



HAL
open science

Vers un outil de visualisation de la dynamique textuelle : l'exemple des phénomènes citationnels et modaux

Marie Chagnoux, Delphine Battistelli

► To cite this version:

Marie Chagnoux, Delphine Battistelli. Vers un outil de visualisation de la dynamique textuelle : l'exemple des phénomènes citationnels et modaux. 2008. hal-00336859v1

HAL Id: hal-00336859

<https://hal.science/hal-00336859v1>

Submitted on 5 Nov 2008 (v1), last revised 6 Nov 2008 (v2)

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Vers un outil de visualisation de la dynamique textuelle : l'exemple des phénomènes citationnels et modaux

Delphine Battistelli* — Marie Chagnoux**

** Equipe LaLIC FRE2919
Université Paris-Sorbonne
Maison de la Recherche
28 rue Serpente
F-75006 Paris
delphine.battistelli@paris-sorbonne.fr*

***IRIT
Université de Toulouse
118 route de Narbonne
F-31062 Toulouse-Cedex4
marie.chagnoux@irit.fr*

RÉSUMÉ. Nous proposons d'exposer ici une méthode d'investigation du phénomène des différentes prises en charge énonciatives (plus particulièrement, le cas des citations) et modales à l'œuvre dans les textes. Nous mettons l'accent sur la structure hiérarchisée des segments textuels qui en résulte ; nous la représentons d'une part sous forme d'arbre et d'autre part sous forme de graphe. Ce dernier permet d'appréhender la dynamique interprétative d'un texte comme un cheminement qui s'opère entre différents niveaux de discours au fur et à mesure de la lecture syntagmatique. Cette approche prend toute sa légitimité dans le cadre d'une plate-forme de navigation textuelle.

ABSTRACT. We propose a methodological framework for analyzing and representing the concept of commitment, which is one of the features characterizing textual structure. We emphasize the hierarchical structure of textual segments commitment conveys to. We represent it first as a tree and then as a graph. The latter enables us to access the enunciative textual dynamics, as it shows the path followed through different discursive levels during the syntagmatic reading of a text. Our approach is well-founded for textual navigation platforms.

MOTS-CLÉS : linguistique textuelle, prise en charge énonciative, représentation sémantique.

KEYWORDS : textual linguistics, enunciation, semantical representation.

1. Introduction

Les applications de TAL qui visent à accéder au contenu sémantique de documents traitent généralement l'information au niveau local : la recherche d'information et le résumé automatique procèdent par extraction de segments textuels plus ou moins longs à partir d'un texte ou d'un corpus de textes, les repérages de citations ou de lieux spatiaux opèrent par extractions d'entités nommées, l'annotation automatique porte sur des segments que des balises isolent du contexte discursif, etc. Or l'hypothèse qu'accéder à l'information consiste à extraire le segment pertinent de sa structure globale pose un certain nombre de problèmes déjà clairement identifiés, comme celui des anaphores auquel est consacrée une vaste littérature (Busquets *et* Hardt, 2005). Plus largement, elle pose le problème de la continuité référentielle, maintenue ou au contraire brisée par rapport au texte de départ, qui naît à la lecture d'un document composé d'une juxtaposition de segments extraits, comme a pu l'exposer (Battistelli *et* Minel, 2006).

Parmi les problèmes liés à la rupture de la continuité référentielle, celui de la prise en charge énonciative et modale constitue l'un des enjeux posés au TAL : non traité, il peut en effet, au même titre que celui des anaphores, conduire à briser une certaine continuité référentielle, voire à induire des raisonnements erronés si le contexte est ignoré. Prenons ainsi par exemple le segment « l'appel à un cessez-le-feu urgent et non pas immédiat constitue un échec » tiré de l'extrait présenté dans la figure 1.

La Maison-Blanche s'est, elle, montrée satisfaite des résultats de la réunion de Rome et a réfuté que l'appel à un cessez-le-feu urgent et non pas immédiat constitue un échec.

Figure 1. Extrait du texte A¹

Si le segment extrait n'est pas inscrit dans une continuité référentielle, les informations inférées sont (i) que l'appel à cessez le feu est un échec ; (ii) et que l'auteur de l'article assume cette information. Or, non seulement l'indice lexical « réfuter » modifie la valeur du segment (autrement dit, l'appel n'est pas un échec), mais de plus la présence d'un second énonciateur « La Maison-Blanche » modifie la prise en charge de l'information : l'auteur de l'article ne prend en charge que le principe de la citation et non son contenu.

Pour permettre l'analyse sémantique des propositions d'un texte, il apparaît donc nécessaire de disposer d'un outil qui permet d'associer le segment à traiter –

¹ Les références complètes des textes dont les extraits illustrent cet article sont données en annexe. Pour une meilleure lisibilité du document, les articles cités sont désignés par une lettre.

ici la proposition – à une représentation de la structure qui véhicule des informations cruciales pour son traitement.

Nous proposons d'exposer dans cet article une méthodologie d'analyse et de représentation de la structure énonciative et modale de textes. Cette méthodologie repose sur le repérage, à partir d'indices linguistiques dits de rupture, de différents segments textuels dont il s'agit, au travers d'une représentation sous la forme d'un graphe, de montrer les relations qu'ils entretiennent en suivant l'ordre syntagmatique de lecture du texte. Nous illustrons les principes de cette méthodologie en nous appuyant sur un corpus d'articles portant sur le conflit israélo-libanais pendant l'été 2007 (*cf.* annexe). Dans une première partie nous montrons pourquoi il est nécessaire de (et comment) prendre en compte les mécanismes énonciatifs et modaux dans les applications du TAL pour accéder à l'information dans les textes et quelle est la nature de ces mécanismes. La seconde partie présente la méthodologie opératoire que nous proposons. La troisième partie expose les différentes étapes d'implémentation (réalisées ou en cours) qui permettent de construire automatiquement le graphe à partir d'un texte. Nous concluons sur le type de représentation du texte, d'ordre hiérarchique, mis en évidence au travers de cette méthodologie.

2. De la nécessité de prendre en compte la structure énonciative et modale d'un texte dans les applications du TAL

2.1. Quelques exemples

Un texte s'inscrit dans une certaine dynamique énonciative et modale :

- il est la trace d'au moins un acte d'énonciation, celui accompli par l'énonciateur principal qui prend en charge l'ensemble du discours ;
- certains segments ne sont pas complètement assumés par cet énonciateur, soit qu'un énonciateur second est convoqué au terme d'une citation, soit que ces segments sont assujettis à un certain degré de plausibilité ou d'intentionnalité.

Aussi, comme l'exemple précédent le montre, extraire une phrase ou une proposition de son contexte énonciatif et/ou modal peut entraîner des imprécisions et des contresens. Dans le cadre d'applications informatiques comme la Recherche d'Information, cela peut également conduire à des conflits. Ainsi, à partir de notre corpus, il est possible d'extraire les deux phrases des figures 2 et 3.

Le Hezbollah a capturé deux soldats israéliens qui patrouillaient dans des jeeps blindées sur le territoire libanais à la frontière avec Israël.

Figure 2. Extrait du texte B

Le mercredi 12 juillet, 8 soldats de Tsahal ont été tués et 2 ont été kidnappés, au cours d'une attaque du Hezbollah à l'intérieur du territoire israélien, près de la frontière israélo-libanaise.

Figure 3. Extrait du texte C

Toutes deux réfèrent à la même situation : la capture des deux soldats au Moyen-Orient le 12 juillet 2006. Cependant, selon l'énonciateur, le lieu de la capture est différent : pour l'association France Palestine, la capture a eu lieu sur le territoire libanais alors que pour l'Ambassade d'Israël en France, elle s'est déroulée à l'intérieur du territoire israélien. Un système qui ne gère pas la prise en charge énonciative ne peut gérer le conflit d'informations.

Même si ce n'est pas toujours le cas dans l'absolu, la reconnaissance de la source peut apparaître comme triviale dans certains cas d'applications inter-documents puisqu'il suffit d'annoter le texte pour le doter de cette information, comme cela est proposé par exemple pour le Web sémantique. La problématique est différente quand le conflit apparaît au sein d'un seul et même texte comme l'illustre l'extrait de la figure 4.

A propos du lieu de l'enlèvement des deux soldats, les versions diffèrent. Les Israéliens affirment qu'ils ont été capturés près de la ferme collective de Zarit en territoire israélien tout près de la frontière libanaise. De son côté, la police libanaise soutient que la capture s'est produite dans la région de Aïta al-Chaab en territoire libanais donc, proche de la frontière libano-israélienne où une unité israélienne avait pénétré le matin même.

Figure 4. Extrait du texte D

Il apparaît ainsi comme nécessaire de préciser l'organisation des phénomènes citationnels (ou modaux) à l'intérieur d'un texte pour indiquer les faits ou événements décrits qui sont effectivement pleinement assumés par l'auteur d'un texte et, si non, avec quel type de distanciation. En cela, il s'agit de dépasser le « simple » repérage des phénomènes citationnels (ou modaux) comme cela peut être réalisé dans la plupart des systèmes actuels (voir partie 2.2.1). En ayant ainsi organisé – sous une forme hiérarchique comme nous le verrons plus loin - les différentes « prises en charge énonciatives et/ou modales » (pour reprendre une terminologie courante), un traitement informatique est à même de rendre compte de la véritable attribution de tel ou tel propos à tel ou tel énonciateur, que ce soit dans le cadre d'un traitement inter-documents ou intra-document. C'est à ce problème d'identification de la structure énonciative et modale d'un texte que nous nous proposons d'apporter des éléments de réflexion sinon de solution opératoires.

