Université de Toulouse II - Le Mirail
Département de Sciences du Langage
Mémoire de Master 2 de Sciences du Langage parcours TAL

Directrice : Marie-Paule Péry-Woodley
Tutrice : Cécile Fabre

Deux approches de la cohésion lexicale pour le repérage de phénomènes discursifs

François Morlane-Hondère

Toulouse, le 29 juin 2009

Table des matières

Introduction

À l'heure du triomphe de la société de l'information (les plus optimistes parlent de *société de la connaissance*), le nombre de documents auquel tout un chacun se retrouve quotidiennement confronté n'a jamais été aussi important. L'avènement des médias de masse (Internet en tête), le développement de la technique et l'internationalisation des rapports humains sont autant de facteurs qui ont contribué à l'apparition de nouveaux genres de discours[1], dont le nombre est aujourd'hui difficilement calculable et tendrait même vers l'infini (Bakhtine, *Esthétique de la création verbale*, 1984, p. 265, cité dans [Adam, 1999] p. 88). La connaissance de ces vecteurs communicationnels devient ainsi un enjeu de taille :

- la sociologie des médias étudie ces nouvelles façons de communiquer afin de mettre au jour les dynamiques socioculturelles qui les sous-tendent,
- un nouveau genre de discours implique un nouvel usage de la langue et constitue donc un nouvel objet d'étude pour le linguiste,
- sur le plan industriel, une connaissance approfondie des types de discours permet d'en établir des représentations qui facilitent grandement l'automatisation de leur manipulation dans des applications comme les systèmes de recherche d'information, dont les évolutions vont de pair avec la constante augmentation des données disponibles (et leur mode d'accès qui est de plus en plus aisé).

Les genres discursifs sont culturels et sont généralement reconnus de façon intuitive par les locuteurs, qui possèdent ce que van Dijk (*Aspects d'une théorie générative du texte poétique*, 1975, cité dans [Péry-Woodley, 2008]) nomme la *compétence textuelle*[2]. Cette intuition s'appuie sur des propriétés qui sont inhérentes aux genres et qui se situent à deux niveaux.

Au niveau du texte, il est possible d'utiliser des indices comme le temps, le mode des verbes ou la présence de marques de subjectivité pour distinguer des sous-groupes. Ainsi, l'impératif sera plus caractéristique des recettes de cuisine ou des notices de montage que des contes

[1] Voir par exemple [Richer, 2006] pour une tentative de définition du blog comme genre discursif.

[2] Van Dijk emploie ici une terminologie différente de la nôtre, puisqu'il considère que le poème, le manuel de mathématiques, l'article de journal ou le questionnaire sont des types de textes, alors que ce sont pour nous des genres de discours.

pour enfants ou des billets de blog, par exemple. Cette caractérisation des textes est due à [Biber, 1988], qui a recensé 67 caractéristiques (*features*) linguistiques permettant d'établir le profil d'un texte et de l'inclure dans le sous-ensemble des textes qui ont une configuration similaire[3]. Il est intéressant de constater que [Adam, 1999], qui ne fait aucune référence à Biber, se montre réticent à l'idée d'une typologie basée sur des indices textuels :

> L'unité "texte" est bien trop complexe et trop hétérogène pour présenter des régularités linguistiquement observables et codifiables, du moins à ce niveau de complexité. (*op. cit.* p. 82)

Au niveau discursif, les genres se distinguent par des propriétés externes comme le medium par lequel ils sont véhiculés (écrit, audio, vidéo...) ou les destinataires auxquels ils s'adressent.

Ce qu'on appelle le *genre* naît de la combinaison de ces deux dimensions :

> Un genre relie ce que l'*analyse textuelle* parvient à décrire linguistiquement à ce que l'*analyse des pratiques discursives* a pour but d'appréhender sociodiscursivement. (*op. cit.* p. 83)

Il est toutefois des caractéristiques des genres qui sont à cheval entre texte et discours, comme la structure discursive qui n'est certes pas un trait *à la Biber*, mais qui s'inscrit directement dans cette problématique puisque, selon [Halliday et Hasan, 1976] (p. 327), chaque genre possède une structure qui lui est propre. Ainsi, malgré leur nom, les structures discursives sont étroitement liées à des propriétés qui se manifestent au niveau textuel. Le projet VOILADIS[4], dans lequel s'inscrit ce mémoire, vise ainsi à dégager automatiquement des structures de discours en s'appuyant sur la répartition du vocabulaire au sein des textes.

Le repérage des structures discursives est une tâche qui a mobilisé de nombreux linguistes depuis les années 60 et qui constitue encore l'un des principaux objets d'étude d'un champ entier de la linguistique (la linguistique discursive). Toutefois, devant la difficulté de définir une unité d'analyse, l'ambition d'établir une *grammaire de texte* a été revue à la baisse. Le repérage des structures discursives n'est pas tombé dans l'oubli pour autant et jouit actuellement d'un regain d'intérêt en raison du développement d'outils d'analyse linguistique rendant envisageable son utilisation en traitement automatique des langues (TAL) dans des applications comme le résumé automatique ou la navigation intradocumentaire.

Ainsi, la présente étude vise à capturer les aspects structurels des textes en s'appuyant sur le repérage de suites de mots liés entre eux par une relation sémantique. Lorsqu'elles se

[3] Bien que ce soit l'approche biberienne qui nous intéresse le plus ici (notamment à cause de la place qu'elle donne aux indices linguistiques), il existe d'autres typologies recensées dans [Charaudeau et Maingueneau, 2002] (p. 592).

[4] Projet du PRES Toulouse coordonné par C. Fabre impliquant des chercheurs des laboratoires IRIT (équipe LiLac) et CLLE-ERSS (axes TAL et S'caladis).

manifestent au sein des textes, ces relations créent des liens de *cohésion* qui permettent de donner une unité à un ensemble de phrases, qui, sans cela, serait considéré comme un *non-texte*. Ainsi, les phénomènes de cohésion sont plus que de simples faisceaux de relations lexicales : nous verrons que l'étude des phénomènes cohésifs peut servir de moyen d'accès aux propriétés discursives des textes, traditionnellement abordées à l'aide *cue phrases* et autres *introducteurs de cadres* ([Charolles, 1978]).

Nous entamerons ce mémoire par un point sur les notions de *dimension textuelle* et de *dimension discursive* et sur les notions qui leurs sont liées, comme la cohésion et la cohérence (ch. 1). Nous verrons ainsi que la principale caractéristique des textes est qu'ils possèdent une cohésion interne dont les manifestations seront décrites au chapitre 2, notamment à travers la typologie développée dans [Halliday et Hasan, 1976]. Les trois chapitres suivants sont consacrés au versant TAL du mémoire : nous présenterons dans un premier temps les données que nous avons mobilisées pour mener à bien nos expériences (ch. 3), la première consistant en une automatisation de la méthode de repérage des liens de discours basée sur une analyse des relations lexicales développée dans [Hoey, 1991] (ch. 4), la seconde visant à détecter les chaînes lexicales qui parcourent un texte pour essayer de voir si la cohésion est un indice permettant de déduire le rôle discursif du titre qui le chapeaute (ch. 5).

Chapitre 1

Les dimensions textuelle et discursive

1.1 Le texte et le discours

Étant donné que ces notions reviennent à de nombreuses reprises tout le long de ce texte, il nous paraît nécessaire d'entamer cette étude par une mise au point terminologique. Cette dernière s'avère d'autant plus nécessaire qu'il n'est pas rare de voir les définitions de *texte* et de *discours* changer du tout au tout d'un auteur à l'autre : [Tanskanen, 2006] (p. 3) fait ainsi remarquer que certains auteurs n'utilisent qu'un seul des deux termes, que d'autres utilisent les deux indifféremment, ou que la définition de l'un chez un auteur X peut définir l'autre chez un auteur Y. ([Adam, 1999] évoque J.-P. Bronckart, "qui a tendance à appeler "texte" ce que presque tout le monde convient d'appeler "discours" et inversement" (p. 84)). En ce qui nous concerne, nous partons de la conception de la relation texte/discours développée chez Adam :

> Discours = Texte + Conditions de Production
> Texte = Discours – Conditions de Production (*op. cit.* p. 39)

Malgré sa séduisante simplicité, cette formule présente le défaut de suggérer que *texte* et *discours* sont des entités que l'on peut isoler de façon rigoureuse, mathématique. Or, le fait de présenter le discours comme un *texte en contexte* sous-entend qu'il existe une unité *texte* isolable de ses "conditions de production et de réception-interprétation" (c'est la définition que donne Adam du contexte (*ibid.* p. 39)). Nous trouvons écho à cette remarque dans [Détrie *et al.*, 2001] :

> Un texte ne livre pas son sens *in abstracto*, il est relié à ses conditions de production, à la typologie à laquelle il s'ancre, et donc à un arrière-fond sociohistorique, autant de domaines qui lui confèrent une cohérence : la matérialité linguistique qui permet de construire la cohésion textuelle (syntaxe, lexique, formes verbales, etc.) doit donc ainsi être reliée à la matérialité extra-linguistique : l'ancrage du texte dans une **situation de communication**, contrainte sociohistoriquement. Le traiter

> dans un cadre étroitement intralinguistique, c'est l'amputer d'une grande part de ses potentialités signifiantes[1]. (*op. cit.* p. 58)

Il est évident qu'Adam ne nie pas cette interdépendance, mais nous préférons insister sur le fait que cette décontextualisation du texte est méthodologique, étant donné qu'un texte produit *ex nihilo* est inconcevable. Ce *texte*-là est un objet artificiel, né de la dissociation des deux facettes d'une même unité pour permettre l'analyse des phénomènes purement linguistiques en se préservant des interférences que pourrait générer la prise en compte d'éléments liés au contexte, donc à la dimension discursive. Nous verrons qu'une telle herméticité n'est pas adaptée à l'étude à laquelle nous nous livrons ici, puisque nous partons d'indices textuels pour mettre au jour des structures discursives (lesquelles dépendent du genre dans lequel s'inscrit le texte). Le fait même d'accepter qu'il y ait une relation entre vocabulaire et structure implique la remise en question de cette scission (Adam considère que cette distinction "tend à s'estomper" (*op. cit.* p. 40–41) et cite certains auteurs pour qui elle n'a plus lieu d'être). Dans la présente étude, nous avons retenu les définitions de [Adam, 1999], qui correspondent *grosso modo* aux acceptions les plus répandues dans la littérature. En effet, même si les notions de *texte* et de *discours* vont parcourir l'ensemble de la présente étude, nous ne pousserons pas plus loin la discussion ni ne chercherons de définitions plus sophistiquées car nous risquerions de perdre de vue notre objectif principal.

