficulties of locating the correct ones. The first three occur within the syntactic boundaries of the sentence and the fourth one exceeds this boundary.

First, a system might encounter difficulties as far as syntactic segmentation is concerned. In (1) and (2), the phrase *à propos de* is placed initially and prefixed, meeting the syntactic properties of T:

(1) À propos de la grenouille, qui va s'en occuper ? 'Concerning the frog, who is going to look after it ?'
(2) À propos, de la rouge ou de la bleue, laquelle préfères-tu ? 'By the way, between the red or the blue, which is your favourite ?'

Only the phrase in (1) is a T and the difference between the two phrases lies in the presence of a comma after *propos*. To be extracted the T have to be described correctly so that it accepts insertions but no comma.

Second, the system might have to overcome lexical ambiguities resulting from the polysemy of the base of the T. Let's consider (3)-(5):

(3) Au chapitre des insectes, il est incollable 'On the matter of insects, he is impossible to catch out'
(4) Au chapitre des insectes, on sent que le scientifique est vraiment dans son élément. 'In the chapter about insects, you feel that the scientist is in his element'
(5) Au chapitre des dépenses, les institutions répertorient (...) 'In the section on expenditure, the institutions list (...)'

All *au chapitre de* are syntactically detached, followed by the same plural determiner and a noun, though only (3) contains a T. In French, *chapitre* has several meanings which happen to combine in the same syntactic environment and to be positioned syntactically alike. Nonetheless, their semantic compatibilities differ: *chapitre* when referring to a book can be followed by the topic and the author; *chapitre*, related to a budget is also followed by specific nouns (*recette, dépenses*, etc.) and may also have the name of an administration. As a result, to optimise the chance for a system to locate the right phrase, *au chapitre de* must <u>not</u> be found in combination with an author's name or a specific noun.

Third, the positional characteristic has to be reviewed and the hypothesis of the initial position rephrased. In (6), though not initially, *à propos de* and *au chapitre de* are T as (1) and (3):

(6) Dis donc, à propos de la grenouille, qui va s'en occuper ? 'Hey, concerning the frog, who is going to look after it ?'
(7) Pourtant, au chapitre (des) insectes, il est incollable. 'And yet, on the matter of insects, he is impossible to catch out'

Indeed, in a text, it is fairly common for T to be preceded by 1 to 3 groups of elements such as: spatial expressions, temporal expressions, articulative conjuncts, etc. Currently 24 groups which combine in different ways have been identified: *Par exemple, au chapitre de X (...), Comme, par exemple, en ce qui concerne X (...), Ainsi, par exemple, en matière de X (...), Il en va ainsi, par exemple, pour ce qui touche à X (...).*

Lastly, the fourth difficulty encountered by a system lies in the recognition of thematic utterances from truncated ones, when the sentence prototypically consists of the marker plus its complement. Two cases can be considered: first, the T is placed initially; second, it appears after another group of elements. (8)-(9) illustrate the configuration in which the T, strictly followed by their complements, are placed initially or are preceded by other elements (10):

[3] All the noun base are masculine.

(8) Il me semble qu'il m'a raconté une anecdote. À propos de Staline. 'It seems to me that he told an anecdote. About Stalin.'

(9) [§] À propos de démocratie: Jabotinski se définissait comme un libéral et défendait avec fermeté le système parlementaire. (...). 'About democracy: Jabotinski defined himself as a liberal and battled in favour of a parliamentary system. (...)'

(10) J'aurais besoin de votre aide demain. Notamment en ce qui concerne cette note de service. 'I will require your help tomorrow. 'Particularly as far as this memo is concerned.'

Generally speaking, the prepositional phrase in truncated sentences and thematic ones share characteristics: it has no verb, it is not followed by a proposition, it can be preceded by a group of elements and ends either with [.], [:] or [...]. It also has differences:

- the intonation goes down in truncated sentences (8) (10) and rises in thematic ones (9);
- the referent is not present in the interdiscourse in truncated sentences whereas it is in thematic ones (9);
- they tend to occupy different syntactic positions: in the middle of a paragraph with truncated sentences (8) (10), at the beginning of one with thematic sentences (9);
- truncated sentences require the help of the linguistic context and depend on a morphosyntactic constituent in the preceding sentence (8) (10), and can introduce an answer; whereas thematic ones do not.
- some groups of elements as well as punctuation tend to favour the interpretation of sentences as truncated or thematic ones: interjections favour a thematic reading, selective markers a truncated one (10); the configuration in which a T is followed by [...] favours a thematic reading: (Dis donc), à propos de femme... '(Hey) concerning women...'
- truncated sentences do not have an integrative property, thematic sentences do.