L'objectif de cet article étant de s'intéresser aux mécanismes qui président à la construction de cette structure, nous ne présentons pas les analyses détaillées de linguistique générale sur lesquelles se fondent nos travaux (Desclés, 1995 ; Guentchéva, 1996 ; Desclés et Guentchéva, 2000 ; Chagnoux, 2006) mais les situons en regard des approches actuelles du TAL et de la linguistique textuelle.

2.2. Analyse des phénomènes de prise en charge énonciative et modale dans divers travaux

2.2.1. Travaux en TAL

Comme nous l'avons souligné en introduction, les travaux actuels en TAL appréhendent rarement les textes comme des structures hiérarchisées et dynamiques. Quant aux outils de traitement qui s'intéressent aux phénomènes citationnels et modaux, ils portent exclusivement sur des structures locales non articulées. Ils sont soit axés sur le repérage des citations et des locuteurs comme dans (Mourad, 2000 ; Giguët *et* Lucas, 2004), soit sur l'annotation d'ordre essentiellement modale de segments textuels du niveau de la phrase comme dans (Sauri *et al.*, 2005 ; Wilson *et* Wiebe, 2005). Nous nous situons davantage dans la continuité de (Jackiewicz, 2006) qui propose une typologie des discours rapportés en précisant que ce type d'analyse peut « enrichir le formalisme de certaines théories qui cherchent à rendre compte des représentations discursives »². Dans une autre perspective, l'approche bibliométrique aborde le repérage de citations par le biais de méthodes statistiques sur des ensembles de références bibliographiques, que ce soit en vue d'évaluer quantitativement des publications scientifiques (Rostaing *et al.*, 1999) ou de créer des cartes relationnelles de documents ou d'auteurs (Small, 1973 ; Sandor *et al.*, 2006). L'objectif est de repérer qu'il y a présence d'une citation sans que la portée du segment citationnel soit importante : la seule présence d'une référence bibliographique (ou d'un lien hypertexte dans le cas des applications web) suffit à détecter la citation. Nous pouvons également évoquer les travaux de classification automatique de citations dans les publications scientifiques (Teufel *et al.*, 2006) qui cherchent à caractériser les citations selon la relation entre les travaux (comparaison, critique, héritage, etc.). Tous ces travaux ont en commun d'extraire ou de repérer au niveau local.

2.2.2. Travaux en linguistique textuelle et modèles du discours

De son côté, la linguistique textuelle s'intéresse depuis toujours à la description des modes de structuration du discours (cf. Péry-Woodley, 2001). La méthodologie mise en œuvre consiste à rendre compte de modes de structuration textuelle au travers de l'analyse de configurations d'indices linguistiques présents dans les textes. Elle se distingue en cela de celle mise en œuvre dans le cadre de modèles du discours comme ceux de la RST (Mann et Thompson, 1988) ou de (Hobbs, 1990) qui insistent sur l'indépendance du type de structures qui les intéressent par rapport à

² Dans une perspective de représentation des connaissances, cette typologie a été formalisée dans (Jackiewicz *et* Laublet, 2007).

toute marque linguistique³. A notre connaissance cependant, dans le premier cas comme dans l'autre, aucun travail n'a été spécifiquement proposé pour l'analyse et la mise en évidence de la structure énonciative et/ou modale d'un texte, sinon dans les travaux de (Wolf et Gibson, 2005), qui s'inscrivent dans la lignée de ceux de (Hobbs, 1990), et prennent en considération des indices marqueurs d'une relation de discours dite d'« attribution ». Mais il semble que l'analyse de cette relation ne soit menée que de manière superficielle puisque le but ultime de ses auteurs est plutôt de la situer en regard de l'analyse de l'ensemble des autres relations de discours définies.

Il reste à souligner que, même si les analyses proposées en linguistique du discours ne visent généralement pas l'automatisation, les systèmes de navigation textuelle, en tant qu'ils proposent de s'appuyer sur les structures discursives identifiées (le plus souvent de type hiérarchique, notons-le), s'intéressent tout particulièrement à ce type de travaux. Ils proposent ainsi une alternative aux systèmes classiques de résumé automatique et de questions-réponses en permettant une navigation à l'intérieur d'un document parmi des segments annotés inscrits dans leurs co-textes (Bilhaut *et al.*, 2003 ; Crestan *et al.*, 2004 ; Minel *et Couto*, 2004). Cependant, il n'existe actuellement à notre connaissance, là aussi, aucun travail qui prenne en compte la *structure* énonciative et modale des textes.

Etant donné que nous cherchons à relier explicitement notre démarche à celle entreprise dans le cadre du développement de telles plate-formes, nous devons donc proposer des solutions opératoires pour la délimitation de segments textuels homogènes constituant tel ou tel cadre énonciatif ou modal. Cela nous conduit nécessairement à aborder alors la question de la distinction entre marqueurs de « cohésion » (ou de « continuité ») et d'« incohésion » (ou de « rupture »). Or ces marqueurs sont en grande partie de nature temporelle. Cela nous inscrit donc dans la problématique de l'analyse de la temporalité au sens large dans les textes, indissociable en réalité de celle de la modalité et de l'énonciation comme le rappelle par exemple (Gosselin, 2005). Nous proposons dans la section suivante de rappeler brièvement les liens étroits à considérer entre temporalité, énonciation et modalité. Ces liens sont à même d'explicitier le fondement de notre méthodologie qui consiste à identifier différents types de « référentiels temporels » selon la terminologie de (Desclés, 1995) sur lesquels sont validées des propositions d'un texte. Un ensemble de propositions validées sur un référentiel donné forme alors un segment textuel homogène⁴.

³ Il est à noter que l'approche de la SDRT (Asher, 1993) est à cet égard quelque peu différente ; il existe en effet depuis plusieurs années une volonté de lier le modèle aux instanciations langagières des relations de discours (cf. Asher *et* Lascarides, 2003).

⁴ Nous utilisons le terme de « segment » pour désigner un extrait du texte quelque soit sa taille : un segment peut contenir une seule proposition, plusieurs phrase ou même plusieurs paragraphes. Ce terme nous permet d'insister sur le principe de cohésion sémantique en nous soustrayant à la terminologie syntaxique.

2.3. Temporalité, énonciation et modalité

Tout texte, en plus d'être ancré dans le temps, est structuré temporellement. Cette structuration temporelle participe d'ailleurs sa cohérence. Aux modes de structuration proprement temporels des discours, s'ajoute le fait que les situations décrites peuvent être sujettes à certains degrés de plausibilité ou de véracité (voir par exemple l'utilisation du mode conditionnel, ou de certains adverbes comme « peut-être ») ; on fait référence ici au domaine de la modalité. Des situations décrites comme potentielles, par exemple, ne s'inscrivent alors plus dans la continuité des situations perçues comme réalisées ou en cours de réalisation. Des opérations de « rupture » d'ordre modal comme temporel structurent les textes. Les domaines de la temporalité et de la modalité sont donc étroitement liés dans la mesure où l'analyse temporelle d'un discours suppose de pouvoir distinguer les situations décrites comme réalisées, potentielles, à venir, atemporelles (du domaine du définitoire par exemple), ... et éventuellement comme non directement prises en charge par l'énonciateur principal mais par des énonciateurs seconds. Elles doivent donc être analysées comme situées sur des référentiels énonciatifs ou modaux différents qui en outre ne sont pas nécessairement reliés entre eux (sinon au moins au référentiel de l'énonciateur principal) ; on parle aussi de « ruptures discursives ». Il subsiste pourtant bien la perception d'une certaine cohérence du texte, qui résulte de la faculté de l'allocutaire ou du lecteur à construire une représentation complexe qui articule différents référentiels entre eux.

La problématique des différentes prises en charge énonciatives et modales à l'œuvre dans les textes relève donc du principe sémantique de cohérence temporelle : il s'agit de rendre compte des articulations complexe entre différents niveaux de discours dans un texte. Elles se traduisent par exemple par la présence de certains changements de suites de temps verbaux, de verbes introducteurs de discours directs et indirects, de guillemets, etc. Ces différents niveaux de discours sont alors à considérer comme autant d'*espaces de validation* (que l'on ouvre et ferme à l'intérieur d'un même texte). Ils ont été déjà évoqués dans la littérature notamment en tant que « référentiels temporels » (Desclés, 1995 ; Battistelli *et al.*, 2006a) ou en tant que « cadres de discours » (Charolles, 1997). Dans le cas où la principale finalité du traitement est de repérer « qui dit quoi » sans chercher justement à décrire le « quoi » (et les liens temporels qui le lient à d'autres événements relatés), on parle plus précisément de repérage des différentes « prises en charges énonciatives et modales ».

2.4. Notre proposition

C'est dans ce dernier paradigme que nous nous situons, tout en nous appuyant sur la typologie des référentiels proposée par (Desclés, 1995) et précisée par (Chagnoux, 2006). Nous proposons en outre, comme nous l'avons déjà souligné, d'articuler explicitement ces différentes prises en charge repérées au sein d'une structure hiérarchisée. Nous illustrons les principes de notre méthode en nous appuyant sur les travaux de linguistique qui distinguent deux types de rupture : des

ruptures énonciatives qui font varier l'énonciateur en charge des propos rapportés (Desclés *et* Guentchéva, 2000) et des ruptures modales qui font varier le degré de prise en charge des propos d'un même énonciateur (Kronning, 2003 ; Gosselin, 2005). Les deux types de ruptures présentent pour un énonciateur la possibilité de se distancier de ses propos, soit en les imputant à un tiers dans le cas des phénomènes citationnels, soit en modélisant le contenu énonciatif dans le cas des phénomènes modaux. Tout segment cohérent est vu comme s'inscrivant sur un « référentiel » au sens où le définit la Physique : un système de coordonnées spatiales et temporelles lié à un observateur. Ainsi, un même événement diffère selon l'emplacement de l'observateur et les repères associés à son référentiel.