Il est intéressant de mentionner la démarche de D. Apothéloz, qui adopte une approche plus interactionnelle : il considère que le texte (*discours-objet*) est un objet pré-construit qui ne s'active que dans le but d'"agir sur des états de la pensée, [de] transformer [...] un ensemble E1 d'états en un ensemble E2 d'états" ([Apothéloz, 1985] p. 610). Cette *activation* (le *motif*) correspondrait à la contextualisation du texte chez Adam. De plus, chez Apothéloz, le discours n'est pas considéré comme une unité d'analyse statique mais comme "un processus consistant à conférer un statut d'opération à des contenus" (*ibid.* p. 611), processus qu'évoque Adam lorsqu'il parle d'"ouvrir le texte [...] sur une situation d'énonciation-interaction toujours singulière et [...] sur l'interdiscursivité dans lequel chaque texte est pris." (*op. cit.* p. 40)

Une deuxième acception de *texte*, qui est celle que nous utilisons par la suite, présente le texte comme "un énoncé complet, le résultat toujours singulier d'un acte d'énonciation" ([Adam, 1999] p. 40), "l'unité de l'interaction humaine" (*idem*). Cela correspond à la définition de Halliday & Hasan ("a unit of language in use" [Halliday et Hasan, 1976] p. 1) ainsi qu'à son acception la plus répandue dans les ouvrages de linguistique générale.

[1] En plus de mettre l'accent sur la complémentarité des plans textuels et discursifs, cette citation a cela d'intéressant qu'elle introduit les notions de *cohésion* et de *cohérence*, que nous étudions au chapitre 1.2.

1.2 Cohésion textuelle et cohérence discursive

Dans la section précédente, nous avons choisi d'aborder les notions de *texte* et de *discours* en les mettant directement face à face (il nous a en effet paru maladroit de les présenter séparément sachant qu'elles sont intimement liées), et nous ne les avons définies que l'une par rapport à l'autre. Cette approche ne nous a pas permis d'aborder les propriétés intrinsèques de chacune de ces notions. C'est notamment le cas de la notion du *texte*, qui n'a, dans un premier temps été défini que comme le reste de la soustraction du contexte au discours, puis comme une unité communicative. Ainsi, nous avons principalement fondé nos conceptions de ces termes sur celles qu'on retrouve dans [Adam, 1999] et qui sont notamment appuyées par une citation de D. Maingueneau qui met l'accent sur une propriété des textes que nous analyserons plus en profondeur dans le reste de cette section :

> En parlant de *discours*, on articule l'énoncé sur une situation d'énonciation singulière ; en parlant de *texte*, on met l'accent sur ce qui lui donne son unité, qui en fait une totalité et non plus une simple suite de phrases. (*op. cit.* p. 41)

Cette définition du texte est encore une fois négative, puisqu'un texte est défini par ce qu'il n'est pas, à savoir "une simple suite de phrases", un *non-texte*. La différence réside dans ce que [Halliday et Hasan, 1976] nomment la *texture* (p. 2). C'est cette propriété qui fait qu'une série de phrases est considérée ou non comme un texte (pour rappel, "a unit of language in use" p. 1). Ce jugement est à la portée de tout locuteur[2] et correspond à ce que [Charolles, 1978] appelle la *compétence textuelle* (p. 8).

Au niveau textuel opère la *cohésion*. Cette notion a été principalement développée dans [Halliday et Hasan, 1976], qui la définissent comme une relation sémantique qui se manifeste à la fois à travers la grammaire et le lexique (d'où la dichotomie *cohésion grammaticale vs. lexicale*). Dans tous les cas, elle a pour propriété de créer des liens (*ties*) de différentes natures entre deux éléments du texte, même (et surtout) si ces derniers n'apparaissent pas dans la même phrase. En effet, même si cette relation peut s'observer entre éléments d'une même phrase, c'est au niveau interphrastique qu'elle est la plus pertinente puisque c'est de sa capacité à jeter des ponts entre les phrases que naît l'impression de texture. Ainsi, il est possible de repérer dans l'exemple suivant plusieurs procédés cohésifs, dont nous ferons une description détaillée au chapitre suivant :

> La voiture refuse de démarrer. Elle vient pourtant d'être révisée et la batterie est

[2] [Fayol, 1986] évoque les travaux de E. Esperet, qui a démontré que les enfants acquièrent la capacité de distinguer un texte d'un non-texte vers l'âge de sept ans.

neuve.

La cohésion est ici signalée par une relation d'anaphore (*Elle* reprend *la voiture*), un marqueur structurel[3] (*pourtant*) et des liens sémantiques associant *voiture* à *révisée* et *batterie*. L'on pourrait également lier *voiture* à *démarrer*, mais étant donné que ces deux mots se trouvent dans la même phrase, leur association n'est pas pertinente du point de vue de la cohésion globale.

La notion de cohésion se trouve toutefois limitée quand il s'agit de rendre compte de l'unité manifeste d'énoncés comme :

La voiture refuse de démarrer. Il a dû faire sacrément froid cette nuit !

Cet exemple ne présente aucune marque de cohésion mais peut pourtant être considéré comme un texte. La raison en est que la liaison qui unit les deux phrases ne se situe pas au niveau du texte mais à celui du discours. On observe un phénomène similaire avec des énoncés d'une phrase, voire d'un mot comme *Port du casque obligatoire*, *Attention à la marche*, ou *Stop*. Étant donné que Halliday & Hasan considèrent que seule la cohésion interphrastique participe à la texture (p. 9), il faut ici aussi chercher les raisons de l'acceptabilité de ces énoncés dans la dimension discursive : ces exemples ne seront considérés comme acceptables que si leur réalisation satisfait un certain nombre de paramètres situationnels ([Péry-Woodley, 2000] p. 14). Ainsi, l'indication *Port du casque obligatoire* ne sera pertinente que si elle apparaît à l'entrée d'un chantier et non dans une maison de retraite, par exemple.

On a ici affaire à un phénomène qui se rapproche de la cohésion du fait qu'il nécessite que deux éléments aient un certain niveau de compatibilité pour pouvoir se manifester. Cependant, alors que la cohésion relève de liens entre unités de niveau textuel, cette nouvelle relation, la *cohérence*, opère principalement au niveau extralinguistique, et mobilise même un ensemble de variables dont le nombre et la nature en font une notion relativement difficile à aborder (et encore plus à formaliser).

Ainsi, l'exemple précédent n'est cohérent que pour un locuteur dont les connaissances du monde sont suffisantes pour savoir qu'une voiture a un moteur, et qu'il arrive qu'un moteur exposé trop longtemps à une température trop basse refuse de démarrer (on peut parler de *scénario*). Et même si les inférences nécessaires à l'interprétation de tels exemples peuvent être facilitées par la présence de marques de cohésion, il est clair que le facteur des connaissances extralinguistiques joue un rôle primordial dans le processus communicationnel. On pourrait également citer le critère des genres discursifs et de la connaissance qu'en ont les locuteurs en reprenant l'exemple souvent cité du théâtre de l'absurde, dont l'enchaînement des répliques en-

[3]Nous verrons plus loin que [Halliday et Hasan, 1976] parlent dans ce cas de "conjonction adversative" (p. 250).

traînera un jugement d'inacceptabilité chez le béotien. Du point de vue psycholinguistique, un texte peut être qualifié d'*incohérent* quand les efforts nécessaires à la production des inférences que nécessite sa compréhension sont trop importants pour le récepteur (ce critère est donc fondamentalement subjectif[4]). [Sanford, 2006] illustre cette notion d'effort à l'aide de deux expériences où sont mesurés les temps de lecture d'une phrase en fonction de son co-texte droit. Soient les énoncés suivants :

1a) Harry took the picnic things fom the trunk.

1b) Harry took the beer from the trunk.

2) The beer was warm.

Les résultats montrent que les temps de lecture de l'énoncé 2 quand il est précédé de 1b sont plus bas que quand il est précédé de 1a. On peut en déduire que l'inférence BEER *is part of* PICNIC THINGS nécessaire pour surmonter la rupture entre les deux énoncés nécessite effectivement une charge cognitive plus importante que dans les cas où l'enchaînement des énoncés est fluidifié[5] par la présence d'un procédé cohésif (ici, la répétition de *beer*). La même expérience a été reproduite avec les énoncés suivants :

3a) John drove to London yesterday.

3b) John took his car to London yesterday.

4) The car broke down halfway.

Les résultats montrent que dans ce cas, les temps de lecture de l'énoncé 4 ne varient pas quelle que soit la phrase qui le précède. On peut expliquer ce phénomène par le fait que l'association de *drive* avec *car* est tellement ancrée chez les locuteurs que la reprise *the car* dans 4 se fait

[4]Il n'est pas rare dans la littérature de voir les notions de cohérence et de cohésion respectivement associées à la subjectivité et à l'objectivité (notamment dans [Tanskanen, 2006] p. 21). Toutefois, si nous ne remettons pas en question le caractère éminemment subjectif de la cohérence, nous verrons qu'il est plus qu'abusif de considérer la cohésion comme un paramètre purement objectif. Voir [Morris, 2004] et [Morris et Hirst, 2005] pour une étude expérimentale sur la part de subjectivité qui intervient dans l'interprétation des liens de cohésion lexicale.

[5]Nous parlons de *fluidification* dans le sens où le lien de cohérence entre deux énoncés consécutifs est instinctivement présupposé par tout locuteur et non créé par la cohésion :

> Il est impossible de voir quelqu'un accomplir deux actions successives sans supposer que ces deux actions constituent un tout : nous imaginons nécessairement qu'elles font partie d'une seule intention globale qui justifie le fait qu'elles aient été entreprises l'une après l'autre. ([Charolles, 1978] dans [Péry-Woodley, 2008])

On pourrait également évoquer les *standards de textualité* de de Beaugrande (que cite [Tanskanen, 2006] p. 27) que sont l'*intentionnalité* et l'*acceptabilité*, qui correspondent respectivement à la volonté du locuteur de produire un discours cohésif et cohérent et à l'attitude de l'interlocuteur qui s'attend à recevoir un discours qui sera également cohésif et cohérent.

sans le moindre effort. Nous verrons plus loin que nous avons ici affaire à un cas de collocation. Cette relation est manifestement beaucoup moins marquée entre *picnic things* et *beer*.

