TALN 2011, Montpellier, 27 juin – 1er juillet 2011

Calcul de réseaux phrastiques pour l'analyse et la navigation textuelle

Dominique Legallois[1] Peggy Cellier[2] Thierry Charnois[3]
(1) CRISCO Université de Caen Basse-Normandie, Campus 1, 14000 Caen
(2) IRISA-INSA de Rennes, Campus Beaulieu 35042 Rennes cedex
(3) GREYC Université de Caen Basse-Normandie, Campus 2, 14000 Caen

Résumé. Le travail présente une méthode de navigation dans les textes, fondée sur la répétition lexicale. La méthode choisie est celle développée par le linguiste Hoey. Son application manuelle à des textes de grandeur conséquente est problématique. Nous proposons dans cet article un processus automatique qui permet d'analyser selon cette méthode des textes de grande taille ; des expériences ont été menées appliquant le processus à différents types de textes (narratif, expositif) et montrant l'intérêt de l'approche.

Abstract. In this paper, we present an automatic process based on lexical repetition introduced by Hoey. The application of that kind of approaches on large texts is difficult to do by hand. In the paper, we propose an automatic process to treat large texts. We have conducted some experiments on different kinds of texts (narrative, expositive) to show the benefits of the approach.

Mots-clés : Réseau phrastique, Appariement de phrases, Analyse textuelle, Navigation textuelle.

Keywords: Sentence network, Bonds between sentences, Textual analysis, Textual navigation.

1 Introduction

Notre travail propose une analyse des textes, fondée à la fois sur la « réduction textuelle » et la répétition lexicale. Nous nous inspirons du modèle linguistique de Hoey (Hoey, 1991). Par « réduction », nous entendons une méthode linguistique qui vise à déterminer dans un texte, quelles sont les phrases les plus pertinentes informationnellement. L'ensemble de ces phrases délesté des phrases informationnellement marginales, constitue un ensemble théoriquement cohérent. Contrairement aux travaux sur les résumés automatiques, [1] nous définissons une méthode qui s'inscrit dans une démarche expérimentale visant, en même temps, à mieux comprendre l'organisation textuelle et à proposer un mode de navigation. Nous ne connaissons que peu d'analyses linguistiques proposant une telle démarche. Outre Hoey, mentionnons les travaux de Thomas (Thomas, 1999) dont l'analyse est fondée sur la proéminence sémantique des propositions, et de Toolan (Toolan, 2009) qui travaille exclusivement sur les textes narratifs afin de conserver les phrases constituants les nœuds importants de l'histoire. Le modèle de Hoey, que nous avons appliqué au français (Legallois, 2004, 2006), ne prend en compte que les textes expositifs (par ex. les textes scientifiques, philosophiques) et exclut en principe le genre narratif [2]. De plus, les appariements de phrases ayant des lexèmes en commun ne peuvent être repérés que sur des textes courts. En effet, l'application manuelle de la méthode sur des textes de plusieurs milliers de lignes est extrêmement problématique.

Dans cet article, nous proposons d'implémenter l'adaptation au français de la méthode de Hoey dans un processus automatisé afin de pouvoir analyser des corpus variés en genre et en taille. Cette mise en œuvre informatique permet notamment au linguiste d'observer et de tester l'application du modèle en corpus. L'implémentation concerne uniquement la répétition lexicale au sens strict (i.e., le même lemme), qui constitue, d'après nos observations, le cas de répétition le plus fréquent. La répétition par anaphore et la synonymie ne sont pas prises en compte [3]. Toutefois, cette limitation n'entrave en rien la pertinence du modèle pour la recherche en linguistique comme les expériences le montrent. Dans la suite de l'article nous présentons le modèle linguistique (Section 2). Puis nous

1. *Cf.* (Knight & Marcu, 2002) ou les récents systèmes présentés aux compétitions TAC 2008-2010 (`www.nist.gov/tac/tracks/`)
2. Nous verrons plus bas que nous avons fait une entorse à ce principe.
3. Dans l'état actuel des techniques, l'identification de la répétition par anaphore reste un problème non résolu sur un texte long.

discutons les expériences que nous avons menées sur différents types de textes : narratifs et expositifs (Section3).