3. Du texte au graphe de visualisation de sa dynamique énonciative et modale : principes méthodologiques

3.1. Préliminaire : à propos des marqueurs de « cohésion » et d' « incohésion »

Dans l'analyse d'un texte, il est opportun de distinguer deux types de marqueurs : des marqueurs de « cohésion » et des marqueurs d' « incohésion ». Les marqueurs de cohésion assurent l'homogénéité d'un segment de texte en inscrivant les situations dans une chronologie ; c'est par exemple le cas des connecteurs, des adverbes, des temps verbaux concordant en systèmes (système d'alternance de verbes à l'imparfait et au passé simple par exemple). Les marqueurs d' « incohésion » indiquent des « ruptures ». Ils peuvent être soit de simples marqueurs comme les indices typographiques tels que les guillemets, suivis ou non de deux points, soit des ensembles de marqueurs⁵, soit encore des temps ou des modes verbaux.

Les figures 5, 6 et 7 montrent le résultat de la segmentation selon ces critères des extraits de textes E, F et G respectivement. Les marqueurs de cohésion qui permettent d'identifier des segments temporels homogènes sont soulignés alors que les marqueurs d'incohésion qui indiquent des ruptures sont en gras. Les segments homogènes ainsi dégagés sont notés *seg_i*.

```
<seg_1>Le 14 juillet, au large de Beyrouth, alors qu'il participait à un blocus naval
et qu'il se trouvait dans les eaux territoriales libanaises le navire israélien Hanit de la
Classe Sa'ar V a été endommagé par un missile C 802, de fabrication chinoise livré
par l'Iran selon l'armée israélienne, tiré par le Hezbollah : 4 marins ont été tués.
</seg_1>
```

Figure 5. Segmentation de l'extrait du texte E

⁵ Lors de la représentation, une configuration syntaxico-lexicale de marqueurs sera notée $\{m^1+...+m^n\}$ alors qu'une liste de marqueurs sera notée $\{m^1, ..., m^n\}$.

L'extrait de texte E offre une alternance de temps du passé (imparfaits et passés composés) relayée par des connecteurs et présente donc de ce point de vue une certaine cohésion. Autrement dit, ce segment textuel est cohésif temporellement. Cet extrait est donc constitué d'un seul segment temporel *seg_1*, délimité de part et d'autre à l'aide d'une balise ouvrante notée `<seg_1>` et d'une balise fermante notée `</seg_1>`.

`<seg_1>`L'Etat hébreu réclame la libération des deux soldats et le retrait du Hezbollah de la zone frontalière. A défaut d'un tel cessez-le-feu dans les 24 heures, `<seg_2>` « toute la région du Proche-Orient sera en danger » . `</seg_2>`, a prévenu le président pro-syrien. `</seg_1>`

Figure 6. Segmentation de l'extrait du texte F

Dans l'extrait de texte F, la présence de guillemets suivis du verbe introducteur *prévenir* marque la présence d'un discours direct. Le texte n'est pas cohésif temporellement : le segment textuel *seg_2* est en rupture avec le reste du segment textuel *seg_1*.⁶

`<seg_1>`A propos du lieu de l'enlèvement des deux soldats, les versions diffèrent. Les Israéliens affirment qu'`<seg_2>` ils ont été capturés près de la ferme collective de Zarit en territoire israélien tout près de la frontière libanaise. `</seg_2>` De son côté, la police libanaise soutient que `<seg_3>`la capture s'est produite dans la région de Aïta al-Chaab en territoire libanais donc, proche de la frontière libano-israélienne où une unité israélienne avait pénétré le matin même`</seg_3></seg_1>`

Figure 7. Segmentation de l'extrait du texte D

Le même type de rupture énonciative est à l'œuvre dans l'extrait D. Trois segments temporels, *seg_1*, *seg_2* et *seg_3*, peuvent être distingués et à chaque segment est associé un énonciateur différent⁷ :

- l'énonciateur premier prend en charge le contenu de *seg_1* : « A propos du lieu de l'enlèvement des deux soldats, les versions diffèrent. Les Israéliens affirment qu' [...] . De son côté, la police libanaise soutient que [...] » ;
- la prise en charge de *seg_2* : « ils ont été capturés près de la ferme collective de Zarit en territoire israélien tout près de la frontière libanaise » est attribuée à « Les Israéliens » ;

⁶ Nous ne faisons pas de différence de traitement entre discours direct et indirect à ce niveau de l'analyse.

⁷ Notre objectif se limite ici uniquement au repérage des ruptures discursives et non à l'identification automatique des entités nommées pour répondre à la question « qui a dit » comme proposé par exemple par (Daille et Morin, 2000).

- la prise en charge de *seg_3* : « *la capture s'est produite dans la région de Aïta al-Chaab en territoire libanais donc, proche de la frontière libano-israélienne où une unité israélienne avait pénétré le matin même* » est attribuée à « *la police libanaise* ».

Même si certains problèmes d'interprétation peuvent se poser – même pour un annotateur humain⁸ - quant à l'attribution de tel ou tel propos à tel ou tel énonciateur, l'analyse de ces textes montre qu'il est possible de s'orienter vers la construction d'un outil de modélisation et de représentation qui associerait chaque segment cohésif à un référentiel (énonciatif ou modal) dans le but de dégager la structure énonciative et modale des textes. Cette structure hiérarchisée rendrait alors compte des ruptures, articulées, qu'un lecteur identifie à la lecture d'un texte.

Nous présentons maintenant de manière plus explicite les principes qui régissent l'identification des segments énonciatifs cohésifs et la construction de la représentation graphique sous forme d'arbre puis de graphe toujours à partir de l'exemple du texte D rappelé dans la figure 8.

A propos du lieu de l'enlèvement des deux soldats, les versions diffèrent. Les Israéliens affirment qu'ils ont été capturés près de la ferme collective de Zarit en territoire israélien tout près de la frontière libanaise. De son côté, la police libanaise soutient que la capture s'est produite dans la région de Aïta al-Chaab en territoire libanais donc, proche de la frontière libano-israélienne où une unité israélienne avait pénétré le matin même.

Figure 8 : *Extrait du texte D.*

3.2. Délimitation de segments textuels cohésifs

Au niveau discursif, le texte est considéré comme une suite de propositions en relation de rupture ou de continuité. La première étape consiste à distinguer les différentes propositions comme dans la figure 9 où le texte est segmenté en propositions numérotées de 1 à 6.

⁸ Nous remercions ici un relecteur anonyme qui nous a signalé que dans le cas du texte de la figure 8, par exemple, se pose le problème de savoir à qui « appartient » la proposition « Et ce ne pourrait être qu'un début ». Ce relecteur suggère qu'en fonction du type de texte et éventuellement d'autres facteurs, ce pourrait être aussi bien à l'auteur du texte qu'à Nasrallah et pose ainsi directement le problème du traitement du discours indirect libre.

P₁	A propos du lieu de l'enlèvement des deux soldats, les versions diffèrent.
P₂	Les Israéliens affirment qu'
P₃	ils ont été capturés près de la ferme collective de Zarit en territoire israélien tout près de la frontière libanaise.
P₄	De son côté, la police libanaise soutient que
P₅	la capture s'est produite dans la région de Aïta al-Chaab en territoire libanais donc, proche de la frontière libano-israélienne
P₆	où une unité israélienne avait pénétré le matin même.

Figure 9. Découpage en propositions de l'extrait de texte D.

Chaque proposition est ensuite qualifiée selon des critères d'annotation fondés sur les types de décrochage de prise en charge et sur la typologie des référentiels définie par (Chagnoux, 2006). Ces critères permettent d'inscrire les propositions sur différents référentiels⁹ :

- Le référentiel énonciatif *RE*, sur lequel le contenu propositionnel est complètement assumé par un énonciateur ; il est dit *global* (noté *RE_g*) quand il s'agit de l'énonciateur premier, et *local* (noté *RE_l*) quand il s'agit d'un discours rapporté, direct ou indirect, ou médiatisé ;
- Le référentiel possible *RP*, sur lequel le contenu propositionnel est considéré comme possible éventuel (dans ce cas, le référentiel est noté *RPE*) ou possible contre-factuel (ici, le référentiel est noté *RPC*) ;
- Le référentiel mental *RM*, où s'inscrivent des propositions introduites par *il pense que*, *il croit que*, etc., à rapprocher des méta-représentations décrites par (Recanati, 2000).

L'identification et l'annotation des segments s'appuient sur la présence de marqueurs linguistiques de différentes natures : temps et modes des verbes conjugués, déictiques, constructions syntaxiques, lexèmes, adverbes, etc. Nous insistons sur le fait que ces catégories fonctionnent en systèmes. C'est l'expression de patrons complexes qui rend compte des opérations de rupture ou au contraire de cohésion et permet (i) d'annoter sémantiquement les propositions ; (ii) de déterminer des contraintes pour propager une annotation à des propositions adjacentes¹⁰.