2. Thematic introducers and text segmentation

This section is about the segmentation of texts in large corpora (Le Monde Diplomatique, Frantext) and deals with frame openers and closers; examples are used to illustrate these.

2.1. T as frame openers and text segmenters

To use T as signals for text segmentation, it is necessary to consider their place in the text and their combination with other markers found in the cotext.

As with time and locative adverbials (Virtanen 1992), a series of initial T can either appear within paragraphs (11) or start paragraphs (12). As such, they create text strategy continuity (Virtanen 1992):

(11) [§][4] Marie s'est résolue à quitter sa maison et à aller vivre en appartement. Pour ce qui est de ses meubles, elle laisse son fauteuil à bascule à sa femme de ménage. Le secrétaire sera pour Ludovic. La salle à manger revient à sa nièce. Enfin, elle gardera la télé. Concernant sa voiture, elle la donne à son petit fils.(...), etc. Quant à son chien, le voisin s'en occupera. 'Marie brought herself to leave her house and to move into an apartment. As regards her furniture, she leaves her rocking chair to her cleaner. The secretaire is for Ludovic. The dining room suite goes to her niece. Lastly, she will keep the TV. As far as her car is concerned, she gives it to her grandchild (...) etc. As for his dog, the neighbour will take care of him.'

(12) [...] Cette situation a des conséquences décisives sur trois variables: X, Y et Z. 'This situation has decisive consequences upon three variables: X, Y and Z:
[§] En ce qui concerne X, (...) 'As far as X is concerned, (...)'
[§] Quant à Y, (...) 'Regarding Y (...)'
[§] Enfin, pour ce qui est de Z, (...) 'Lastly, concerning Z (...)'

On the one hand, (11) starts with a sentence, which does not tell how the information will be broken down: the referents are implicit. As each referent is introduced by a T, *pour ce qui est de, concernant, quant à*, a frame opens while closing the previous one. The last T, *quant à*, signals the last item in the list. The frame instantiated by *pour ce qui est de* exemplifies the integrative property of T: the referent *meubles* 'furniture' is a hyperonym encompassing the different pieces of furniture listed. *Enfin* 'lastly' heralds firstly the end of the furniture list, and secondly the end of the scope of the T. (12) provides an example of a derived thematic progression: the paragraph finishes with an introducing sentence explicitly specifying that three *variables* X, Y and Z, will be reused and developed in the following paragraphs. Each three T, *en ce qui concerne, quant à, pour ce qui est de*, prefixed and at the beginning of a paragraph or following a connector (*enfin*), signals a shift from one variable to the next. The paragraph structure is of importance as the idented line tells the reader that s/he has just dealt with a sense unit and that s/he is going to start on a new one. However, if the notion is to be of use as a segmentation indicator, it does not have to be overstressed: first because typographical paragraphs and thematic ones do not always coincide and second, because stylistic and balancing parameters enter the matter (Bessonnat 1988, Virtanen 1992). In both these examples T open a series of thematic frames which are examples of thematic strategic continuity. But each example is different as to the level of

[4] [§] indicates the beginning of a paragraph.

structuration. In (11) a structuration indicates a shift from one particular point to another one within the paragraph in accordance with the first sentence; and in (12) the structuration is running in more than one paragraph[5], indicating a thematic break each time.

As different groups of elements are likely to precede T, it will be necessary to find out which ones play a part in the structuration of text. Here, the role of *enfin* 'lastly' has to be outlined. In (11), it signals the end of a list as well as the end of the integration under *pour ce qui est de*. In (12), it signals, with redundancy, the shift into the last variable; it also opens the last frame with the T *pour ce qui est de*. The scope of a connector such as *enfin* is different when within a paragraph or placed initially in a paragraph (Bessonnat 1988).

To conclude, T are linguistic features which open new frames and consequently close the already instantiated one; they mark textual boundaries; they segment texts on different levels (within a paragraph with a linear thematic succession or across paragraphs with a hierarchical thematic one); they can be preceded by groups of elements, the place of which brings scope variation; and in addition the place of T indicates a textual organisation on the part of the text producers, which makes them reliable indicators for text summarisation.