2 Modèle linguistique

Dans cette section nous rappelons le modèle linguistique présenté dans (Legallois, 2006) en donnant quelques exemples. La méthode consiste à identifier les phrases d'un texte qui partagent au moins trois unités lexicales, aussi appelées *lexèmes* (par ex. : nom, verbe, adjectif, adverbe). Ce partage est fondé sur le concept de répétition lexicale : répétition à l'identique de mots (*chômage/chômage*) ou sous une forme « dérivée » (*travail/travailler*), mais aussi reprise anaphorique, synonymique (*meurtre/assassinat*), hypo/hyperonymie (*végétal/tulipe*), relation « implicative » (*conduire/voiture*), suite ordonnée (*lundi/mardi/...*). Le principe P s'applique alors :

P : si une phrase d'un texte non narratif partage au moins trois lexèmes avec une ou plusieurs autres phrases du même texte quel que soit l'empan entre ces phrases – alors la suite de ces phrases sera cohérente, c'est-à-dire interprétable dans le contexte développé par le texte.

Lorsque deux phrases partagent au moins trois lexèmes, on dit qu'il y a *appariement* entre ces phrases. Considérons un appariement entre deux phrases comme un chemin entre elles. On appelle *réseau phrastique* un ensemble d'au moins trois phrases tel que quelles que soient deux phrases de ce réseau, on peut trouver une succession de chemins menant de l'une à l'autre. On appelle *hypotexte* l'ensemble des réseaux phrastiques. Notons que les phrases *marginales*, c'est-à-dire ne participant à aucun appariement, n'apparaissent pas dans l'hypotexte. L'hypotexte constitue une réduction du texte. Nous donnons ci-dessous un exemple de réseau phrastique traité dans (Legallois, 2004) portant sur un article de plus d'une centaine de phrases [4]. La numérotation correspond à l'ordre des phrases dans le texte intégral ; on voit donc que l'empan de ce réseau est très conséquent :

[1] *« un faisceau de lumière envoyé par un projecteur sur la scène obscure d'un* ***théâtre***$_1$*, celui de l'****inconscient***$_2$ » : cette **métaphore**$_3$, couramment employée pour décrire la **conscience**$_4$, découle directement du **postulat**$_5$ **fondateur**$_6$ de la psychologie **cognitive**$_7$.
[30] Pour expliquer les contenus de la **conscience**$_4$, il faut donc comprendre la **pièce**$_1$ qui se joue au plus profond de notre **inconscient**$_2$.
[90] Le **modèle**$_3$ initial du "**théâtre**$_1$ de la **conscience**$_4$" **postule**$_5$, ~~nous l'avons dit~~, que cette dernière est dépourvue de toute fonction dans la dynamique de l'apprentissage implicite.
[99] [Mais] La **métaphore**$_3$ de la **conscience**$_4$, simple spectateur d'une **pièce**$_1$ dont l'**inconscient**$_4$ serait l'auteur et le metteur en **scène**$_1$, a-t-elle encore lieu d'être ?
[106] La remise en question d'une vie mentale **inconsciente**$_4$ s'oppose ~~en effet~~ non seulement aux **principes**$_5$ **fondateurs**$_6$ de la psychologie **cognitive**$_7$, mais, au-delà, à une conception qui est devenue quasi-universellement acceptée dans le grand public par la banalisation des conceptions d'inspiration psychanalytique.

Ce réseau est tout à fait interprétable avec quelques « aménagements » mineurs. En effet, dans l'exemple certains mots sont barrés (par ex. ~~en effet~~), et d'autres mis entre crochets (par ex. [Mais]). Il s'agit d'une restitution de certaines formes dont se dispensait pour des raisons normatives, la cohésion locale du texte initial. Cette restitution permet en outre de faciliter la lecture des enchaînements phrastiques du réseau. L'ensemble des réseaux pour ce texte forme donc un hypotexte constitué de 40% du nombre total des phrases.