A partir de cette typologie, il est possible de distinguer trois référentiels énonciatifs distincts dans le texte D :

⁹ L'objectif de cet article étant davantage de s'intéresser aux mécanismes de construction des structures hiérarchiques qu'à l'analyse menée en amont, nous ne présentons qu'une typologie partielle et synthétique de référentiels.

¹⁰ Nous présentons p.19 deux exemples de règles pour illustrer ces principes.

- le référentiel énonciatif global (RE_g), celui de l'énonciateur principal, sur lequel sont validées P₁, P₂ et P₄ ;
- un premier référentiel énonciatif local (RE_{l1}), celui d'un nouvel énonciateur, créé à partir de P₂ et sur lequel est validée P₃ ;
- un second référentiel local (RE_{l2}), celui d'un autre nouvel énonciateur, créé à partir de P₄, sur lequel se situent P₅ et P₆.

Ces trois référentiels forment autant de segments cohésifs : le premier segment est constitué des propositions P₁, P₂ et P₄, le deuxième de la proposition P₃ et le troisième des propositions P₅ et P₆.

3.3. Construction du graphe reliant les différents segments d'un texte

3.3.1 Préambule

La tâche suivant celle de délimitation de segments cohésifs consiste à spécifier les relations entre ces segments. Comme nous l'avons mentionné plus haut, le fait de considérer une structure textuelle hiérarchisée est classique en analyse du discours (Mann et Thompson, 1988 ; Hobbs, 1990 ; Asher, 1993). Ainsi que le souligne entre autres (Cornish, 2006), le « scripteur » réalise linguistiquement une organisation qui passe par la constitution de segments, la hiérarchisation de ces segments et leur mise en relation. Autrement dit, un texte est vu comme un ensemble de segments – constitués de propositions non nécessairement contiguës dans le texte – qui entrent en relation d'inclusion ou de succession¹¹ et il peut donc être représenté à l'aide d'une structure d'*arbre*.

Ce qui nous intéresse ici est de décrire plus spécifiquement, en nous appuyant sur cette structure d'*arbre*, le *cheminement* qui s'opère entre différents niveaux de discours dans un texte au fur et à mesure de sa lecture syntagmatique. Chaque création de référentiel correspond à une rupture de prise en charge énonciative et/ou modale et nous voulons visualiser cette dynamique des ruptures dans l'ordre où elles interviennent. Le cheminement opéré prend alors l'allure d'un *graphe*. Ce type de structure de données pour représenter la dynamique énonciative et modale de textes présente alors selon nous les deux avantages suivants :

- celui de pouvoir représenter les segments discursifs homogènes d'un texte ainsi que les relations hiérarchiques qu'ils entretiennent entre eux ;
- celui d'offrir également la possibilité de visualiser le parcours – donc nécessairement dynamique – de lecture d'un texte entre référentiels distincts en explicitant les indices linguistiques indiquant les ruptures.

¹¹ Ces relations peuvent éventuellement faire l'objet d'une typologie sémantique comme c'est le cas pour les relations dites de discours (narration, élaboration, ...) proposées par exemple par (Mann et Thompson, 1988) ; il reste qu'elles renvoient toutes à une structuration arborescente du texte. Nous renvoyons à (Busquets *et al.*, 2001) pour une présentation comparative de différents modèles hiérarchiques du discours.

3.3.2 Définition d'un graphe

Les nœuds du graphe correspondent aux référentiels identifiés dans le texte. Ils sont numérotés pour indiquer le chemin à l'intérieur du graphe et étiquetés sous la forme RT_{sn} où T désigne le type de référentiel, s désigne le statut du référentiel (global, noté g ou local, noté l) et n le numéro d'ordre d'apparition du référentiel désigné. L'étiquette du référentiel est suivie des différentes propositions qui s'inscrivent dans ce référentiel. Ainsi, $RE_{ll}\{P_3\}$ désignera par exemple le premier référentiel énonciatif local qui contient la proposition P_3 . Le nœud numéroté 1 occupe un statut particulier d'un point de vue théorique. Il correspond en effet toujours au référentiel énonciatif global RE_g , celui de l'énonciateur principal, dont on fait l'hypothèse qu'il est toujours présent, même si aucun marqueur linguistique ne l'introduit explicitement (*cf.* entre autres Desclés, 1980).

Les arcs correspondent aux transitions entre référentiels et sont étiquetés avec les marqueurs qui déclenchent une transition d'un référentiel à un autre. Ils sont donc tous orientés. Ils sont de deux types, selon qu'ils portent une information de type linguistique ou non. Quand un ensemble de marqueurs linguistiques est repéré comme porteur d'une rupture, il étiquette l'arc. Un arc ne portant pas ce type d'information correspondra toujours à un retour d'un nœud vers le nœud immédiatement précédent (et ce faisant correspondra soit à une fermeture de référentiel soit fera partie d'un ensemble d'arcs menant nécessairement à la fermeture d'un référentiel).

Il est à noter que lors de la construction de toute structure sous-jacente à un texte, la principale difficulté réside dans l'identification automatique de la fermeture des segments repérés. Cette difficulté se pose également dans les autres modèles qui rendent compte de la structure textuelle (*cf.* par exemple Charolles *et* Vigier, 2005 ; Le Draoulec *et* Péry-Woodley, 2005) et diverses stratégies sont mises en œuvre pour résoudre ce point. Pour notre part, nous proposons :

- le retour systématique au premier nœud du graphe en fin de texte, ainsi, quel que soit le référentiel sur lequel se situe la dernière proposition du texte, il est considéré comme nécessaire de revenir à RE_g en fermant l'un après l'autre, selon l'ordre parcouru au préalable, tous les référentiels restés ouverts (c'est donc la lecture du point final dans un texte qui déclenche ce processus de fermeture de tous les référentiels s'ils ne l'ont pas déjà été, illustrant ainsi le fait que l'ensemble du texte est énoncé par l'énonciateur principal) ;




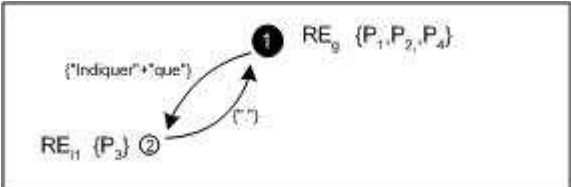
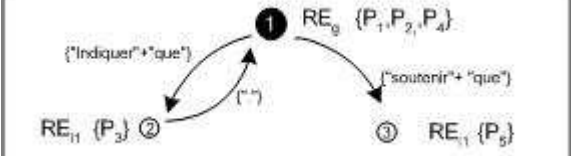
- de considérer l'absence ou la présence de marqueurs d'homogénéité entre la proposition en cours d'examen et la précédente pour décider du traitement à effectuer ; ainsi, des règles permettent de propager une annotation sémantique d'une proposition à une proposition adjacente en fonction de la présence ou de l'absence de marqueurs¹².

¹² Un exemple de règle de propagation est donné p.19.

Un graphe ne propose donc qu'un seul chemin de parcours indiqué par la numérotation des nœuds et en suivant l'orientation des arcs et il commence et finit toujours sur le nœud 1 RE_g. L'algorithme qui permet d'associer un graphe à un texte est implémenté dans un système de règles présenté p.19. Il peut être résumé comme suit : tout texte commence et finit sur RE_g, toute prise en compte d'une nouvelle proposition répond à l'une des actions suivantes : (i) si une marque de rupture est identifiée, il y a ouverture d'un nouveau référentiel ; (ii) si une marque de rupture de type «fermeture » du référentiel est identifiée, il y a retour vers le référentiel précédent ; (iii) selon le type de référentiel, sans marqueur explicite, il y a soit continuation dans le référentiel en cours soit fermeture du référentiel précédent.

3.3.3 Exemple de construction d'un graphe à partir d'un texte court

Nous détaillons, en les illustrant, les étapes de construction d'un graphe en partant de l'extrait de texte D associé aux connaissances visualisées dans la figure 10.

<p>Toute énonciation ouvre un référentiel énonciatif (dit global), quand bien même celui-ci n'est pas explicitement marqué dans le texte. On construit ici RE_g, premier nœud du graphe.</p>	
<p>Puisqu'il n'y a pas d'indice de rupture, on situe sur ce référentiel les propositions P₁ et P₂.</p>	
<p>Bien que située sur RE_g, P₂ contient des indices de rupture, <i>indiquer</i> suivi de <i>que</i>, qui déclenchent l'ouverture d'un nouveau référentiel énonciatif (local) RE₁₁.</p>	
<p>C'est sur ce référentiel que s'inscrit la proposition P₃. L'absence de marqueurs d'homogénéité (adverbe ou construction syntaxique) entre P₃ et P₄ et la typographie permet de fermer le référentiel RE₁₁ et de revenir sur RE_g pour situer P₄.</p>	
<p>Comme P₂ ouvrait un nouveau référentiel, P₄ ouvre un second référentiel énonciatif (local) RE₁₂, grâce aux indices <i>soutenir</i> suivi de <i>que</i>, auquel appartient P₅</p>	