Ces expériences sont particulièrement intéressantes car elles mettent en lumière le fait que des phénomènes considérés comme linguistiques (la répétition dans la première expérience, la collocation dans la deuxième) influencent directement les raisonnements inférentiels, considérés comme discursifs, ce qui remet sérieusement en question l'opposition cohérence *vs.* cohésion. Concernant cette dernière remarque, nous citons M. Charolles qui abonde dans ce sens :

> En l'état actuel des recherches [...] il ne semble plus possible techniquement d'opérer une partition rigoureuse entre les règles de portée textuelle et les règles de portée discursive. Les grammaires de texte font éclater les frontières généralement admises entre la sémantique et la pragmatique, entre l'immanent et le situationnel, d'où, à notre avis, l'inutilité présente d'une distinction cohésion-cohérence que d'aucuns proposent en se fondant justement sur un partage précis de ces deux territoires. ([Charolles, 1978] p. 14)

Le point de vue de Halliday & Hasan à ce sujet est encore aujourd'hui l'objet de conflits d'exégèse : étant donné qu'ils n'emploient que rarement le terme *coherence*, certains auteurs considèrent que ce concept correspond en fait à ce qui est désigné par la notion de *texture*, dans le sens où les deux ont pour propriété de pouvoir distinguer un texte d'un non-texte. D'autres considèrent que la cohérence se retrouve chez Halliday & Hasan dans le concept de *registre* :

> Just as one can construct passages which seem to hang together in the situational-semantic sense, but fail as texts because they lack cohesion, so also one can construct passages which are beautifully cohesive but which fail as texts because they lack consistency of register – there is no continuity of meaning in relation to the situation. The hearer, or reader, reacts to both of these things in his judgement of texture. ([Halliday et Hasan, 1976] p. 23)

On pourrait également mentionner le concept de *situation*, qui semble largement recouper la définition de *coherence* :

> The term SITUATION, meaning the 'context of situation' in which a text is embedded, refers to all those extra-linguistic factors which have some bearing on the text itself. (*ibid.* p. 21)

> The question is, what are the external factors affecting the linguistic choices that the speaker or writer makes. (*ibid.* p. 21)

Le concept de *registre* semble néanmoins subsumer celui de *situation* :

> A text is a passage of discourse which is coherent in [. . .] two regards : it is coherent with respect to the context of situation, **and therefore consistent in register**"' (*ibid.* p. 23) [nous soulignons]

Il a été démontré de façon empirique que cohésion et cohérence sont "à la fois indépendants et entrelacés" ([Tanskanen, 2006] p· 21 [notre traduction]), puisque la plupart des textes nécessitent leur association pour pouvoir être considérés comme tels[6]. Toutefois, nous avons vu qu'il est possible, et même fréquent, de rencontrer et d'interpréter sans le moindre problème des textes totalement dépourvus de cohésion. On pourra objecter que les exemples donnés dans la littérature se limitent à deux voire trois échanges[7], qu'ils se situent quasi-systématiquement dans le domaine de la conversation et qu'il est probable que tout texte d'une longueur supérieure ferait vite apparaître des marques de cohésion, mais cela n'enlève rien au fait que les énoncés non cohésifs sont bel et bien attestés, ce qui n'est pas le cas d'énoncés dépourvus de cohérence comme l'exemple suivant, cité dans [Brown et Yule, 1983] (p. 197 ; emprunté à Enkvist, *Coherence, pseudo-coherence and non-coherence*, 1978, p. 110) :

> I bought a Ford. A car in which President Wilson rode down the Champs Elysées was black. Black English has been widely discussed. The discussions between the presidents ended last week. A week has seven days. Every day I feed my cat. Cats have four legs. The cat is on the mat. Mat has three letters.

Il est clair que les nombreux liens cohésifs présents dans cet exemple n'en font pas une unité de communication.

Ainsi, la cohérence semble être une condition première à l'acceptabilité des textes, et la cohésion sa manifestation de surface (on pourrait parler d'épiphénomène si la nécessité de la cohésion n'était pas avérée). [Widdowson, 1978] illustre la primauté de la cohérence sur la cohésion à l'aide des exemples suivants :

1) -What is the police doing ?

[6]Nous verrons toutefois que les opinions divergent quant aux rapports qu'entretiennent ces deux dimensions. Alors que des auteurs comme [Meurer, 2003] mettent empiriquement en évidence un lien entre le nombre de marques de cohésion et le jugement de cohérence, [Mosenthal et Tierney, 1985] recensent d'autres études qui ont démontré que la cohésion ne pouvait à elle seule rendre compte de la cohérence (p. 243).

[7]On peut notamment citer l'exemple souvent repris de [Widdowson, 1978] (p. 29) :

A : That's the telephone.

B : I'm in the bath.

A : O.K.

-They are arresting the demonstrators.

2) -What is the police doing ?

-The fascists are arresting the demonstrators.

(*op. cit.* p. 27)

Dans le premier échange, la reprise du sujet *the police* est explicitement marquée par le pronom *they*, qui n'a de signification que s'il est rattaché à un syntagme nominal (c'est un élément cohésif *par nature* [Halliday et Hasan, 1976] p. 19). Ici, l'effort interprétatif nécessaire au rattachement de *they* à *the police* est minimal, si ce n'est qu'il faut identifier *the police* comme une institution et *they* comme un sous-ensemble de ses membres. Dans l'exemple 2, le lien est moins évident puisque le pronom est remplacé par un SN qui possède une certaine autonomie au niveau référentiel. Il revient au lecteur/auditeur d'inférer que ce syntagme est anaphorique, en s'appuyant éventuellement sur ses connaissances du monde (le fait qu'il est répandu dans certains milieux de considérer les membres de la police comme des fascistes). Ainsi, la relation de synonymie entre *police* et *fascists* est créée discursivement[8], et participe à la cohésion du texte. Nous verrons dans le chapitre suivant quels sont les autres procédés qui permettent au locuteur/scripteur de satisfaire le besoin de cohésion exprimé par les impératifs communicationnels en tissant des liens entre les mots d'un texte.

[8]Dans la sémantique interprétative de F. Rastier, ce cas serait considéré comme une afférence de sème.

Chapitre 2

Les manifestations cohésives de la cohérence

La première moitié de ce chapitre est consacrée à la typologie des relations permettant de rendre compte de la cohésion dressée par [Halliday et Hasan, 1976]. Nous verrons que l'une d'elles, la collocation, présente un intérêt significatif au vu de l'objectif que nous nous sommes donnés dans ce mémoire : elle présente le paradoxe d'être considérée comme la relation qui permet le mieux de rendre compte de la cohésion [Morris et Hirst, 2004] et d'être à la fois la plus mal définie. Nous verrons au chapitre 3 que le choix des bases lexicales que nous avons mobilisées a été en partie orienté par le fait que le repérage de ce type de relation nous permettra d'accéder à un pan entier des phénomènes cohésifs habituellement négligé car difficilement appréhendable.

2.1 La classification de Halliday & Hasan

Nous entamons ce chapitre par une section consacrée à la typologie des procédés cohésifs qu'ont dressée M.A.K. Halliday & R. Hasan dans leur ouvrage *Cohesion in English*, que de nombreux auteurs reconnaissent comme un standard dans le domaine de la linguistique textuelle. Nous adoptons toutefois un point de vue légèrement différent du leur, dans le sens où leur classification fait la part belle aux procédés cohésifs que sont la référence, la substitution, l'ellipse et la conjonction (relevant tous de la cohésion grammaticale), dont les descriptions respectives occupent entre 45 et 80 pages de leur ouvrage, alors que seulement 18 pages sont consacrées à la cohésion lexicale (alors que selon [Hoey, 1991], ce sont les relations sémantiques entre les éléments lexicaux d'un texte qui permettent le mieux de générer de la cohésion (p. 9)). Or, étant donné que la présente étude se situe dans une approche résolument lexicale,

nous ne ferons qu'évoquer les principales propriétés des procédés de cohésion grammaticale (que nous prenons le temps de décrire un minimum pour donner un point de vue global sur la typologie) pour nous focaliser sur la cohésion lexicale, et en particulier sur ce que Halliday & Hasan nomment la *collocation*.

2.1.1 La cohésion grammaticale

La référence

Sont considérées comme des marques de référence[1] les utilisations de pronoms, déterminants ou adverbes dont l'interprétation passe par le rattachement à un élément du co-texte, voire du contexte. Dans le premier cas, la reprise peut se faire soit sur un élément du co-texte gauche, auquel cas on parle d'*anaphore* :

Je suis allé voir *une comédie musicale*. *Elle* était formidable.

Soit, et c'est beaucoup plus rare en français, sur le co-texte droit (on parle alors de *cataphore*) :

Il est venu *le temps des cathédrales* !

Dans le cas où la reprise se fait sur un élément situé hors du texte, Halliday & Hasan parlent de référence *exophorique* (p. 31), par opposition à la référence *endophorique*, dont on vient de voir les deux types de manifestation, l'anaphore et la cataphore, dans les exemples précédents :

(Deux personnes passent devant une affiche de *Notre-Dame de Paris*)

C'est *cette* comédie musicale que tu es allé voir ?

Cette catégorisation s'applique ensuite à toutes les marques de référence, divisées en référence *personnelle* (pronoms personnels, déterminants et pronoms possessifs), *démonstrative* (pronoms et adverbes démonstratifs) et *comparative* (adjectifs et adverbes de comparaison).

La substitution et l'ellipse

N.B. Étant donné que l'ellipse est considérée comme une forme de substitution où une unité est remplacée par une unité vide, nous avons décidé de regrouper ces deux procédés dans la même sous-section. Il est également à noter que [Hoey, 1983] classe la substitution comme une sous-classe de la répétition.

[1] [Brown et Yule, 1983] p. 192 préfèrent parler de *co-réference*.

La substitution se décline sous ses formes nominale, verbale ou *clausale* (ces trois catégories s'appliquent également à l'ellipse). La substitution nominale est typiquement marquée en anglais par *one*, que l'on peut directement traduire en français par *un* et que l'on retrouve dans des énoncés comme :

Le voisin m'a montré sa nouvelle télé à écran plasma. J'en veux une !