3 Mise en œuvre informatique et résultats

3.1 Textes et outils

Textes Afin de tester la pertinence de la méthode et avoir un premier retour sur le modèle, nous avons choisi d'utiliser, pour nos expériences, des textes variés en genre (2 textes narratifs et 2 textes expositifs), en taille (de 108 phrases pour le plus petit à 6075 phrases pour le plus grand) et d'époques différentes (de la fin du 19e siècle au début du 21e). La table 1 détaille ces caractéristiques. Notons que pour les expériences, les phrases sont décrites par les lexèmes qui les composent. Les *grammèmes*, par exemple les prépositions et articles, ne sont pas conservés.

4. Article de Pierre Perruchet paru dans La Recherche, juillet-aout 2003

Titre	Auteur	Année	Genre	Nb de phrases
Le crime, phénomène normal	Émile Durkheim	1894	Epositif	108
Jeunes cherchent place	Nicole Baldé-Georgin	2000	Expositif	2713
J'irai cracher sur vos tombes	Boris Vian	1946	Narratif	3030
Mémoires de Guerre (Tome 2)	Général de Gaulle	1956	Narratif	6075

TABLE 1 – Caractéristiques des textes utilisés pour les expériences.

De plus, pour optimiser la découverte des appariements, ce ne sont pas les mots eux-mêmes mais leurs lemmes qui sont utilisés dans la description des phrases. Nous utilisons le catégoriseur Cordial pour cette tâche [5].

Visualisation des réseaux phrastiques Les réseaux phrastiques sont des graphes dont les nœuds sont les phrases du texte et les arcs relient les phrases appariées. Les arcs sont étiquetés par les lexèmes qui sont communs aux phrases. Les réseaux sont visualisés dans un outil d'affichage de graphes développé en java. Notons que le calcul des appariements se fait par intersection d'ensembles de lexèmes qui composent les phrases. Un exemple est donné à la figure 1. Sur cet exemple on voit une partie d'un réseau phrastique issu du texte *Jeunes cherchent place* de Nicole Baldé-Georgin. On y voit les quatre phrases contenant les mots *adulte*, *social* et *espace*. Notons que les quatres phrases contiennent également toutes les mots *jeune* et *être*. Une analyse textuelle plus détaillée est donnée à la section 3.2.