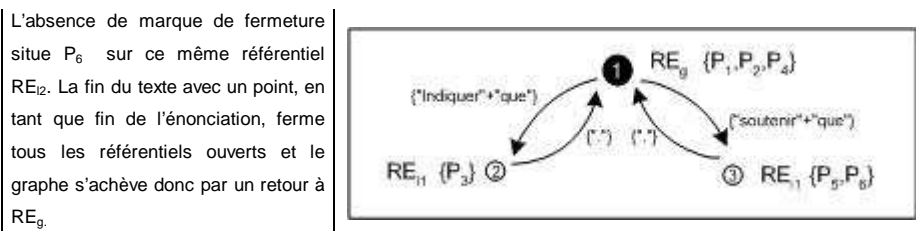


Figure 10. *Étapes de construction du graphe pour l'extrait du texte D.*

Comparable à la fois dans la stratégie de représentation des emboitements de segments textuels et dans la construction de type incrémentale adoptées, notre construction de la représentation d'un texte se distingue néanmoins de celle adoptée dans le cadre de la SDRT (Asher, 1993). Ce sont en effet les nœuds qui sont typés sémantiquement et non les arcs qui ne sont porteurs d'aucune interprétation sémantique ; dans la SDRT, les « boîtes » ou SDRS ne sont pas typées et ce sont les relations entre ces dernières qui le sont (à l'aide des relations dites de discours). Par ailleurs, nous mettons pour notre part l'accent sur les mécanismes d'ouverture et de fermeture des espaces de validation des propositions à la fois à un niveau conceptuel mais aussi opérationnel selon nous. Ce sont en effet des marqueurs (ou des ensembles de marqueurs) explicites qui déclenchent l'ouverture ou la fermeture de ces espaces et non l'interprétation de la sémantique des propositions comme proposé dans le cadre de la SDRT. Enfin, la SDRT ne propose pas d'analyser les phénomènes de prise en charge énonciative et modale à proprement parler, si ce n'est dans le cadre de travaux portant sur des situations d'échanges dialogiques (Asher *et* Lascarides, 2003 ; Maudet *et al.*, 2004), ce qui, reviendrait, notons-le, dans le cadre de notre approche, à considérer deux énonciateurs principaux. Ne nous situant pas dans cette perspective, nous pensons donc ne pas pouvoir aller plus avant – du moins à ce stade – dans un comparatif avec la méthodologie de la SDRT.

3.3.4 Exemple de construction d'un graphe à partir d'un texte long

Tout au long de cet article, nous n'avons proposé que de courts extraits de textes : la figure 11 illustre la représentation – obtenue manuellement – d'un texte complet¹³ comptant 46 propositions et 17 segments identifiés automatiquement.

¹³ Le texte D figure en annexe.

4. Automatiser le calcul de la représentation: implémentation et perspectives

4.1 Une architecture modulaire

Pour passer automatiquement du texte à sa représentation graphique, nous avons proposé une architecture modulaire qui présente les avantages (i) de pouvoir évaluer séparément chaque module ; (ii) dans une perspective à plus terme de modifier un module ou d'intégrer un module externe. Cette dernière section présente les différentes étapes de cette architecture résumée par la figure 12.

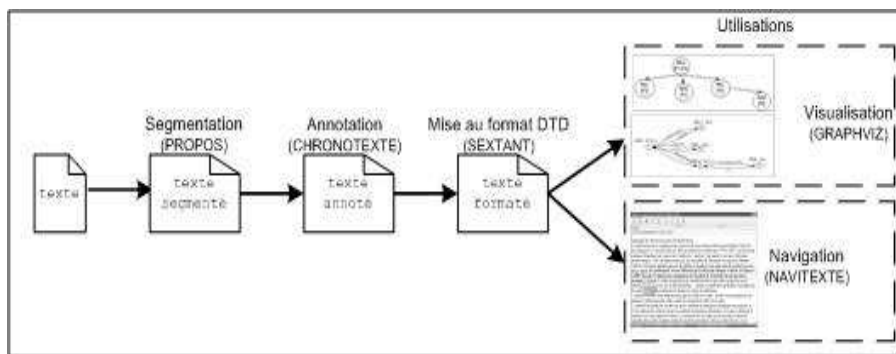


Figure 12. Chaîne de traitement pour l'automatisation de la génération de graphes

Quatre étapes sont nécessaires:

1. la première étape consiste à découper le texte (à l'aide du système PROPOS réalisé par (Wonsever, 2004)) pour définir les différentes unités propositionnelles qui vont constituer les différents segments textuels ;
2. la seconde étape annote les propositions du texte (à l'aide du système CHRONOTEXTE réalisé par (Chagnoux, 2006)) en fonction de marques d'ouverture ou de fermeture de référentiels ;
3. la troisième étape (non réalisée à l'heure actuelle) exploite les annotations pour reconstruire la dynamique discursive ;
4. la dernière étape (non réalisée à l'heure actuelle) exploite les résultats du fichier soit dans une perspective de visualisation, soit dans une perspective de navigation.

4.1.1 Découpage en propositions

L'organisation discursive est directement concernée par le problème de la segmentation. (Péry-Woodley, 2005 : 178) note à ce sujet que « toute structuration passe en effet par une segmentation, segmenter impliquant à la fois diviser et regrouper en fonction d'un critère organisationnel ». Cela passe par la recherche de « critères de regroupement d'unités (en segments), l'identification de ruptures ou de discontinuités (entre segments), l'étude de relations (entre segments) qui les hiérarchisent et forment des segments de niveaux supérieurs » (*ibid.*).

Bien que cette première étape soit primordiale pour l'analyse sémantique d'un texte, il existe actuellement peu d'outils pour segmenter automatiquement un texte en propositions et, à l'instar de (Mourad, 2000) ou (Nakamura-Delloye, 2007), les travaux de sémantique exigent parfois de construire un segmenteur en guise de prélude. Au lieu de construire notre propre outil, nous avons préféré utiliser *PROPOS* (Wonsever, 2004), un segmenteur en propositions, dont l'intérêt est de s'appuyer sur la présence de marqueurs linguistiques pour délimiter les propositions. La figure 13 présente le résultat du traitement par *PROPOS* d'un extrait du texte D : la phrase donnée en 1. est traitée comme décrit en 2. ; en 3., nous offrons une vue résumée de ce traitement pour une meilleure lisibilité.

- | |
|---|
| <ol style="list-style-type: none"> 1. Ses cadres affirment qu'ils ont les moyens de procéder à une escalade de grande envergure si Tsalal poursuit ses raids sur le sol libanais. 2. <PHRASE_TEXTE><PROP COD=pli1 NIVEAU=1 INDEP VERBE = affirment>Ses cadres affirment </PROP><PROP COD=pr1 NIVEAU=2 INSER IP=mot (que) VERBE = ont>qu'ils ont les moyens de procéder à une escalade de grande envergure ><PROP COD=gf21 NIVEAU=3 INDEP VERBE = poursuit>si Tsalal poursuit ses raids sur le sol libanais. </PROP></PROP></PHRASE_TEXTE> 3. [P1 : Ses cadres affirment] [P2 : qu'ils ont les moyens de procéder à une escalade de grande envergure [P3 : si Tsalal poursuit ses raids sur le sol libanais.]] |
|---|

Figure 13. Découpage en propositions par *PROPOS*.

Cet exemple témoigne du fait que *PROPOS* est apte à gérer les phénomènes d'inclusion syntaxique¹⁴ : P3 est bien incluse dans P2. La complexité de l'enchevêtrement des propositions fait que la qualité d'un outil de segmentation dépend fortement de sa capacité à identifier la hiérarchie propositionnelle : à ce titre, *PROPOS* fournit parfaitement les éléments nécessaires à la suite du traitement.

4.1.2 Identification des référentiels énonciatifs

Une fois le texte découpé en propositions, il est annoté par *CHRONOTEXTE* (Chagnoux, 2006). Conçu au départ pour attribuer des valeurs aspecto-temporelles aux propositions d'un texte, *CHRONOTEXTE* est un outil qui identifie les différents référentiels grâce à une base¹⁵ de 76 règles heuristiques et de 708 marqueurs organisée en 119 classes implémentées dans *Semantext* (Ben Hazez,

¹⁴ Nous insistons ici sur la différence entre les inclusion de type syntaxique (par exemple la présence d'une proposition relative) et les inclusions de type sémantique (par exemple la présence d'une citation).

¹⁵ (Chagnoux 2006) décrit en détail la modélisation des connaissances linguistiques organisées dans ces classes de marqueurs.

2002), une plate-forme de collecte et d'organisation de connaissances linguistiques pour l'annotation des textes. CHRONOTEXTE fonctionne en appliquant en cascade différents modules : les connaissances générées par un module sont capitalisées et utilisées pour le module suivant. Un moteur de règles coordonne l'exécution des règles quand une occurrence d'indicateur est reconnue : le système détermine l'ensemble des règles candidates à l'exécution puis les active une à une selon un ordre prioritaire. Dès lors, l'organisation des règles permet de définir l'ordre hiérarchique dans lequel elles s'appliqueront et d'éviter le déclenchement et l'application de plusieurs règles : les règles sont examinées dans l'ordre d'écriture dans la table des règles et correspondent à l'algorithme de construction de la structure. En sortie du système, toutes les propositions d'un texte sont balisées suivant le référentiel auquel elles appartiennent. A titre d'illustration, l'exemple 1 donné plus haut suit les étapes de traitement suivantes :

- le système repère tout d'abord les référentiels modaux. P3 vérifie la règle suivante (`$proposition$ starts-with /&si/ AND ($proposition$ contains /&present/ AND (($proposition justbefore $proposition$ contains /&present/ OR ($proposition justbefore $proposition$ contains /&present/))`)¹⁶ qui peut être glosée comme « si une proposition est introduite par un marqueur qui relève de la classe SI et si la proposition contient un verbe conjugué au présent et si la proposition précédente ou suivante contient un présent ». Par conséquent, le système attribue P3 à un référentiel possible éventuel (noté RPE) ;
- le système repère dans un second temps les référentiels énonciatifs locaux. Ainsi P2 vérifie la règle suivante (`$proposition justbefore $proposition$ contains /&ouverture-citation/ AND ($proposition$ startswith / &que/)` qui peut être glosée comme « si la proposition précédente contient un marqueur appartenant à la classe *&ouverture-citation*¹⁷ et si la proposition courante est introduite par un marqueur de la classe *&que* ». Par conséquent, P2 est identifiée comme appartenant à un référentiel énonciatif local (noté REL) ;
- puisque nous partons du postulat que tout texte s'inscrit dans un référentiel énonciatif global par défaut, tant qu'aucun marqueur ne permet de les assigner à un autre référentiel, toutes les propositions relèvent de ce référentiel énonciatif global. Le système vérifie donc la règle (`NOT ($proposition$ contains (/ $Refentiel-possible/ OR / $Referentiel-mental/ OR /Referentiel-enonciatif2/ OR /&Referentiel-Nactualise/)`) : une fois que tous les autres types de propositions ont été reconnues, le système

¹⁶ Les règles de CHRONOTEXTE sont implémentées dans le langage propriétaire de SEMANTEXT, c'est-à-dire dans la syntaxe proposée par le système par opposition à des langages de programmation.