Une se substitue ici à *télé à écran plasma*. Contrairement à la référence, l'élément substitué ne correspond pas à celui qui a été instancié, comme c'est le cas dans l'exemple suivant, où *elle* correspond à *la nouvelle télé à écran plasma du voisin* :

Le voisin m'a montré sa nouvelle télé à écran plasma. Elle est gigantesque.

La substitution verbale se manifeste en anglais par le remplacement d'un syntagme verbal par *do*, dont la polyvalence ne trouve pas d'équivalent en français. Soit l'exemple suivant, extrait de [Hoey, 1991] (p. 6) :

-Does Agatha sing in the bath ?

-No, but I do.

L'équivalent français de la deuxième réplique ne ferait probablement pas apparaître de forme verbale et se rapprocherait plus de l'ellipse :

-Est-ce qu'Agathe chante sous la douche ?

-Non, mais moi si.

On peut cependant observer un phénomène similaire avec des locutions comme *en faire autant*, *en faire de même*, etc.

La substitution *clausale* correspond au remplacement d'une proposition entière et s'exprime avec *so* :

-Is there going to be an earthquake ?

-It says so.

([Halliday et Hasan, 1976] p. 130)

On pourrait traduire la deuxième réplique (!) de l'exemple ci-dessus par un énoncé elliptique comme *Il paraît*, ou par ce qui correspond le mieux, à nos yeux, au *so* anglais, c'est-à-dire le démonstratif *ce* dans *C'est ce qu'on dit*.

La conjonction

La conjonction est un type de cohésion qui relie deux phrases à l'aide d'adverbes ou de syntagmes prépositionnels. Ces éléments conjonctifs ne portent pas sur un mot ou un groupe de

mots particuliers, mais expriment des relations qui peuvent englober des ensembles de phrases. Ces relations se divisent en quatre types[2].

La conjonction *additive* se rapproche de la coordination et s'exprime avec des marqueurs comme *et* ou *de plus* :

> Il avait escaladé la montagne toute la journée sans s'arrêter. *Et* il n'avait rencontré personne.

La conjonction *adversative* exprime une certaine rupture avec les attentes du lecteur/auditeur concernant la suite du texte. Elle se manifeste par *cependant*, *toutefois*, *mais*, etc.

> Il avait escaladé la montagne toute la journée sans s'arrêter. *Mais* il était à peine fatigué.

La conjonction *causale*, comme son nom l'indique, exprime une relation de cause à effet (*ainsi*, *par conséquent*, etc.) :

> Il avait escaladé la montagne toute la journée sans s'arrêter. *Tant et si bien qu'*à la nuit tombée, la vallée était loin derrière lui.

La conjonction *temporelle* exprime une succession dans le temps (*après*, *ensuite*, *deux jours plus tard*, etc.) :

> Il avait escaladé la montagne toute la journée sans s'arrêter. *Puis*, à la nuit tombée, il se reposa.

2.1.2 La cohésion lexicale

La réitération

Ce que Halliday & Hasan nomment la *réitération* englobe plusieurs phénomènes cohésifs qui vont de la répétition pure et simple d'une unité lexicale à sa reprise par un mot générique (*general word*).

L'ensemble des procédés relevant de la réitération sont illustrés dans l'exemple suivant, que nous empruntons à [Halliday et Hasan, 1976] (p. 279).

> I turned to the ascent of the peak. $\left\{ \begin{array}{l} \textit{The ascent} \\ \textit{The climb} \\ \textit{The task} \\ \textit{The thing} \end{array} \right\}$ is perfectly easy.

[2]Les exemples qui suivent sont adaptés des exemples donnés dans [Halliday et Hasan, 1976] (p. 238–239).

L'ordre dans lequel sont présentés ces procédés n'est pas anodin puisqu'il reflète l'augmentation du degré de généralité des unités lexicales utilisées pour reprendre *ascent*. Ainsi, le premier procédé utilisé est la répétition simple, accompagnée par le marqueur anaphorique *the*, qui indique que cette seconde occurrence du mot *ascent* renvoie au même référent que dans la phrase précédente (la question de la référence n'est toutefois pas pertinente en ce qui concerne l'étude de la cohésion lexicale ([Halliday et Hasan, 1976] p. 282–284)). Le deuxième procédé est l'utilisation d'un synonyme, ici *climb*. Étant donné que la question de savoir dans quelle mesure deux mots sont plus ou moins synonymes reste difficile à trancher (et que de toute façon ce n'est pas l'objet de *Cohesion in English*), Halliday & Hasan règlent le problème en englobant avec les synonymes la catégorie plus vague des *quasi-synonymes* (*near-synonyms*). Le troisième procédé fait appel à une autre relation lexicale qui est celle d'hypéronymie. Avec *task*, on s'éloigne du sens de *ascent* en remontant à un niveau de sens plus global, étant donné que *task* subsume à la fois *ascent* et *climb*. À un niveau supérieur, *thing* englobe un ensemble de réalités encore plus important et se distingue par le fait qu'il est lui-même quasiment vide au niveau sémantique. Cette propriété est celle des éléments primaires que l'on retrouve au sommet des réseaux lexicaux traditionnels. Ces mots, les *general nouns*, sont à la limite du lexique et de la grammaire, dans la mesure où leur ensemble est quasiment fermé (*ibid.* p. 280). Halliday & Hasan les distinguent cependant de mots comme *one* et *do* vus précédemment dans le cadre de la substitution, qui constituent les classes ultimes de la *hiérarchie lexicale* pour ce qui est des noms et des verbes, mais qui constituent une classe fermée, et qui relèvent par conséquent de la grammaire (c'est la raison pour laquelle la substitution n'a pas été traitée avec la cohésion lexicale).

La version originale de l'exemple ci-dessus fait figurer un dernier niveau où *the ascent* est remplacé par *it*. Sachant que ce cas a été traité dans la partie *référence*, nous n'avons pas jugé utile de le reproduire ni de le commenter ici.

La collocation

Nous avons vu avec la réitération que les relations lexicales que sont la synonymie et l'hypéronymie permettaient de tisser des liens de cohésion entre deux mots. Le cas de la synonymie se rapproche de celui de la répétition simple, dans le sens où cette relation implique une certaine identité sémantique entre deux mots. Dans le cas de l'hypéronymie, la reprise d'un mot par l'un de ses super-ordonnés modifie le niveau de grain auquel on considère le premier mot, mais son sens est conservé. Ainsi, on peut désigner un chat par ses synonymes *matou*, *minet*, etc. ou bien par ses hypéronymes *félin*, *animal*, *bête*, etc., et dans tout les cas, le sens de *chat* est conservé (on parle de *transitivité* dans le cas de l'hypéronymie). On peut alors aisément

parler de réitération, dans le sens où la forme du mot n'est certes pas répétée (comme dans le cas de la répétition), mais le contenu sémantique, lui, est réitéré. Toutefois, il s'avère que les paires de mots dont l'identité du point de vue sémantique est loin d'être aussi évidente peuvent également être porteuses de liens de cohésion.

Le premier exemple donné dans *Cohesion in English* est celui de l'antonymie. Il pourrait à première vue sembler paradoxal qu'un lien de cohésion puisse naître de l'association de mots dont les sens sont par nature opposés, mais il s'avère que les antonymes partagent en réalité la plus grande partie de leur sens et ne se distinguent que par rapport à l'une de leurs dimensions, comme celle du sexe dans *garçon* et *fille*, qui partagent les traits [+humain] et [+jeune]. Ainsi, il est indéniable que des paires de mots opposés comme *garçon* et *fille* partagent plus de sens que des mots qui ne présentent à première vue aucune opposition comme *trottoir* et *oreille*, et que cette proximité sémantique est suffisante pour créer un lien de cohésion. Il est à noter que Halliday & Hasan évoquent séparément de l'antonymie la relation qui unit les membres d'une classe fermée comme celle des jours de la semaine ou des grades d'une hiérarchie, alors que certains auteurs s'accordent à considérer cette relation comme un contraste non binaire, un sous-genre de l'antonymie.

La deuxième relation lexicale évoquée est celle de la méronymie (la relation *partie-tout*), qui relie des mots comme *voiture* et *frein*, *porte* et *poignée*, *livre* et *couverture*, etc. L'identité sémantique des membres de ces paires est ici quasiment nulle : dans l'univers extralinguistique, une porte comporte généralement une poignée, mais le sens de *poignée* n'est pas contenu dans celui de *porte*, et réciproquement.

Halliday & Hasan évoquent pour finir le cas des co-hyponymes comme *table* et *chaise*, qui partagent le même hypéronyme *meuble*. C'est là que s'arrête ce début de classification de ces paires de mots qui véhiculent de la cohésion à l'aide d'une relation sémantique qui est, dans la plupart des cas, extrêmement difficile à caractériser. Soient les exemples suivants, traduits de [Halliday et Hasan, 1976] (p. 285) :

rire/blague	jardin/creuser	essayer/réussir	porte/fenêtre	bateau/rame
lame/aiguisé	malade/docteur	abeille/miel	roi/couronne	soleil/nuage

Ces paires de mots sont manifestement liées, mais le lien qui les rattache échappe à toute catégorisation. Il serait en effet difficile de mettre un nom sur les relations qu'entretiennent les paires ci-dessus sans tomber dans la caractérisation *ad hoc*, comme ce serait le cas d'une hypothétique relation ATTRIBUT TYPIQUE DE qui relierait *roi* et *couronne*[3], ou d'une autre que l'on pourrait gloser par ACTIVITÉ POSSIBLE DANS pour *jardin* et *creuser*. On pourrait éventuellement repérer

[3][Chaffin et Herrmann, 1984] parleraient dans ce cas de relation *agent-objet* (p. 136).

un semblant de relation de cause/conséquence dans les paires *rire/blague* et *essayer/réussir*, de la co-hyponymie dans *porte/fenêtre* et de l'antonymie dans *soleil/nuage* et *malade/docteur* (antonymie converse), mais la plupart des paires entretiennent des relations tellement spécifiques qu'il serait impossible de toutes les nommer.

Ainsi, Halliday & Hasan regroupent tous ces couples *marginaux* dans ce qu'ils nomment, de façon un peu malencontreuse, la *collocation*. Car en effet, comme [Hoey, 1991] le fait remarquer (p. 7), le terme *collocation* désignait bien avant la parution de *Cohesion in English* un phénomène statistique attribué au linguiste J. R. Firth caractérisant deux mots qui apparaissent ensemble dans un même co-texte plus souvent qu'une distribution aléatoire ne le laisserait prévoir. Or, ce critère n'est pas pris en compte dans la catégorie des collocations telle que l'envisagent Halliday & Hasan (qui relève plus de la psychologie que de la statistique). Cette classe accueille ainsi tous les couples d'unités lexicales véhiculant de la cohésion mais ne relevant ni de la répétition, ni de la synonymie, ni de l'hypéronymie (puisque ces cas-là sont considérés comme de la réitération). C'est cette variété[4] de relations qu'embrasse la collocation qui en fait une catégorie difficile à aborder[5] mais riche en promesses.