2.2. Frame closers

The opening boundary is relatively easy to identify as it overlaps with the presence of the T in a text. Things are different, though, as far as the closing boundary is concerned. Indeed, even if a T is at the beginning of a paragraph, it does not necessarily means that it integrates all the propositions of that paragraph. Markers likely to induce the closing of a thematic frame are for example: a) T themselves among which some tend to indicate the end of a list, *quant à* (11) (12), b) logical connectors (11) (13) (16), c) spatial and temporal adverbials (14), resumptive anaphors (15) (17a), discourse markers (17c), sentence adjuncts (17b), aspectual changes, typographical means, etc.

(13) [...] Cette situation a des conséquences décisives sur trois variables: X, Y et Z. 'This situation has decisive consequences upon three variables: X, Y and Z:
[§] En ce qui concerne X, (...) 'As far as X is concerned, (...)'
[§] Quant à Y, (...) 'Regarding Y (...)'
[§] Enfin, Z, (...) 'Lastly, Z (...)

(14) [§] En ce qui concerne l'allocation des ressources, en Côte d'Ivoire, l'excédent (...). En Corée du Sud (...) [§] '[§] As far as the resource allowance is concerned, in Ivory Coast, the excess (...). In South Korea (...) [§]'

(15) [..] En ce qui concerne les déchets toxiques, les autorités affirment (...). Mais cet argument ne tient pas (...) '[§] As far as toxic waste is concerned, authorities claim (...) But such an argument does not hold (...)'

(16) À propos des sous-marins, les auteurs (...) À leurs yeux, la marine (...) 'Concerning submarines, the authors (...) According to them, the Navy (...)'

(17) [§] S'agissant de la coopération, le président s'est félicité (...) 'As regards co-operation, the president was pleased (...)'
a) Cette co-operation (...) 'This co-operation (...)'
b) Personnellement, nous aurions (...) 'Personally, we would have (...)'
c) Mais revenons à X (...) 'But to return to my point (...)'

These frame closers, all text segmenters interacting with one another and taking part in discourse management, have a double property: while indicating a break they also signal textual continuity, then coherence. In (13) the connector *enfin*, not followed by a T, closes the preceding frame and opens the last one in the derived thematic progression. (14) illustrates the case when more than one frame appears at the beginning of a sentence, being subordinate to another. In this instance, *en ce qui concerne l'allocation des ressources* gives the general topic of the paragraph while the locative adverbials, *en Côte d'Ivoire* and *en Corée du Sud*, are subordinate and deal with particular points related to the topic. In (15)-(17) the presence of the writer is noticeable through resumptive anaphors preceded (or not) by a logical connector (15) (17a), sentence adjuncts indicating an attitude (17b), discourse markers (17c) which signal the reintroduction of a previous topic of the discourse.

There are also cases where no frame closers appear on the textual surface. The linguistic tools which have been listed so far are then of little use. A solution can be found in looking at changes in the text vocabulary. Such experiments have been conducted by Ferret, Grau and Masson (1998). They worked out two methods, the outcome of which varies according to the type of texts. Ideally, the combination of properly recognised linguistic tools (T and connectors) and a statistic method should optimise the results produced in the segmentation of text, as well as improve the thematic coherence of the summary.

3. Technical overview

This section describes the architecture of the ContextO software and gives a technical overview of how texts are processed.

[5] Unless with spatial or temporal adverbials it is rare to find a text in which T run from the beginning till the end.

3.1. The architecture

The ContextO software runs under the Filtext platform (figure 1) (Minel et al., forthcoming) both of which have been developed by the LaLIC team in Paris IV. It 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 contextual rules written according to the contextual exploration method (Desclés 1997). In ContextO, the database is used to store relevant linguistic data in order to locate T. This linguistic knowledge is grouped into classes and individual classes can make compound classes in order to better describe markers. Here are examples of classes and sets of items belonging to them :

&marqueurs_liaison = {cependant, Cependant, enfin, Enfin}
&prep_de = {de, d', du, des}
&ponctuation = {,}
&a_propos_min = {à propos, en fait}
&introducteurs_thématiques = {&avant_motbase_pdg_en_maj+ce qui +&vb_être,
&avant_motbase_pdg_en_maj+ce qui +&vb_concerne}[6]

Contextual rules are contained in Java methods and use the class information to process texts. The notion of context is a contextual one : a context is determined by indicators (here T) and constituted of sentences not necessarily adjacent. It is in that delimited space that constraints are applied. For example to locate the T *à propos de* et *en ce qui concerne* in a text, the rules R1 and R2 have to be written :

R1 extracts sentences in which no comma appears after *à propos* but still allows for insertions between *à propos* and *de*:

R1: Items belonging to &a_propos_maj = T
IF placed initially
AND IF &a_propos_maj is not followed by an item of &ponctuation
AND IF &prep_de appears within 5 words

R2 extracts sentences in which &introducteurs_thématiques are preceded by *enfin, cependant*

R2: Items belonging to &introducteurs_thématiques = a T
IF in the left context there is an item belonging to &marqueurs_liaison

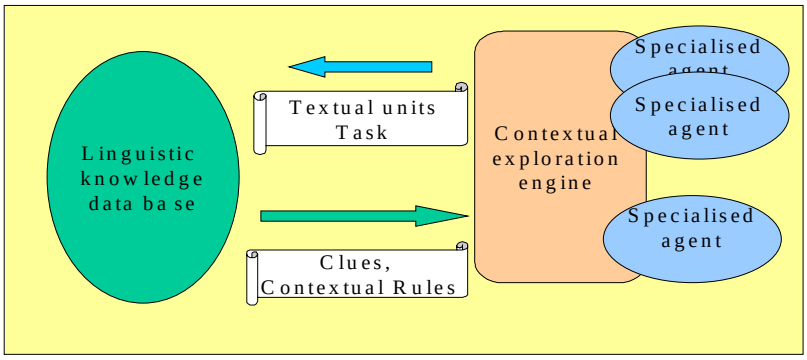

Figure 1. FilText Architecture

3.2. Processing overview

The following sentences used as a corpus will show how a text is actually processed by the ContextO software when the specific task 'recognition of thematic frames' has been chosen: *À propos de la grenouille, qui va s'en occuper ? , À propos, de la rouge ou de la bleue, laquelle préfères-tu ?, À propos notamment des chemins de fer, l'État a décidé (...), Cependant, en ce qui concernait ce problème particulier, on aurait pu (...).* The text is first segmented into its component sentences. Each sentence is then analysed by applying the individual rules. The results of applying our example rules R1 and R2 to the corpus will be as followed:

[6] This implies that the mentioned classes as well as the items pertaining to them have been declared beforehand: &vb_concerne = {concerne, concernait, concernera}, &vb_être : {est, était, sera}, &avant_motbase_pdg_En_maj = {En, Pour, Pour tout}. Only the relevant *tenses* of the verbs are declared

Sentences	R1	R2
À propos de la grenouille, qui va s'en occuper ?	True	False
À propos, de la rouge ou de la bleue, laquelle préfères-tu ?	False	False
À propos notamment des chemins de fer, l'État a décidé (...)	True	False
Cependant, en ce qui concernait ce problème particulier, on aurait pu (...)	False	True

Sentences where either rules R1 or R2 are true will receive a 'frame' tag and thus be extracted.

The approach presented is based on linguistic knowledge in order to analyse texts and more specifically to automatically segment texts without relying on a field of application. We first detailed the analytical steps required for a system to identify T and not a phrase of matching characters. The second part, focused on frame openers and closers. T can be used to open text segments which can be closed, for example, by other linguistic tools. As linguistic tools are not always available on the textual surface it is suggested that combining the use of T and statistic methods would lead to more promising results and should improve the thematic coherence of summarised texts. The last part gave a technical overview of the implementation of the captured linguistic knowledge in the ContextO software developed by the LaLIC team. The next step of this research will consist in increasing the list of T, in refining the contextual rules and in testing these rules in full texts.

References

Bessonnat D 1988 Le découpage en paragraphes et ses fonctions. *Pratiques* 57: 81-105.

Charolles M 1997 L'encadrement du discours - Univers, champs, domaines et espace. *Cahier de recherche linguistique* 6.

Desclés JP 1997 Système d'exploration contextuelle. Co-texte et calcul du sens, Presses Universitaires de Caen, pp. 215-232.

Ferret O, Grau B, Masson N 1999 Thematic segmentation of texts: two methods for two kinds of texts. In *Proceedings of the ALC-COLING'98*, Montréal, pp. 392-396.

Minel JL, Descles J.P. 2000 Résumé automatique et filtrage des textes. In Pierrel J.M. (ed), *Ingénierie des langues*. Paris, Hermès, pp. 253-270.

Minel et al. (forthcoming) Résumé automatique par filtrage sémantique d'informations dans des textes. Présentation de la plate-forme FilText. *Technique et science informatique* 3.

Virtanen T. 1992 *Discourse Functions of Adverbial Placement in English*. Abo, Abo Akademi University Press.