¹⁷ Dans CHRONOTEXTE, *&ouverture_citation* est le nom de la classe contenant plus de 90 verbes susceptibles d'introduire les propos d'un tiers comme *s'exclamer, dire, hurler, etc.*

considère que les propositions restantes relèvent du référentiel énonciatif global (noté REG)¹⁸.

Par conséquent, en sortie, l'extrait est balisé comme suit :

```
<REG> Ses cadres affirment </REG> <REL> qu'ils ont les moyens de procéder à  
une escalade de grande envergure </REL> <RPE> si Tsahal poursuit ses raids sur  
le sol libanais.</RPE>
```

A ce stade, les propositions sont balisées une à une et des règles récursives ont défini des conditions de propagation qui permettent à l'annotation sémantique d'une proposition d'être associée à la proposition suivante afin d'annoter l'ensemble des propositions du texte.

4.1.3 Construction de la dynamique discursive

La troisième étape utilise la hiérarchie propositionnelle et les annotations sémantiques pour reconstruire la dynamique discursive en regroupant les différentes propositions au sein de segments cohérents.

A l'aide de la DTD définie par (Couto, 2006) pour la plateforme Navitexte (Couto *et* Minel, 2006), il est possible à la fois de prendre en compte les balises attribuées à chaque proposition pour rapprocher les propositions relevant d'un même référentiel. La figure 14 présente le balisage associé à l'extrait de texte analysé précédemment).

Le « *Corps* » contient les différentes propositions du texte, hiérarchisées en référentiels. Chaque segment est défini grâce à une *Unité Textuelle* (UT) à laquelle est associé un attribut « *Nature* » qui définit le type de référentiel (ici, énonciatif global, énonciatif local ou possible éventuel). Il faut noter que chaque type de référentiel convoque des attributs différents permettant d'encoder toutes les propriétés qui lui sont associées. Ces données ne sont pas exploitées dans ce court extrait mais par exemple dans le cas d'un énonciatif local, la DTD prévoit un attribut permettant à terme d'associer un énonciateur particulier à chaque segment.

La « *Tete* » permet de traiter le problème de la discontinuité : l'opération de « *Séquence* » rétablit la cohésion entre des segments discontinus dans le texte. Le court extrait donné ici ne présente pas de segments discontinus, la « *Tete* » est donc vide.

```
<?xml version="1.0" encoding="iso-8859-1"?>  
<!DOCTYPE Texte SYSTEM "DocumentNaviTexte.dtd">  
<Texte>  
  <Tete>  
  </Tete>  
</Corps>
```

¹⁸ Il s'agit là d'un exemple de règle de propagation qui permet d'annoter des propositions qui n'ont pas encore été traitées par le système.

```

<UT Type="Référentiel" Nro="1">
  <Attribut Nom="Nature">Enonciatif_global</Attribut>
  <UT Type=" Proposition" Nro="1">
    <Chaine>Ses cadres affirment qu'</Chaine></UT>
</UT Type=" Référentiel " Nro="2">
  <Attribut Nom="Nature">Enonciatif_local</Attribut>
  <UT Type=" Proposition " Nro="2">
    <Chaine>ils ont les moyens de procéder à une escalade de grande
    envergure</Chaine></UT>
</UT Type=" Référentiel " Nro="3">
  <Attribut Nom="Nature">Enonciatif_possible_eventuel</Attribut>
  <UT Type=" Proposition " Nro="3">
    <Chaine>si Tsahal poursuit ses raids sur le sol
    libanais.</Chaine></UT></UT></UT></UT>
</Corps>
</Texte>

```

Figure 14. Fichier XML associé à l'extrait de texte D

4.1.4 Applications visées

Actuellement, le fichier généré peut être exploité de deux manières en fonction des deux principales applications que nous visons, la visualisation des graphes en tant qu'objets graphiques ou leur exploitation en tant que structures de données.

Premièrement, la génération automatique de graphes à l'aide d'un outil de visualisation comme GRAPHVIZ¹⁹ fait émerger des phénomènes invisibles à la lecture : la visualisation de la structure énonciative et modale d'un texte facilite ainsi la détection d'éventuelles régularités de distribution de phénomènes, comme par exemple une tendance des phénomènes citationnels (eux-mêmes possiblement imbriqués les uns dans les autres) à apparaître de manière plus fréquente pour des thèmes sujets à débats (Battistelli *et* Chagnoux, 2006). Cette visualisation rend ainsi possible la présentation explicite d'un document selon différents *niveaux de profondeur* dans la prise en charge énonciative et modale par l'énonciateur principal des différentes propositions : à un premier niveau (celui du nœud 1), toutes les propositions sont considérées comme totalement prises en charge par l'énonciateur principal ; à un second niveau (celui qui correspond à tous les nœuds immédiatement adjacents au nœud 1), les propositions sont soit prises en charge par des énonciateurs seconds soit se situent sur des référentiels modaux qui correspondent à des faits décrits par l'énonciateur principal comme non attestés ; et ainsi de suite jusqu'à atteindre la profondeur maximale de la structure d'arbre sous-jacente au texte ainsi mise en évidence. Il devient alors possible de visualiser une certaine forme de *mise à distance* par l'énonciateur principal des faits et

¹⁹ GRAPHVIZ (<http://www.graphviz.org/>) permet de générer automatiquement un graphe à partir d'un fichier XML. Nous travaillons actuellement à l'étape de transformation qui permet de passer automatique du fichier XML à la représentation graphique.

événements décrits dans un texte. Nous donnons dans la figure 15 un exemple construit à la main d'une telle visualisation (il correspond à l'analyse du texte donné en figure 8). Nous avons par ailleurs déjà fait remarquer que le graphe de la figure 11 comprend une structure semi-profonde ancrée au nœud 14. La représentation visuelle sous la forme d'un graphe (comme donnée à voir par exemple dans la figure 11) permet quant à elle de visualiser le cheminement (puisque les nœuds sont numérotés) entre différents référentiels opérés par le lecteur et donc une certaine dynamique interprétative de lecture. De manière générale, nous avons noté que ce type d'outil de visualisation de phénomènes discursifs semble être « réclamé » par une partie des chercheurs en linguistique textuelle (Péry-Woodley, 2005 : 188).

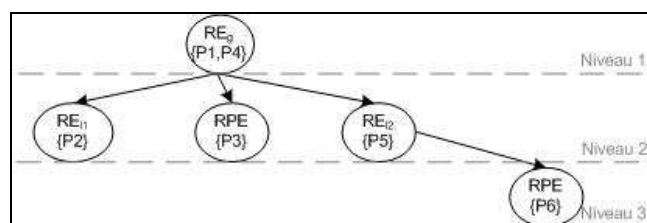


Figure 15: Visualiser les niveaux de prises en charge énonciative et modale dans un texte

Deuxièmement, le format d'encodage des informations dans la DTD adoptée permet un certain nombre d'opérations de navigation à l'intérieur d'un texte dans la plate-forme Navitexte de (Couto *et* Minel, 2006). Il est ainsi possible de naviguer à l'intérieur des différentes propositions d'un même référentiel quand bien même elles ne sont pas adjacentes ; il est aussi possible de naviguer entre des segments textuels à des niveaux de profondeur variable de la structure d'arbre sous-jacente au texte. Les principes de navigation s'appuient sur l'identification des valeurs sémantiques attribuées à chacun des segments du texte.

4.2 Etat de développement de la chaîne de traitement, difficultés rencontrées et perspectives d'évaluation

Le principe de la chaîne de traitement décrite par la figure 12 est actuellement évalué sur un corpus de 25 textes (5 textes journalistiques, 5 textes tirés de Wikipedia, 5 incipit de romans, 5 textes scientifiques et 5 textes fournis par une entreprise). Les tâches de segmentation et d'annotation sont opérationnelles et nous travaillons actuellement d'une part à la mise au format automatique dans la DTD choisie et d'autre part à la visualisation de l'arbre et du graphe *via* Graphviz.