2.2 La relation de collocation

2.2.1 Collocation : sens large *vs.* sens restreint

Nous revenons dans cette section sur le concept de *collocation*, déjà évoqué dans la typologie de Halliday & Hasan, afin d'en donner une définition un peu moins orientée et peut-être plus valorisante que celle d'une *catégorie-poubelle* où échouent les relations qui n'ont pu être classées ailleurs. La définition que nous mobilisons dans le cadre de cette étude correspond à une approche en corpus dite *contextualiste*, dans le sens où la signification d'un mot est considérée comme largement tributaire du contexte dans lequel il apparaît et est ainsi étroitement liée au concept de cohésion ([Dubreil, 2008]). Cette approche est considérée par E. Dubreil comme étant la "définition large" de la collocation, par opposition aux définitions qui considèrent que le phénomène n'englobe que les expressions figées comme *peur bleue*, *vitesse fulgurante* ou *buveur invétéré*, soit des expressions idiomatiques plus ou moins opaques aux structures syntaxiques fortement contraintes (*op. cit.*) :

nom + adjectif (épithète) : *amour platonique*

nom (sujet) + verbe : *la colère s'apaise*

[4]Le mot est faible. [Hoey, 1991] parle même de "ragbag of lexical relations" (p. 7).

[5]La fugacité des liens qui unissent les éléments de cette classe a poussé Hasan à l'exclure de son champ d'étude dans ses travaux ultérieurs.

verbe + nom (objet) : *commettre une agression*

...

Nous excluons ces phénomènes de notre champ d'étude, étant donné que la contiguïté des deux collocats ne permet pas d'établir des liens de cohésion à la portée suffisamment importante pour être considérée comme pertinente (si l'on suit la conception de Halliday & Hasan, ce type de collocation est même exclu d'office des vecteurs de cohésion puisque les collocats sont confinés aux frontières de la phrase[6]). Le concept qui nous intéresse se caractérise également par le fait qu'il est exclusivement lexical. Nous laissons en effet de côté les collocations grammaticales comme *eu égard de*, que certains auteurs regroupent sous le concept de *colligation*.

Au final, la définition que donne [Dubreil, 2008] de la *collocation textuelle* semble répondre à nos attentes puisqu'elle synthétise trois propriétés de la collocation qui sont indissociables de l'approche que nous envisageons :

- ne sont prises en compte que les collocations lexicales, "construites avec des mots pleins",
- les collocats sont susceptibles d'apparaître dans des enchaînements,
- le critère de bonne formation syntaxique (le fait de correspondre à un schéma pré-défini, comme les exemples ci-dessus) n'est pas pris en compte.

2.2.2 Une relation mal définie

Cette section rapporte le point de vue de divers auteurs qui se sont retrouvés confrontés à la notion de collocation au sens de Halliday & Hasan, sans forcément l'identifier comme telle. [Morris et Hirst, 2004] appellent *non-classical relations* les relations qui sortent du cadre des typologies traditionnelles (*synonymie*, *antonymie*, *hypéronymie*, etc.). Cette dénomination traduit un certain malaise à l'égard de ces relations, dont la nature est si difficilement définissable qu'on a du mal à leur attribuer un nom. Ainsi, dans le dictionnaire analogique paru chez Larousse ([Niobey *et al.*, 2007]), des mots comme *miauler*, *ronronner* et *chatière* sont associés à l'entrée *chat* par la relation RELATIF À (les thésaurus anglo-saxons parlent de *related terms* ([Morris et Hirst, 2004])).

Nous avons vu précédemment que la psycholinguistique s'était également intéressée à des phénomènes de ce type :

> La représentation de "donner un coup de poing" renferme également quelques-unes des conséquences possibles telles que : blessures éventuelles, douleur, désir d'être soulagé, etc. Pendant la lecture [...], le lecteur "active" la représentation de

[6] Les énoncés comme *Do not feed*, *For sale*, et *National Westminster Bank* sont toutefois considérés comme des textes dans *Cohesion in English* (p. 294), alors que, comme nous l'avons vu plus haut, ils ne tirent leur acceptabilité que du fait qu'ils apparaissent dans des contextes bien particuliers, et non grâce à leurs propriétés linguistiques.

> “donner un coup de poing” en faisant travailler sa mémoire et prévoit ce qui peut arriver ensuite. ([Haberlandt, 1982] p. 734)

Ainsi, quand on rencontre dans un texte un concept qui est inclus dans la représentation d'un concept précédemment activé, on parle de *script*, ou de *scénario*. Ces derniers impliquent une lecture plus fluide du texte, et en cela ils se rapprochent du concept de cohésion.

Dans le domaine de la lexicologie, [Coseriu, 1970] cite l'exemple suivant (emprunté à Ch. Bally, *L'arbitraire du signe. Valeur et signification*, 1940) :

> Le mot “bœuf” fait penser : 1) à “vaches, taureau, veau, cornes, ruminer, beugler”, etc. ; 2) à “labour, charrue, joug”, etc. ; enfin 3) il peut dégager et dégage en français, des idées de force, d'endurance, de travail patient, mais aussi de lenteur, de lourdeur, de passivité.

Il se pose la question de savoir dans quelle mesure toutes ces associations peuvent être considérées comme *linguistiques* et conclut de cet exemple que :

- *vache*, *taureau* et *veau* entretiennent un rapport d'opposition avec *bœuf* dans un même champ lexical (ils sont co-hyponymes),
- *cornes* et *ruminer* peuvent être perçus comme des traits distinctifs dans la définition de *bœuf*,
- *beugler* et *joug* entretiennent un rapport d'*implication lexicale* avec *bœuf*[7],
- *charrue*, *labour* et les idées de force, etc. évoquées dans le 3) de l'exemple de Bally sont exclues du domaine de la linguistique.

Concernant cette dernière affirmation, Coseriu considère que “‘charrue’ et ‘labour’ sont un objet et un état de choses qui se trouvent souvent *dans le contexte réel* de l'objet ‘bœuf’ (il n'y a aucun rapport lexicalement nécessaire et définissable entre les lexèmes ‘charrue’, ‘labour’ et le lexème ‘bœuf’).” On peut cependant avancer sans trop de risque le fait que si deux objets du monde extralinguistique se retrouvent souvent associés “dans le contexte réel”, cette association se reflétera au niveau linguistique. Et l'on peut, par la même occasion, faire l'hypothèse que si les relations qu'embrasse le concept de collocation sont si difficiles à appréhender, c'est qu'elles relèvent de l'extralinguistique et qu'elles désignent ainsi des relations entre objets du monde réel et non entre mots[8] : deux mots peuvent avoir des sens similaires, opposés, imbriqués, etc.,

[7] Il est d'ailleurs étonnant de voir que la relation *bœuf/joug* est considérée plus proche de *bœuf/beugler* que de *bœuf/charrue*, étant donné que *joug* et *charrue* renvoient tous deux à des instruments alors que *beugler* désigne le cri de l'animal, qui en est ainsi indissociable et qui correspond alors effectivement à ce qu'on pourrait considérer comme une *implication lexicale*.

[8] Cette idée est également présente dans [Morris et Hirst, 1991] (p. 22).

mais une association comme celle de *miel* avec *abeille* ne relève pas du lexique[9]. On peut ainsi s'avancer à énoncer les trois propriétés suivantes du phénomène de collocation :

1. il est dans un premier temps pragmatique : "deux objets du monde extralinguistique se retrouvent souvent associés *dans le contexte réel*",
2. il s'inscrit dans la dimension psychologique à partir du moment où cette association est perçue par les locuteurs,
3. il se reflète au niveau linguistique à travers les productions des locuteurs et peut se mesurer à l'aide de méthodes statistiques (information mutuelle, t-score, etc.).

Plus proche de nous et directement en lien avec l'approche TAL, [Morris et Hirst, 2004] démontrent que la plupart des relations relevées par des locuteurs lors d'une tâche de lecture sortent du cadre des relations dites *classiques* que sont la synonymie, l'antonymie, l'hypéronymie et la méronymie. On retrouve par exemple :

– des sous-ensembles liés à une notion non-lexicalisée comme la suite *funérailles/né/marié*, que l'on peut considérer comme appartenant à un ensemble ÉTAPES DE LA VIE,
– des paires nom/adjectif typiques : *professeur/brillant* (on peut ici parler de collocation au sens restreint du terme),
– des couples de mots qui se retrouvent souvent associés et qui peuvent entretenir un rapport qui a trait à un lieu (*funérailles/chapelle*), à une notion de cause/conséquence (ou assimilée : *alcoolique/sevrage*), etc.

Les auteurs mettent ainsi en lumière l'importance de ces relations comme procédés cohésifs et se demandent dans quelle mesure ces relations sont préexistantes ou créées textuellement. Nous verrons que cette différence a son importance dans la problématique des ressources mobilisées pour repérer ces relations.

[9]Une telle affirmation peut également s'appliquer à la relation de méronymie, puisque l'inclusion de la *partie* dans son *tout* est extralinguistique.

Chapitre 3

Les données mobilisées

Après avoir abordé les points les plus théoriques du mémoire dans les deux chapitres précédents, nous entrons maintenant dans la partie applicative. Comme c'est souvent le cas en TAL, la question des ressources est de première importance dans notre étude. En effet, la performance d'un système de repérage des liens de cohésion comme le nôtre dépend principalement de la richesse de la ressource sur laquelle il s'appuie, puisque les relations qu'il projettera sur les textes sont celles qui ont été définies *a priori* dans la base de données de référence.

Ainsi, afin de pouvoir couvrir la plus grande partie possible du spectre des relations sémantiques permettant de générer de la cohésion, nous avons décidé de combiner trois ressources de différentes natures :

- une base de voisins distributionnels (les *Voisins de Wikipédia*) (3.1.1),
- un réseau lexical obtenu à partir des sorties du programme *JeuxDeMots* (3.1.2),
- un dictionnaire de synonymes (*Dicosyn*) (3.1.3).