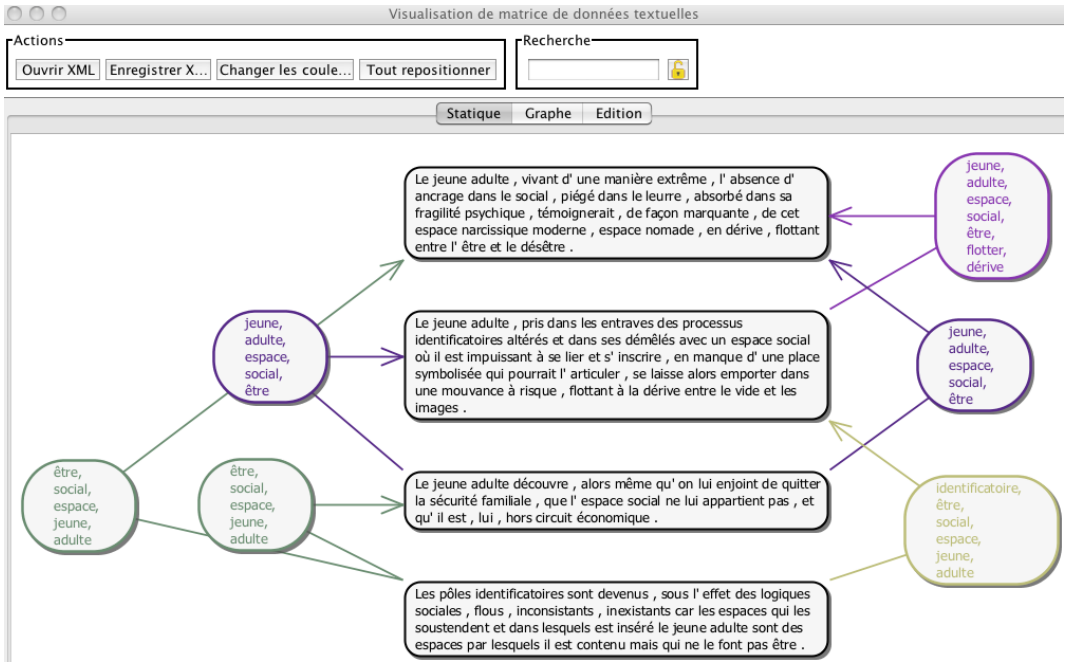

FIGURE 1 – Visualisation d'une partie d'un réseau phrastique extrait de l'essai de Nicole Baldé-Georgin.

Navigation dans les appariements Parfois les réseaux phrastiques sont très grands rendant leur affichage difficile à analyser. Afin d'explorer de larges textes en recherchant les phrases contenant des mots en commun nous avons utilisé Abilis (Allard *et al.*, 2010), un système d'information qui s'appuye sur la théorie de l'*analyse de concepts logique* (Ferré & Ridoux, 2004). Ce genre de système permet non seulement de naviguer dans les descriptions d'objets mais aussi d'élaborer des requêtes complexes. Dans notre application les objets sont les phrases. Chaque phrase est décrite par l'ensemble des *lexèmes* qui la constituent. La première fenêtre (en haut) du système d'information permet de faire une requête et de sélectionner ainsi toutes les phrases vérifiant cette requête. Dans l'exemple de la figure 2 montrant le tome 2 des mémoires du Général de Gaulle, la requête sélectionne toutes les phrases contenant le mot *juif* [6]. Le résultat de la requête (3 phrases) apparait dans la fenêtre droite. La fenêtre de

5. `http://www.synapse-fr.com/Cordial_Analyseur/Presentation_Cordial_Analyseur.htm`

6. Dans l'exemple, la requête est très simple. Des requêtes plus complexes combinant les opérateurs classiques de logique (*et, ou, non*) peuvent être faites, par exemple la requête `('juif' or 'persécution') and not 'horreur'` renvoie toutes les phrases conte-

gauche propose d'autres mots qui apparaissent dans les phrases dans lesquelles *juif* est employé (par exemple, les mots *cours* et *persécution*). On constate que les trois phrases du texte contenant *juif*, contiennent également toutes le mot *persécution* (indiqué par le chiffre entre parenthèses) et deux d'entre elles le mot *cours* utilisé pour indiquer une expression temporelle.

FIGURE 2 – Exemple de navigation dans les phrases des *Mémoires de guerre* du Général de Gaulle.

3.2 Analyse des résultats

Analyse quantitative À la table 2, on trouve un résumé des informations quantitatives observées sur les différents textes étudiés. Relativement au nombre de phrases, on constate un plus grand nombre d'appariements chez de Gaulle et chez Baldé-Georgin que chez Vian. Ceci semble s'expliquer par plusieurs raisons : Vian possède des phrases plus courtes (en moyenne 5 lexèmes par phrase), donc assez peu candidates aux appariements, et un vocabulaire plus diversifié. Au contraire, Baldé-Georgin et de Gaulle se caractérisent respectivement par une thématique homogène, et par la répétition des situations (par ex. la description des fronts pendant la guerre). Tous deux, en outre, sont composés de phrases relativement longues (en moyenne 10 à 11 lexèmes par phrase). De plus, quel que soit le genre, on note que très peu de réseaux phrastiques sont trouvés. Mais, ces réseaux sont de grande taille (relativement au nombre de phrases) pour Baldé-Georgin et de Gaulle, et les hypotextes correspondant couvrent plus de 65% de l'ensemble des phrases de ces œuvres. Les caractéristiques textuelles signalées précédemment (taille des phrases et homogénéité thématique) expliquent ces phénomènes qui sont accentués par la présence de lexèmes très fréquents comme *être, avoir, pouvoir*.