Concernant les deux premières tâches de segmentation et d'annotation, il nous faut encore résoudre un certain nombre de difficultés : lors de l'annotation automatique des propositions, il est parfois difficile de distinguer automatiquement la nature de tous les référentiels. Ainsi par exemple, le repérage des assertions introduites par « selon » ou « pour » en position antéposée produit encore trop de

bruits en identifiant des propositions comme « selon l'appartenance sociale » ou « pour le développement ». Les règles qui traitent les phénomènes modaux doivent encore être affinées pour permettre une meilleure annotation. Il faudrait également à terme pouvoir regrouper les segments discontinus pris en charge par un même énonciateur que celui-ci soit par exemple « Le porte-parole du gouvernement américain » ou « La Maison Blanche ». Le graphe que nous proposons présente en effet au moins autant de nœuds que de prises en charges énonciatives repérées. Or tous ces nœuds ne correspondent pas nécessairement à autant d'énonciateurs. Ainsi, résoudre les relations d'anaphore pronominale et nominale pourrait conduire à créer une seconde vue du texte où tous les propos pris en charge par un énonciateur seraient regroupés sous un nœud, en donnant ensuite la possibilité là aussi de visualiser l'ensemble de ces propos en se positionnant sur ce dernier nœud.

En ce qui concerne l'évaluation, comme l'explicite (Chaudiron, 2004), la difficulté qu'il y a à construire des protocoles d'évaluation est particulièrement accrue dans le domaine du traitement ou de la modélisation de phénomènes d'ordre sémantique. Nous n'avons pour le moment pas mis en place de protocole explicite d'évaluation par des lecteurs humains de notre méthodologie et du type de représentations proposées (arbres ou graphes). Cependant, dans le cadre de deux cours d'ingénierie linguistique²⁰, nous avons pu constater au travers d'exercices visant à vérifier la compréhension de l'importance de la prise en compte des phénomènes citationnels et de modalisation pour des systèmes de Recherche d'Information (tels que les systèmes de Questions/Réponses ou les systèmes de navigation textuelle en particulier), que la méthodologie que nous proposons était assimilée et reproductible par les étudiants sur différents corpus constitués par eux-mêmes.

Nous travaillons par ailleurs à la prise en compte de notre méthodologie de repérage des phénomènes citationnels pour le calcul des expressions temporelles adverbiales déictiques dans le cadre du développement d'un outil d'aide à la lecture de textes (Battistelli *et al.*, 2006b). Dans la perspective soulevée par (Jacques *et Aussenac-Gilles*, 2007), nous souhaitons également soumettre nos patrons lexico-syntaxiques ainsi que nos règles à l'épreuve de plusieurs corpus afin de mesurer la variabilité de la performance de notre système. Enfin, le format d'encodage des informations qui permet un certain nombre d'opérations de navigation à l'intérieur d'un texte dans une plate-forme comme celle de (Couto *et Minel*, 2006) offre une méthode alternative de validation de la représentation que nous avons proposée. Une opération de navigation envisageable serait par exemple celle qui consisterait à proposer de « passer du référentiel énonciatif global aux référentiels énonciatifs locaux ». Les premiers résultats que nous avons obtenus montrent par ailleurs que notre méthodologie permet de proposer une typologie de textes basée sur les mécanismes énonciatifs (Battistelli *et Chagnoux*, 2006).

²⁰ Il s'agit de cours dispensés à l'Université Paris 4 dans le cadre de deux Master 2 professionnels : *Médias Informatisés et Stratégies de Communication* (Celsa) et *Ingénierie de la Langue pour la Gestion Intelligente de l'Information* (ISHA).

5. Conclusion

Dans cet article, nous avons proposé une méthodologie et un outil permettant de construire une visualisation dynamique et hiérarchisée et nous l'avons mis en œuvre pour illustrer la dynamique discursive qui fait varier les prises en charges énonciatives et modales au sein d'un texte. Les objectifs de cet outil sont de diverses natures : (i) du point de vue théorique, il propose un modèle d'organisation du discours basé sur les notions de cohésion et de rupture ; (ii) du point de vue linguistique, il offre une modélisation des marqueurs assurant ou non la cohésion textuelle ; (iii) du point de vue informatique, il permet de construire automatiquement une visualisation à partir d'un texte annoté selon une DTD donnée ; (iv) du point de vue des applications de TAL, il améliore l'extraction de segments textuels en associant à chaque segment des propriétés qui assurent la continuité référentielle. Relativement à ce dernier point, nous pensons qu'il illustre l'apport – et la nécessaire prise en compte – des travaux en linguistique textuelle dans le domaine du TAL qui semble ainsi « revenir » à des méthodes de représentation de la sémantique de textes entiers après s'être longtemps cantonnée à des analyses locales²¹. Nous situant dans une démarche d'analyse empirique des modes de structuration discursive – ici, énonciative et modale –, notre approche se distingue ainsi de la « simple » approche d'annotation (locale) ; nous nous inscrivons dans les principes méthodologiques de la linguistique textuelle et donc – par là-même – dans ceux qui fondent le développement de systèmes de navigation textuelle qui s'appuient directement sur les structures discursives mises en évidence.

Par ailleurs, quand à l'instar de (Paveau, 2007), de nombreux travaux arguent que l'analyse du discours ne peut se faire sans intégrer les questions du contexte et de la culture et en appelle à la pragmatique ou à la cognition sociale pour construire des méthodes d'analyse du discours, il nous semble important de prouver qu'il est possible de proposer des analyses sémantiques fines et automatisables en faisant du discours son propre contexte et d'essayer de fixer les limites de la notion de texte autoporteur.

6. Bibliographie

- ASHER N. (1993). *Reference to Abstract objects in Discourse*. Kluwer Academic Publishers.
- ASHER N., LASCARIDES A. (2003). *Logics of Conversation*, Cambridge University Press.
- BATTISTELLI D., CHAGNOUX M. (2006). Vers une typologie de mécanismes discursifs. *Actes de la journée ATALA "Typologie de textes"*, Paris.

²¹ Voir par exemple l'analyse de ce phénomène présentée dans (Battistelli *et* Minel, 2006) ou dans (Nazarenko, 2005).

- BATTISTELLI D., CHAGNOUX M., DESCLES J.-P. (2006a). Référentiels et ordonnancements temporels dans les textes. *Cahiers Chronos*, 18, Amsterdam/New York.
- BATTISTELLI D., MINEL J.-L. (2006). Les systèmes de résumé automatique : comment assurer une continuité référentielle dans la lecture des textes. G. Sabah (Ed.), *Compréhension des langues et interaction*, 295-330.
- BATTISTELLI D., MINEL J.-L., SCHWER S. (2006b). Représentation des expressions calendaires dans les textes : vers une application à la lecture assistée de biographies. *TAL* vol. 47:2.
- BEN HAZEZ S. (2002). Un modèle d'exploration contextuelle des textes : filtrage et structuration d'informations textuelles, modélisation et réalisation informatique (système SemanText). *Thèse de l'Université Paris-Sorbonne*.
- BILHAUT F., HO-DAC M., BORILLO A., CHARNOIS T., ENJALBERT P., LE DRAOULEC A., MATHET Y., MIGUET H., PERY-WOODLEY M.-P., SARDA L. (2003). Indexation discursive pour la navigation intradocumentaire : cadres temporels et spatiaux dans l'information géographique. *Actes TALN'03*, Batz-sur-Mer, 315-320.
- BUSQUETS J., VIEU L., ASHER N. (2001). La SDRT : une approche de la cohérence du discours dans la tradition de la sémantique dynamique. *Verbum*, tome XXIII, 1.
- BUSQUETS J., HARDT D. (dir.) (2005). Modèles et algorithme pour la résolution d'anaphores, *TAL* vol. 46:1.
- CHAGNOUX M. (2006). Temporalité et aspectualité dans les textes français. Modélisation sémantico-cognitive et traitement informatique. *Thèse de l'Université Paris-Sorbonne*.
- CHAROLLES M., VIGIER D. (2005). Les adverbiaux en position préverbale : portée cadrative et organisation des discours, *Langue Française*, 148, 9-30.
- CHAUDIRON, S. (2004). *L'évaluation des systèmes de traitement de l'information*. Lavoisier.
- COUTO J. (2006). Modélisation des connaissances pour une navigation textuelle assistée. La plate-forme logicielle NaviTexte. *Thèse de l'Université Paris-Sorbonne*.
- COUTO J., MINEL J.-L. (2006). SEXTANT, un langage de modélisation des connaissances pour la navigation textuelle. *Actes ISDD'06*, Caen, 80-90.
- CORNISH F. (2006). Relations de cohérence en discours : critères de reconnaissance, caractérisation et articulation cohésion-cohérence, Journée d'étude du CRISCO *Organisation des textes et cohérence des discours*, Université de Caen, texte publié en ligne en 2006
- CRESTAN É., DE LOUPY C., MANIGOT L. (2004). Analyses sémantiques pour la navigation textuelles ; *Actes CIDE 7* La Rochelle, 22-25 juin 2004, 293-308.
- DAILLE B., MORIN E. (2000). Reconnaissance automatique des noms propres de la langue écrite : les récentes réalisations. *TAL* vol. 41:3, 601-622.
- DESCLES J.-P. (1980). Construction formelle de la catégorie grammaticale du temps et de l'aspect, *Notion d'aspect*, 198-237, J. David et R. Martin (eds), Klincksieck, Paris.
- DESCLES J.-P. (1995). Les référentiels temporels pour le temps linguistique. *Modèles linguistiques* 16, 9-36.