Au-delà d'une simple accumulation de ressources tout-venant, l'intérêt d'une telle démarche se situe dans la complémentarité escomptée des données mobilisées : nous verrons dans ce chapitre que chacune d'entre elles possède un mode de constitution qui lui est propre, ce qui entraîne une certaine variété dans les types de relations détectées qui ne peut nous être que profitable.

La première section de ce chapitre est ainsi consacrée à la description de ces trois ressources et aux modifications que nous leur avons faites subir avant de les amalgamer. Puis nous abordons, dans la deuxième section, la question du texte sur lequel les expériences ont été menées, le *texte de travail* (un article de l'encyclopédie Wikipédia), et du genre dans lequel il s'inscrit.

3.1 Les ressources lexicales

3.1.1 Les voisins de Wikipédia

Description

Le concept de *voisins distributionnels* est issu du domaine de l'analyse distributionnelle, développée par L. Bloomfield dans les années 30. Elle s'appuie sur l'idée selon laquelle la seule observation des contextes d'apparition des mots d'une langue (leur *distribution*) permettrait d'en établir la grammaire. Un corollaire de cette théorie est que les mots qui ont des distributions similaires, appelés *voisins*, ont des chances de partager des éléments de sens ([Harris, 1991]). Cette analyse peut ainsi permettre de rapprocher des mots sémantiquement proches comme *château* et *forteresse* en se basant sur le fait que ces deux noms apparaissent en position objet de verbes comme *construire*, *bâtir*, *détruire*, qui s'avèrent eux aussi sémantiquement proches. C'est cette propriété de l'analyse distributionnelle qui fait tout l'intérêt d'une ressource comme les voisins de Wikipédia.

Cette dernière a été générée par le programme Upéry ([Bourigault, 2002]) à partir d'un corpus constitué de l'ensemble des articles de la version francophone de Wikipédia, soit plus de 470 000 articles pour 194 millions de mots[1]. Le processus permettant d'obtenir une base de voisins distributionnels à partir d'un corpus se divise en trois étapes :

1. le corpus est étiqueté par TreeTagger[2],
2. il est ensuite traité par Syntex, qui extrait les relations syntaxiques sous forme de triplets <gouverneur, relation, dépendant>. Ainsi, le programme va repérer dans la phrase *Pierre mange un biscuit* les triplets <manger, suj, Pierre> et <manger, obj, biscuit>. Quand la relation de dépendance syntaxique se fait via une préposition, cette dernière prend la place de la relation au sein du triplet (*biscuit au chocolat* se représente <biscuit, à, chocolat>),
3. afin de faciliter leur traitement par le logiciel Upéry, les triplets obtenus sont ramenés sous la forme de couples <prédicat, argument> où le prédicat correspond au gouverneur auquel on accole la relation, et où l'argument correspond au dépendant (<biscuit, à, chocolat> devient <biscuit_à, chocolat>). Cette formalisation va permettre d'opérer un double rapprochement : celui des prédicats partageant les mêmes arguments, mais aussi celui des arguments partageant les mêmes prédicats. Concrètement, dans notre base de

[1] Le corpus a été recueilli courant avril 2007. À l'heure où nous écrivons (mai 2009), le nombre d'articles est de 809 185. Le recueil du corpus, son traitement ainsi que la création de la base de voisins sont dus au travail de Franck Sajous (CLLE-ERSS).

[2] Université de Stuttgart (www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/)

voisins, le prédicat *manger_obj* est rapproché de *se nourrir_de* via les arguments *pousse*, *bourgeon*, *crustacé*, etc., et l'argument *biscuit* est rapproché de *sucre* via les prédicats *fabrique_de*, *marque_de*, *production_de*, etc. Lors de cette étape, le programme attribue à chaque paire de voisins deux scores de proximité qui indiquent dans quelle mesure les distributions des deux mots rapprochés sont similaires. Les calculs utilisés pour obtenir ces scores sont l'indice de Jaccard et la mesure de Lin ([Lin, 1998]). Le premier consiste, pour deux arguments *a1* et *a2*, à diviser le nombre de prédicats qu'ils ont en commun *P* par la somme de leurs productivités[3] respectives *n1* et *n2* à laquelle on soustrait *P*, soit :

$$sim_{jacc}(a1, a2) = \frac{P}{n1 + n2 - P}$$

L'autre méthode, le calcul du Lin, se déroule en deux étapes. Une première consiste à calculer la *quantité d'information* (QI) de chaque prédicat et argument, c'est-à-dire le rapport du nombre d'arguments/prédicats avec lesquels ils se combinent dans le corpus, sur la totalité des arguments/prédicats avec lesquels ils pourraient se combiner. Soit un prédicat *p*, *a* le nombre de ses arguments dans le corpus et *n* le nombre de ses arguments potentiels (le nombre total de noms que contient le corpus) :

$$QI = -log\left(\frac{a}{n}\right)$$

Il s'agit ensuite de rapprocher les arguments/prédicats entre eux en s'appuyant sur le nombre de prédicats/arguments qu'ils partagent et de diviser la QI de ces cooccurrents syntaxiques communs aux deux voisins (multipliée par 2) par la somme des QI respectives de chacun des voisins. Ainsi, si l'on considère deux arguments *a1* et *a2*, et *P* le nombre de prédicats avec lesquels ils se combinent dans le corpus, le calcul de leur similarité est le suivant :

$$sim_{lin}(a1, a2) = \frac{2 * QI(P(a1) \cap P(a2))}{QI(P(a1)) + QI(P(a2))}$$

NB : Ces calculs valent aussi bien pour la mesure de la similarité de deux arguments, que pour celle de deux prédicats.

Cette méthode possède l'avantage de permettre les rapprochements intercatégoriels qui font cruellement défaut aux ressources actuellement disponibles[4]. Ainsi, le verbe prédicat *manger_obj* se retrouve également associé à des prédicats nominaux comme *bouillon_de*, *cuisson_de*, *recette_de* ou *plat_de* via des arguments communs comme *viande*, *poulet*, *poisson*,

[3]La productivité d'un argument ou d'un prédicat correspond au nombre de prédicats ou d'arguments différents avec lesquels il entretient une relation syntaxique dans le corpus.

[4]Le projet EuroWordNet avait tenté une approche intercatégorielle qui restait toutefois limitée aux mots présentant des similarités morphologiques ([Fabre et Bourigault, 2006]).

spaghetti, etc.

La base ainsi obtenue[5] compte environ quatre millions de couples, dont le degré de pertinence (au niveau sémantique) varie énormément : les structures syntaxiques présentes dans la langue générale, dont relève le corpus Wikipédia[6], sont relativement variées, ce qui génère une certaine quantité de bruit parmi les résultats. En effet, si les exemples de rapprochements que nous avons utilisés plus haut font apparaître des liens sémantiques comme la synonymie (*manger_obj/se nourrir_de*), la collocation (*manger_obj/plat_de*) ou la méronymie (*biscuit/sucre*), ces cas restent difficiles à distinguer des rapprochements non pertinents de façon automatique parmi la masse de résultats obtenus.

Traitement

Comme nous l'avons vu dans l'introduction de ce chapitre, chacune des trois bases lexicales que nous comptons fusionner a été constituée d'une façon particulière et possède certaines caractéristiques qui lui sont propres. Cela se traduit par des singularités dans la structure de chacune d'elles qu'il nous faudra gérer afin que la ressource obtenue suite à leur fusion soit la plus homogène possible. Les premières modifications que nous avons apportées à la base de voisins sont assez triviales. Il convient toutefois d'en faire une courte description afin de garantir la reproductibilité de notre expérience.

Dans la chaîne de traitement qui permet d'obtenir une base de voisins à partir d'un corpus s'insère une opération bien connue du TAL, la lemmatisation. Chaque voisin de la base a été réduit à sa forme canonique, ce qui ne pose *a priori* pas de problème étant donné que c'est également le cas pour les autres ressources (on pourrait argumenter sur le fait que certains mots ont des distributions différentes selon qu'ils sont au singulier ou au pluriel). Toutefois, cette lemmatisation pose problème dans le cas des syntagmes nominaux, puisque chacun de leurs éléments se retrouvent dépourvus de marques de flexion, ce qui rend des syntagmes comme *affaire étranger*, *pression artériel* ou *corée de nord* inexploitables en l'état puisqu'il faudrait rétablir les accords en fonction des occurrences *usuelles* de chacun d'eux (en plus d'accorder l'adjectif en fonction du nom, il faudrait également pouvoir définir que *affaire étrangère* est également à passer au pluriel). Et malgré le fait que cette rectification pourrait s'effectuer assez facilement à l'aide de la ressource adéquate, nous ne l'avons pas réalisée pour la bonne et simple raison que nous avons décidé de ne pas prendre en compte la plupart les locutions (nous

[5]Une ressource générée selon la même procédure à partir d'un corpus constitué de dix ans d'articles du journal *Le Monde* est consultable à l'adresse suivante : http ://www.irit.fr :8080/voisinsdelemonde/

[6]Pour des études concernant l'analyse distributionnelle automatique sur corpus de langue générale, se référer à [Bourigault et Galy, 2005] et [Fabre et Bourigault, 2006].

justifions ce choix à la section 3.2). Nous avons toutefois rétabli l'élision dans des cas comme *se améliorer*, *se opposer*, *se unir*, etc.

Comme nous l'avons vu au début de cette section, l'un des avantages d'une base de voisins calculée par Upéry est la double mise en relation prédicat/prédicat et argument/argument. Cela implique la cohabitation de deux types de paires au sein d'une même base, dont la différence réside dans le fait que le champ *relation* de chaque argument affiche un caractère *underscore* au lieu d'une préposition ou d'un nom de relation (*suj* pour *sujet*, *obj* pour *objet* ou *mod* pour la relation épithète), comme c'est le cas pour les prédicats :

```
construire   V   obj   bâtir       V   obj   0.13   0.19
château      N   _     forteresse  N   _     0.17   0.27
```

Ces deux lignes correspondent aux relations entre les prédicats *construire_obj* et *bâtir_obj*, et les arguments *château* et *forteresse* telles qu'on peut les trouver dans la base qui nous a servi de point de départ. Dans les deux cas, les trois premiers champs sont consacrés à la première entité de la paire et les trois suivants se rapportent à la deuxième entité. Chacune d'entre elles est ainsi caractérisée par, dans l'ordre :

1. son lemme,
2. sa catégorie grammaticale,
3. pour les prédicats, la relation qui lui est attachée.