Titre	**Nb lexèmes/phrase (en moy.)**	**Nb appariements**	**Nb réseaux phrastiques**	**% de phrases dans l'hypotexte**
Le crime, phénomène normal	10	146	1	58%
Jeunes cherchent place	10	42 053	1	65%
J'irai cracher sur vos tombes	5	3 191	6	28%
Mémoires de Guerre (Tome 2)	11	139 677	1	71%

TABLE 2 – Résumé des résultats quantitatifs par texte.

Première analyse textuelle L'application du modèle à l'essai de Nicole Baldé-Georgin (environ 200 pages), qui porte sur le malaise des jeunes, nous a permis de mettre en évidence un réseau phrastique représentatif de l'argumentation de l'auteur. Nous reportons ci-dessous, en faisant le travail d'aménagement mentionné dans la section 2, ce réseau phrastique constitué de 4 phrases partageant 3 lexèmes (*adulte, espace et social*) :

nant le mot 'juif' ou le mot 'persécution' mais pas le mot 'horreur'.

[725] Le jeune **adulte**$_1$, vivant d'une manière extrême, l'absence d'ancrage dans le **social**$_2$ piégé dans le leurre, absorbé dans sa fragilité psychique, témoignerait, de façon marquante, de ~~eet~~ l'**espace**$_3$ narcissique moderne, **espace**$_3$ nomade, en dérive, flottant entre l'être et le désêtre.
[1911] ~~Le jeune **adulte**~~$_1$, pris dans les entraves des processus identificatoires altérés et dans ses démêlés avec un **espace**$_3$ **social**$_2$ où il est impuissant à se lier et s'inscrire, en manque d'une place symbolisée qui pourrait l'articuler, *il* se laisse alors emporter dans une mouvance à risque, flottant à la dérive entre le vide et les images.
[2472] ~~Le jeune **adulte**~~$_1$ *il* découvre, alors même qu' on lui enjoint de quitter la sécurité familiale, que l'**espace**$_3$ **social**$_2$ ne lui appartient pas, et qu'il est, lui, hors circuit économique.
[2561] Les pôles identificatoires sont devenus, sous l'effet des logiques **sociales**$_2$, flous, inconsistants, inexistants car les **espaces**$_3$ qui les soustendent et dans lesquels est inséré le jeune **adulte**$_1$ sont des **espaces**$_3$ par lesquels il est contenu mais qui ne le font pas être.

On constate ici une chose assez remarquable : l'appariement à partir de 3 lexèmes permet de découvrir des relations lexicales binaires, mais fortement cohésives. Par exemple, *dérive* et *flottant* qui apparaissent dans les phrases 725 et 1911 ; mais aussi *être*, nominalisé dans 725 et entrant dans la construction factitive de 2561. Ce petit texte possède sa logique, que l'on peut résumer ainsi : l'absence d'ancrage social favorise le narcissisme du jeune et empêche sa réalisation psychique et sociale (725) ; ce rejet de l'espace social contribue à son enfermement (1911) et à sa prise de conscience de sa disqualification (sociale, donc, et économique). Les processus qui permettent habituellement le travail de réalisation psychique et sociale (les pôles identificatoires) ne sont pas à même de rendre indépendant et intègre le jeune adulte. Le texte exprime donc l'effet sur le devenir des jeunes, d'une sorte de coercition sociale qui les amènerait à fuir vers le narcissisme. D'un point de vue plus argumentatif, on peut concevoir que 725 constitue une thèse dont 1911 et 2472 sont l'élaboration (le développement) et 2561 la conséquence générale.