- DESCLES J.-P., Guentcheva Z. (2000). Énonciateur, locuteur, médiateur. Erikson Ph et Monod-Becquelin A (éds). *Les rituels du dialogue*. Nanterre : Société d'ethnologie.
- GOSSELIN L. (2005). *Temporalité et modalité*. De Boeck et Larcier, Editions Duculot, Bruxelles.
- GUENTCHEVA Z. (éd.) (2000). *L'énonciation médiatisée*. Louvain-Paris : Peeters.
- GIGUET E., LUCAS N. (2004). La détection automatique des citations et des locuteurs dans les textes informatifs. *Le Discours rapporté dans tous ses états : question de frontières*, Paris, L'Harmattan.
- JACKIEWICZ A. (2006). Relations intersubjectives dans les discours rapportés, *TAL* vol. 47:2, 65-87.
- JACKIEWICZ A., LAUBLET P. (2007). Web sémantique et linguistique du discours. OntoDiscours : un tournant énonciatif, *Actes TIA'07*, Journée Ontologies et textes, INRIA Sophia Antipolis.
- JACQUES M.P., AUSSENAC-GILLES N. (2007). Variabilité des performances des outils de TAL et genre textuel, Cas des patrons lexico-syntaxiques. *TAL* vol. 48.1.
- HOBBS J.R. (1990). The coherence and structure of discourse, *Literature and Cognition*, Leland Stanford Junior University, Calif. *CSLI Lecture Notes* 21, 83-114.
- KRONNING H. (2003). Modalité et évidentialité. Birkelund, M., Boysen, G. & Kjærsgaard, P. S. (éds). *Aspects de la Modalité, Linguistische Arbeiten* 469, 131-151.
- LE DRAOULEC A., PERY-WOODLEY M.-P. (2005). Encadrement temporel et relations de discours. *Langue Française* 148, 45-60.
- MANN W.C., THOMPSON S.A. (1988). Rhetorical Structure Theory : toward a functional theory of text organization. *Text* 8(3), 243-281.
- MAUDET N., MULLER P., PREVOT L. (2004). Tableaux conversationnels en SDRT. *Actes TALN'04*, Workshop SDRT, Fès, 22/04/2004.
- MOURAD G., MINEL J.-L. (2000) « Filtrage sémantique du texte, le cas de la citation », *Actes 3ème Colloque International sur le Document Electronique (CIDE'2000)*, Lyon, 41-56.
- NAKAMURA-DELLOYE Y. (2007). Alignement automatique de textes parallèles français-japonais, *Thèse de l'Université Paris 7*.
- NAZARENKO A. (2005). Sur quelle sémantique reposent les méthodes automatiques d'accès au contenu textuel ? *Sémantique et Corpus*, chapitre 6, Lavoisier.
- PAVEAU M.-A. (2007). Discours et cognition : les prédiscours entre cadres internes et environnement extérieur, *Corela*, Numéro spécial, Cognition, discours, contextes.
- PERY-WOODLEY M.-P. (2001). Présentation du numéro : Cohérence et relations de discours à l'écrit, *Verbum*, XXIII, 1.
- PERY-WOODLEY M.-P. (2005). Discours, Corpus, Traitements automatiques. *Sémantique et Corpus*, chapitre 5, Lavoisier.

- ROSTAING H., BOUTIN E. MANNINA B. (1999) Evaluation of Internet resources : Bibliometric technique applications. *Actes Cybermetrics'99, ISSI'99* , University of Colima, Mexico, 9 juillet 1999.
- SANDOR A., KAPLAN A., RONDEAU G. (2006). Discourse and citation analysis with concept matching. *Schedae*, prépublication n°19 (fascicule 1), 147-152.
- SAURI R., VERHAGEN M., PUSTEJOVSKY J. (2005). Annotating and recognizing Event Modality in Text. *Actes FLAIRS'06*, Melbourne Beach, Florida.
- SMALL H. (1973). Co-citation in the scientific literature. *Journal of the American Society for Information Science*, 24, 265-269.
- TEUFEL S., SIDDHARTHAN A., Tidhar D. (2006). Automatic classification of citation function. *Actes EMNLP'06*, Sydney , Australie.
- WILSON T., WIEBE J. (2005). Annotating Attributions and Private States, *Actes ACL Workshop on Frontiers in Corpus Annotation II: Pie in the Sky*, Ann Arbor, Michigan.
- WOLF F., GIBSON E. (2005). Representing Discourse Coherence : A Corpus-Based Study, *Computational Linguistics*, vol. 31, 249-288.
- WONSEVER D. (2004). Repérage automatique des propositions par exploration contextuelle. *Thèse de l'Université Paris-Sorbonne*.

7. Annexe

7.1 Texte D associé à la figure 11

Hier après-midi, quarante-huit heures après l'ouverture du front libano-israélien, le Hezbollah qui, selon son chef Hassan Nasrallah, disposerait de plus de 10 000 roquettes et missiles pointés sur Israël, avait déjà frappé 165 fois le territoire de l'Etat hébreu. Et ce ne pourrait être qu'un début. Ses cadres affirment qu'ils ont les moyens de procéder à une escalade de grande envergure si Tsahal poursuit ses raids sur le sol libanais. «Haïfa, que nous n'avons pas encore pris pour cible, contrairement à ce que disent les Israéliens, pourrait figurer très prochainement parmi nos objectifs, déclare le député du parti chiite Hussein Haj Hassan. Mais, nous sommes également prêts à étendre nos opérations à une échelle beaucoup plus vaste si les attaques contre nos civils perdurent.» La milice chiite a-t-elle pour autant les moyens d'atteindre Tel-Aviv comme l'a écrit le *Jane's Defence Weekly*, hebdomadaire britannique spécialisé dans les questions de défense ? Les responsables islamistes refusent d'infirmier ou de confirmer. Mais un observateur libanais qui connaît bien le Parti de Dieu soutient qu'il dispose de l'armement nécessaire à des tirs en profondeur. «Le Hezbollah peut aussi se tourner vers un autre plan et décider de frapper à l'extérieur de l'Etat hébreu», ajoute-t-il.

Pour l'heure, les miliciens arrosent le nord d'Israël de leurs roquettes. Disséminés dans le sud du Liban, ils sont lourdement armés. Leur arsenal compterait, entre autres, missiles antichars, mortiers, lance-roquettes B7, une artillerie antiaérienne ou encore des unités d'artillerie. Un équipement fourni par Téhéran et qui transite par Damas. Depuis le début

de son opération, Tsahal a veillé à empêcher tout transfert d'armes vers la frontière libanaise en pilonnant les principaux points de passages vers le sud.

Mais l'autre force du Hezbollah, qui a vu le jour après l'invasion israélienne de 1982, réside dans sa capacité de mobilisation de la population. L'immense majorité des Libanais chiïtes sont acquis à la cause de la résistance islamique et vouent un véritable culte à Nasrallah, élu par ses pairs en 1992. Vendredi, le leader charismatique a échappé de peu au bombardement de son QG dans la banlieue de Beyrouth et promis à la télévision «la guerre ouverte».

Alors que le Parti de Dieu a intégré pour la première fois le gouvernement en juillet, il consacrait jusqu'ici l'essentiel de son énergie à la lutte contre l'occupation israélienne. Pour ce faire, il dispose de dizaines de milliers d'hommes en âge de se battre. Parmi eux, 5 000 miliciens, dont quelques centaines de commandos seraient particulièrement aguerris. Ses objectifs : libérer les fermes de Chebaa, un territoire de 25 km² au sud du pays, défendre le pays contre toute agression de Tsahal et obtenir le retour de cinq Libanais détenus en Israël. Des objectifs exclusivement nationaux, donc, tout comme l'est son discours. Malgré tout, ses affinités affichées avec l'Iran, la Syrie ou le Hamas, sèment le doute.

«Lors des récentes réunions de la conférence de dialogue national [au Liban, ndlr], Nasrallah a promis aux leaders politiques qu'il veillerait à ce que l'été soit calme, rappelle l'éditorialiste du quotidien *An-Nahar*, Rosana Boumonsef. Or, nous assistons actuellement à une offensive israélienne sans précédent depuis 1982. Alors de deux choses l'une. Soit le Hezbollah a mal calculé les répercussions qu'aurait la capture de deux soldats israéliens. Soit, il a été poussé par l'Iran à venir en aide au Hamas, actuellement considérablement affaibli à Gaza.»

7.2 Références bibliographiques des textes du corpus cités dans cet article

- **Texte A** : "Le Liban abandonné", site de *France 5*, <http://www.france5.fr/cdanslair/006721/434/136151.cfm>. (consulté le 25/07/06).
- **Texte B** : "L'armée israélienne pénètre sur le territoire libanais, les avions survolent Beyrouth.", site de l'association *France Palestine 5*, <http://www.france-palestine.org/article4161.html>. (consulté le 25/07/06).
- **Texte C** : "L'offensive du Hezbollah – Chronologie des événements.", site de *L'ambassade d'Israël en France*, <http://paris1.mfa.gov.il/mfm/web/main/missionhome.asp?MissionID=31&>. (consulté le 15/09/06)
- **Texte D** : "L'armée israélienne investit le Liban sud.", Arsenault, C., site de *RFI*, http://www.rfi.fr/actufr/articles/079/article_45021.asp. (publié le 12/07/06 et consulté le 25/07/06)
- **Texte E** : "Conflit israélo-libanais de 2006.", site de l'encyclopédie *Wikipedia*, http://fr.wikipedia.org/wiki/Conflit_isra%C3%A9lo-

libanais de 2006. (publié le 13/07/06, actualisé le 14/07/06 et consulté le 25/07/06)

- **Texte F** : "Israël frappe lourdement le Sud-Liban", site du *Figaro*, édition en ligne du 15/10/07.
http://www.lefigaro.fr/international/20060715.FIG000000419_israel_pilone_le_hezbollah_et.html, (publié le 15/07/06, actualisé le 16/07/06 et consulté le 25/07/06)