Les deux derniers champs de l'exemple ci-dessus sont consacrés aux deux indices de similarité dont nous avons parlé plus haut, le Jaccard et le Lin.

Dans notre démarche d'unification des ressources, nous avons dû éliminer dans un premier temps les champs *relation* de chaque prédicat et argument étant donné que ce type d'information est absent des autres bases lexicales[7]. Nous avons toutefois bien conscience que cette opération ampute la base de voisins de l'une de ses caractéristiques les plus intéressantes. En effet, comme le font remarquer [Bourigault et Galy, 2005], le contexte syntaxique associé à chaque prédicat permet une description des relations qui va au-delà de la simple association de lemmes. Ainsi, alors que Dicosyn (cf. 3.1.3) met en relation les verbes *souligner* et *insister*, l'analyse fournie par Upéry rapproche les prédicats *souligner_obj* et *insister_sur*, ce qui signifie qu'il est quelque peu erroné de rapprocher hors contexte les lemmes *souligner* et *insister* dans le sens où la proximité distributionnelle (et donc, idéalement, sémantique) de ces deux verbes ne se constate que quand *insister* est associé à la préposition *sur*. Nous devrons donc ici

[7]La base JeuxDeMots compte un certain nombre de paires liées par les relations *Agent typique* et *Patient typique*, qui peuvent, dans une certaine mesure, être assimilées aux relations *suj* et *obj* qu'assigne Upéry. Nous avons choisi de ne pas prendre en compte ce type de relations pour des raisons que nous expliquons en 3.1.2.

nous passer de ce genre de précisions en misant sur le fait qu'un tel degré de finesse n'est pas indispensable à la tâche que nous visons.

Une fois les champs *relation* de chaque paire de prédicats/arguments retirés, la base passe de 3 922 657 à 2 766 809 paires. Cela est dû au fait qu'une partie d'entre elles, dans la base originale, ne se distinguaient que par le type de relation syntaxique associées à leurs prédicats, comme dans les deux paires suivantes qui, au final, n'en forment plus qu'une[8] :

```
absence     N     (de)     rester     V     (suj)
absence     N     (de)     rester     V     (dans)
```

Une autre conséquence a été le fait de voir apparaître des paires où les deux mots reliés étaient identiques, comme dans l'exemple ci-dessous. Ces dernières ont été effacées : le traitement de la répétition a été implémenté dans le programme de repérage des liens, ce qui rend ce genre de relations inutiles.

```
employer     V     (suj)     employer     V     (obj)
employer     V     (pour)    employer     V     (dans)
```

Une dernière opération a été de *désymétriser* la base. En effet, toute paire présente dans la base possède sa forme réfléchie, comme dans l'exemple ci-dessous :

```
absence     N     (de)     rester     V     (suj)
rester      V     (suj)    absence    N     (de)
```

Ces deux paires sont redondantes, puisqu'elles représentent une seule et même relation : lors de la projection des voisins, le programme interprète une relation *A/B* comme étant également une relation *B/A*, ce qui rend superflue la cohabitation dans la base de deux paires *A/B* et *B/A*. Ainsi traitée, la base est ramenée à un total de 1 455 395 paires, ce qui reste particulièrement volumineux.

Nous avons vu que chaque couple de voisins de cette base possédait deux scores de proximité, ce qui signifie que toutes ces paires n'ont pas le même degré de pertinence : plus deux prédicats/arguments ont des distributions similaires et plus ils ont de chance d'être sémantiquement proches. Or, les paires de voisins qui ne manifestent aucune relation sémantique n'ont pas d'intérêt pour nous puisque nous visons le repérage de relations de cohésion. Il convient ainsi de fixer un seuil afin d'obtenir un rapport optimal entre la proportion de paires dont le rapprochement syntaxique est motivé sémantiquement et le nombre total de voisins.

[8] Les scores de Jaccard et de Lin de la paire restante sont ceux de la paire qui avait le Lin le plus élevé.

Cette question du seuillage est extrêmement délicate puisque plus le seuil est élevé et moins le nombre de paires récupérées est important. De plus, étant donnée la taille de la base, il est impossible d'évaluer *manuellement* la pertinence des paires contenues dans un sous-ensemble des voisins obtenu pour un seuil donné. Et il est également impossible de s'appuyer sur une ressource de référence, puisque chaque base de voisins est le reflet du corpus sur lequel elle a été calculée. Toutefois, afin de se faire une idée de la proportion de relations sémantiques pertinentes que l'on peut retrouver dans une base lexicale comme la nôtre (et afin de nous aider, éventuellement, à fixer un seuil), nous avons mesuré le recouvrement entre les voisins de Wikipédia et Dicosyn. Cette expérience avait déjà été effectuée dans [Bourigault et Galy, 2005] avec une base de voisins calculée sur dix ans d'articles du journal *Le Monde*, et le taux de recouvrement se situait aux alentours de 1%, ce qui est très faible. Sur les voisins de Wikipédia, il est un peu supérieur puisque nous l'avons calculé à 1,5%. Dans les deux cas, les calculs ont été effectués sur l'ensemble des voisins. Or, comme l'on pouvait s'y attendre, cette valeur varie considérablement quand on isole des sous-parties de la base en fonction d'un seuil donné : afin de mieux se rendre compte de ce phénomène, nous avons mesuré le taux de recouvrement entre Dicosyn et différentes sous-parties de la base de voisins. Les résultats, présentés aux figures 3.1 et 3.2, demandent à être commentés.

Sur l'axe des abscisses se trouvent les valeurs auxquelles ont été faites les mesures : ces dernières ont été effectuées tous les 0.01, de 1 (la valeur maximale pour le Jaccard et le Lin) à 0,03 pour le Jaccard et à 0,10 pour le Lin (leurs valeurs minimales respectives). Pour chacun des indices, le nombre de paires rapportées augmente de façon exponentielle, si bien que le nombre de relations pour les seuils inférieurs à environ 0.40 est indiscernable. On peut voir que c'est en utilisant la mesure de Lin que l'on obtient le taux de recouvrement maximal, qui est de 7,04% avec un seuil à 0,34, alors qu'il est de 6,56% pour un Jaccard de 0,28. Il est toutefois impossible de filtrer la base sur ces seuils, étant donné que le nombre de paires que l'on obtiendrait serait respectivement de 22 081 et 12 868 (dont 1554 et 844 paires de synonymes), ce qui est loin d'être satisfaisant car cela reviendrait à sous-exploiter considérablement la base. Ainsi, l'augmentation du nombre de voisins que l'on désire garder implique un compromis entre rappel et précision puisque passé un certain seuil, l'augmentation du nombre de paires de la base suppose une baisse de sa concentration en paires synonymiques.

3.1.2 JeuxDeMots

Description

Développée au LIRMM (Université de Montpellier 2), la base JeuxDeMots correspond à un réseau lexical de nature un peu particulière. En effet, cette ressource est alimentée par les

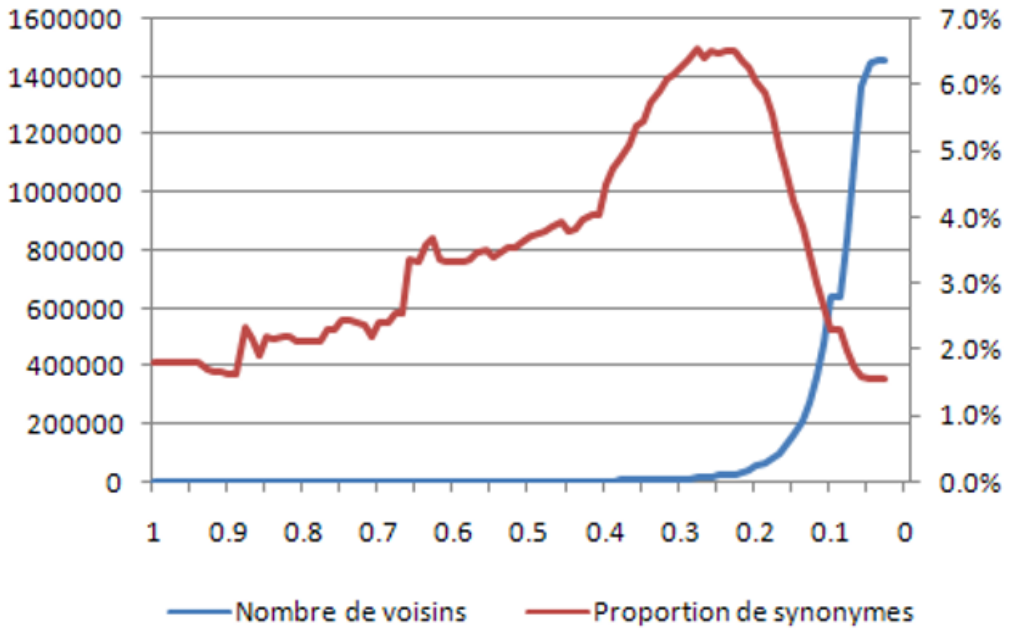

FIG. 3.1 – Mesure des taux de recouvrement entre Dicosyn et des sous-parties de la base de voisins filtrée en fonction de l'indice de Jaccard.