Deuxième analyse textuelle À titre expérimental, nous avons appliqué le modèle à un texte narratif : le tome 2 (L'Unité) des *Mémoires de Guerre* du Général de Gaulle. La démarche a été fructueuse puisqu'elle a permis d'identifier un phénomène qui interroge autant l'analyse de discours que l'histoire :

Occurrence 1
[644] Ce dernier [Hériot], ayant renvoyé sa croix de la légion d'honneur, pour marquer sa réprobation de voir décorer des « volontaires » combattant les russes, était arrêté peu après, tandis que MM. Paul **Reynaud**$_1$, **Daladier**$_2$, **Blum**$_3$, **Mandel**$_4$, le général **Gamelin**$_5$, etc., demeuraient au fond des prisons où **Vichy**$_6$ les avait jetés au lendemain de son avènement, sans que la justice les eût condamnés, ni même normalement inculpés. *[645]* **Au cours de l'été, s'aggravait la persécution des juifs, menée par un « commissariat » spécial de concert avec l'envahisseur.** *[646]* En septembre, comme le Reich **exigeait**$_7$ de la France une **main-d'oeuvre**$_8$ sans cesse plus nombreuse et que les ouvriers volontaires n'y suffisaient pas, on procédait à une levée **obligatoire**$_9$ de travailleurs. (p. 301).

Occurrence 2
[1596] **Au cours de l'hiver, redoublait la persécution des juifs malgré l'indignation publique, les protestations des évêques - comme Mgr Saliège à Toulouse, le cardinal Gerlier à Lyon - la réprobation du pasteur Boegner, président de la fédération protestante de France.** *[1597]* Le 30 janvier 1943, était créée la milice, dont Darnand, déjà incorporé dans la police allemande, devenait le secrétaire général et qui s'employait activement à traquer les patriotes. *[1598]* Le 16 février, s'instituait le service du travail **obligatoire**$_9$, procurant au "gouvernement" le moyen de fournir sans limite à l'ennemi la **main-d'oeuvre**$_8$ qu'il **exigeait**$_7$. (p. 351).

Occurrence 3
[3079] **Pendant la même période s'étalent les honteuses horreurs de la persécution juive.** *[3080]* Enfin, c'est l'époque où le Reich se fait livrer les prisonniers politiques de **Vichy**$_6$, notamment : Herriot, **Reynaud**$_1$, **Daladier**$_2$, **Blum**$_3$, **Mandel**$_4$, **Gamelin**$_5$, Jacomet, en arrête d'autres comme Albert Sarraut, François-Poncet, le colonel De La Roque, se saisit de hauts fonctionnaires, d'hommes d'affaires, d'officiers généraux et transfère en Allemagne ces personnalités afin qu'elles lui servent d'otages ou, un jour, d'éléments d'échange (p. 434).

Le calcul automatique des appariements a permis de détecter que les phrases 644 et 3080 sont appariées, ainsi que les phrases 646 avec 1598 et 645 avec 1596. En étudiant ces six phrases en contexte, on observe un réseau phrastique avec les phrases 645, 1596 et 3079 (mises en gras) qui contiennent toutes les trois les lexèmes *juif* et *persécution* et une expression temporelle. Ces phrases (séparées par un empan de 131 pages, éd. La Pléiade) constituent les seules références de de Gaulle à l'arrestation en masse des juifs en France. La quasi répétition des trois phrases est manifeste, tant au niveau lexical que syntaxique ; elle n'a pourtant jamais été perçue par les

exégètes de de Gaulle (historiens, spécialistes de lexicométrie, etc.) alors que les *Mémoires* ont été et reste un texte très « fréquenté », en raison de son caractère documentaire et de ses qualités littéraires [7]. Cette répétition concerne l'événement le plus tragique de la seconde guerre mondiale ; le relatif silence de de Gaulle (seulement trois phrases sur plus de mille pages) s'explique par le fait que cette tragédie ne peut, par son ampleur et l'impuissance de la Résistance, participer à la construction de l'Ethos gaullien – l'image du sauveur de la France (cf. (Legallois, 2010) pour une interprétation développée). L'appariement permet de plus de montrer la façon dont de Gaulle écrivain a rédigé ces passages : il est certain que de Gaulle a écrit les passages 2 et 3 en ayant sous les yeux le passage 1. Autrement dit, la répétition des phrases incriminées a été opérée de façon consciente et délibérée. Il s'agit là, selon nous, d'un fait particulièrement éclairant sur la façon dont l'après-guerre en France tout en reconnaissant l'horreur de la tragédie, en a fait un événement marginal.

4 Conclusion

Nous avons développé et utilisé des outils informatiques pour mettre en œuvre automatiquement le modèle linguistique proposé. Un premier outil calcule automatiquement les appariements et les réseaux que l'on peut ensuite afficher dans un outil de visualisation. Un second outil (système d'information logique) nous permet d'affiner l'exploration du texte par navigation. Cette première implémentation du modèle linguistique permet une exploration de corpus de grande taille en « réduisant » le texte aux phrases informationnellement les plus pertinentes et en faisant apparaître la cohésion interphrastique du texte à travers les appariements entre phrases. Les premières expériences sur différents types de textes (narratif et expositif) montrent que le modèle permet de mettre en évidence des relations discursives intéressantes qui échappent à la « simple » analyse lexicométrique, ou à l'utilisation de concordanciers classiques.

Pour le calcul des appariements, les traitements linguistiques sont basés sur l'identité des lemmes des lexèmes. Nous envisageons d'améliorer ce calcul en prenant en compte les formes dérivées (*stabiliser*/*stable*) avec l'outil *DeriF* (Namer, 2003) et les expressions temporelles et spatiales en utilisant des analyses locales comme celles que nous avons développées dans (Bilhaut *et al.*, 2003). De façon plus exploratoire, l'identification de marques anaphoriques sur des phrases consécutives devrait aussi permettre d'obtenir d'autres appariements intéressants.

Références

ALLARD P., FERRÉ S. & RIDOUX O. (2010). Discovering functional dependencies and association rules by navigating in a lattice of OLAP views. In *Concept Lattices and Their Applications*, p. 199–210 : CEUR-WS.

BILHAUT F., CHARNOIS T., ENJALBERT P. & MATHET Y. (2003). Passage extraction in geographical documents. In *Intelligent Information Processing and Web Mining*, p. 121–130.

FERRÉ S. & RIDOUX O. (2004). An introduction to logical information systems. *Information Processing & Management*, **40**(3), 383–419.

HOEY M. (1991). *Patterns of Lexis in Text*. Describing English Language. Oxford University Press.

KNIGHT K. & MARCU D. (2002). Summarization beyond sentence extraction : A probabilistic approach to sentence compression. *Artificial Intelligence*, **139**(1), 91–107.

LEGALLOIS D. (2004). Cohésion lexicale et réseaux phrastiques dans la construction du texte expositif. In S. PORHIEL & D. KLINGER, Eds., *L'unité texte* : Association perspectives.

LEGALLOIS D. (2006). Des phrases entre elles à l'unité réticulaire du texte. In *Langages*, volume 163, p. 56–73.

LEGALLOIS D. (2010). Construire un « hors-mémoires » : notes sur trois phrases des mémoires de guerre du général de gaulle. In *Les temps modernes*, volume 661, p. 116–122.

NAMER F. (2003). Morphologie et lexique. *Cahiers de Grammaire*, **28**, 31–48.

THOMAS J. (1999). *Prominence in discourse, a study based on french texts*. Allied Publishers limited.

TOOLAN M. (2009). *Narrative Progression in the Short Story : A corpus stylistic approach*. John Benjamins Publishing Company.

7. Ce texte est au programme du baccalauréat de français. On sait que la décision a fait polémique.