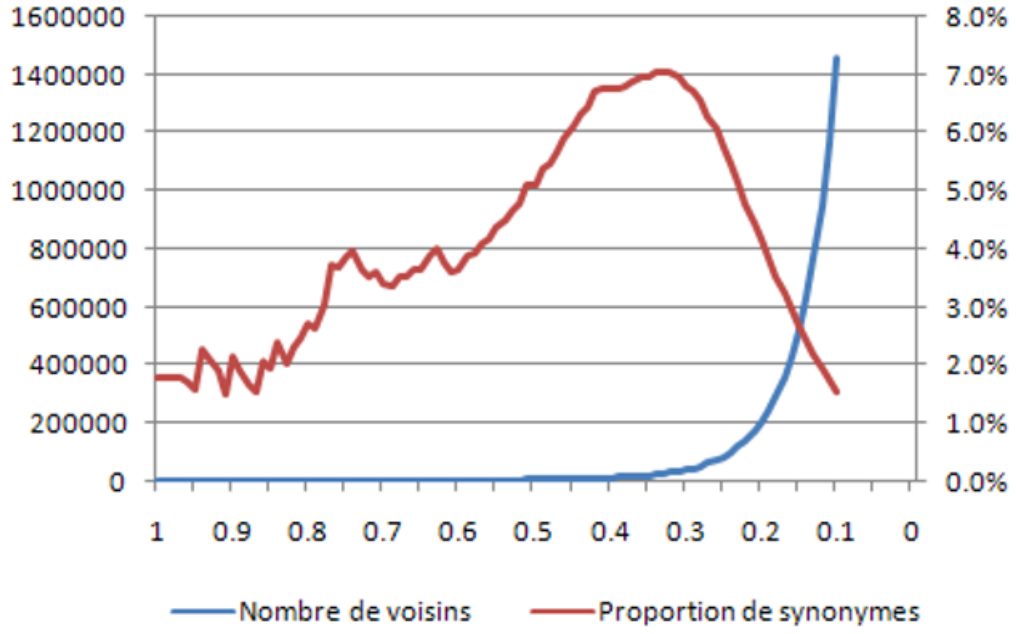

FIG. 3.2 – Mesure des taux de recouvrement entre Dicosyn et des sous-parties de la base de voisins filtrée en fonction de la mesure de Lin.

contributions d'internautes (quel que soit leur niveau de connaissance de la langue) à travers un jeu[9] dont le principe est de fournir des mots susceptibles d'être reliés à un mot donné par une relation donnée. Le réseau peut être consulté via une interface en ligne[10] ou directement téléchargé[11].

Une partie se déroule de la façon suivante : le programme présente au joueur une consigne comme "Donner des idées associées au terme suivant" ainsi qu'un terme sélectionné plus ou moins aléatoirement[12] parmi une base d'environ 152 000 termes continuellement enrichie des mots inconnus cités comme réponses par plusieurs joueurs. Le joueur a ensuite soixante secondes pour fournir ses réponses, lesquelles seront comparées à celles des autres joueurs, les relations communes étant ajoutées au réseau ou renforcées si elles étaient déjà présentes. Si la partie n'avait pas encore été proposée, les réponses du joueur sont sauvegardées en attendant qu'un autre joueur y confronte les siennes. Interviennent ensuite un ensemble de barèmes de pondération selon la fréquence à laquelle les mots sont donnés comme réponse pour une relation ou le degré d'efficacité du joueur, qui accumule des *points d'honneur* au fil des parties (plus le joueur est efficace, plus les relations qu'il propose seront considérées comme pertinentes).

L'un des intérêts majeurs de cette méthode est qu'elle permet d'accéder aux avantages d'une ressource construite manuellement sans pour autant en avoir les inconvénients, qui sont le coût et le temps nécessaire à sa constitution. Les relations obtenues grâce à JeuxDeMots ont été comparées avec celles d'une ressource construite à la main de façon traditionnelle, à savoir EuroWordNet : étant donné qu'il a été décidé que les liens tissés dans JeuxDeMots ne seraient pas vérifiés par un expert avant d'être validés, les résultats montrent que le taux de précision d'EuroWordNet est bien plus élevé que celui de JeuxDeMots ([Lafourcade, 2007]). Ce résultat est imputé à la nouveauté de la ressource, dont la précision augmentera d'elle-même à l'aide du système de pondération (et notamment grâce à la notion d'érosion, qui fait qu'un lien perd de son poids quand il n'est pas mobilisé). Les résultats restent toutefois encourageants puisque pour les 100 termes les plus fréquemment proposés pas les joueurs, 97% des associations ont été considérées comme correctes.

À l'heure où nous écrivons (mai 2009), les statistiques du site nous indiquent que la base compte 166 551 termes pour 198 763 relations (contre environ 23 000 termes et 100 000 relations pour EuroWordNet). Ces nombres sont en constante augmentation : environ 6000 relations sont ajoutées chaque mois.

[9] http ://www.lirmm.fr/jeuxdemots/jdm-accueil.php

[10] http ://www.lirmm.fr/jeuxdemots//rezo.php

[11] http ://www.lirmm.fr/ lafourcade/JDM-LEXICALNET-FR/

[12] Quand un mot a déjà reçu suffisamment de réponses pour une relation donnée, ses chances d'être sélectionné diminuent (il est considéré comme *tabou*).

Relation	Description	Exemple
FAMILLE	dérivation morphologique	*alcool/alcootest*
SYNONYME	synonymie	*peureux/froussard*
CONTRAIRE	antonymie (NB : pas seulement *contraire*)	*dur/mou*
SPÉCIFIQUE	hyponymie	*arbre/chêne*
GÉNÉRIQUE	hypéronymie	*vache/ruminant*
PARTIE	méronymie	*pied/orteil*
TOUT	holonymie	*processeur/ordinateur*
IDÉE ASSOCIÉE	collocation	*chat/miauler*
DOMAINE	domaine relatif au terme	*fourchette/restauration*
CHOSE>LIEU	lieux typiques où peut se trouver l'objet	*clavier/bureau*
ACTION>LIEU	lieux typiques où peut se réaliser l'action	*manger/cantine*
CARACTÉRISTIQUE	propriétés que peut avoir l'objet	*poire/juteuse*
LOCUTION	locutions en rapport avec le terme	*corde/corde à sauter*
MAGN	amplification du terme	*bon/délicieux*
ANTIMAGN	affaiblissement du terme	*haine/antipathie*
MANIÈRE	manières dont peut être exécutée une action	*marcher/rapidement*
ACTION>AGENT	agents typiques d'une action	*réparer/mécanicien*
ACTION>PATIENT	patients typiques d'une action	*cuisiner/repas*
ACTION>INSTRUMENT	instrument grâce auquel une action peut être réalisée	*chasser/carabine*
SENS/SIGNIFICATION	les acceptions possibles d'un terme	*peinture/tableau*
SENTIMENT	sentiments ou émotions éprouvés par rapport au terme	*actrice/admiration*

TAB. 3.1 – Liste des relations permettant de relier deux termes dans JeuxDeMots.

En dehors de la méthode astucieuse de collecte de données qui caractérise JeuxDeMots, l'aspect de cet outil qui a suscité notre intérêt repose dans le type de relations qu'il propose de tisser. En effet, en plus de relier les mots de sa base par des relations *classiques*, le programme mobilise des relations comme celle d'*idée associée*, qui se rapproche de la définition de la collocation. L'éventail des relations proposées a été rapporté dans le tableau 3.1.

Nous ne nous livrerons pas ici à une description détaillée de ces catégories, qui, bien qu'elles aient toutes leur place dans un réseau lexical, ont tendance à empiéter les unes sur les autres. C'est notamment le cas des relations MAGN et MANIÈRE, qui peuvent recouper la catégorie

LOCUTION (*peur/peur bleue*[13], *manger/manger sur le pouce*), ou de la relation *idées associées*, qui est tellement vague qu'elle englobe souvent une partie des autres relations.

Comme l'indiquent les doubles traits, nous avons distingué plusieurs sous-ensembles parmi les relations :

- la catégorie FAMILLE se distingue des autres du fait qu'elle ne fait intervenir aucun critère sémantique (sauf si l'on suppose que les dérivés morphologiques d'un mot auront forcément un lien sémantique avec leur mot d'origine, ce qui est une supposition plutôt risquée au vu de couples comme *front/fronton*, *bouteille/embouteillage*…),
- la deuxième catégorie regroupe les relations lexicales dites *classiques*,
- la troisième catégorie englobe (peut-être de façon un peu abrupte) les relations que l'on pourrait considérer comme des collocations au sens large.

Cette distinction reste évidemment arbitraire et n'aura aucune incidence sur la façon dont nous traiterons les couples de mots. De plus, le problème de recoupement des relations que nous avons évoqué plus haut n'en sera probablement pas un pour nous : étant donné que les chaînes lexicales se forment en s'appuyant sur des relations qui sont souvent distinctes, il est probable que nous ne tiendrons compte que du fait qu'un mot est relié à un autre, quelle que soit la relation.

Traitement

Dans sa forme *brute*, le réseau est un fichier texte contenant les nœuds et les relations :

- un nœud se présente sous la forme suivante :
  `eid=35423:n="crépiter":t=1:w=54`
  Il se définit selon 4 caractéristiques :
  1. son identifiant (*35423*),
  2. le mot qu'il représente (*crépiter*),
  3. son type (*terme*, *concept*, *partie du discours*… le chiffre *1* signifie ici que le nœud est un terme, comme la plupart des nœuds),
  4. son poids (*54*), qui grossit selon la fréquence à laquelle le mot revient dans le jeu.
- une relation a la forme suivante :
  `rid=408443:n1=52660:n2=54271:t=25:w=50`
  Elle se définit selon 5 caractéristiques :
  1. son identifiant (*408443*),

[13] Qui, soit dit en passant, constitue une collocation au sens strict et qui diffère de la notion de collocation que nous employons dans le reste de ce chapitre.

2. l'identifiant du premier nœud de la relation (*52660*, qui correspond ici au mot *clé*),
3. celui du deuxième (*54271*, soit le mot *ouvrir*),
4. son type (*synonymie*, *antonymie*, *méronymie*... ici *25*, une relation *instrument*),
5. son poids (*50*).

Afin d'obtenir un format compatible avec les autres ressources, la première démarche à effectuer est de mettre en correspondance les nœuds et les relations à l'aide d'un système de gestion de base de données ou d'un programme *ad hoc* (c'est cette dernière solution pour laquelle nous avons opté).

Comme on peut le constater dans la description que nous avons faite de la structure des éléments du réseau, les informations que sont le lemme et la catégorie grammaticale de chaque mot ne sont pas codées directement dans les nœuds. La correspondance entre un mot et son lemme se fait via une relation particulière (non sémantique donc). Les lemmes ne sont toutefois pas codés de façon particulière, ce qui implique que les mots qui ont la même forme que leur lemme sont reliés à eux-mêmes. Les catégories grammaticales sont présentes sous la forme de nœuds. Elles sont au nombre de 137, soit le type de nœud le plus fréquent après le *terme*. Ce nombre s'explique par le fait que ces catégories sont particulièrement précises, puisqu'en plus de coder la catégorie grammaticale, elles comprennent le temps auquel est conjugué le verbe, le genre, le nombre, etc. Chaque nœud-terme est ainsi relié à l'un de ces 137 nœuds-catégories via une relation *pos* (*part of speech*, qui correspond ici au numéro 4). Par exemple, la relation

```
rid=356062:n1=52584:n2=146894:t=4:w=50
```

relie

```
eid=52584:n="vrillée":t=1:w=50
```

à

```
eid=146894:n="Adj:Fem+SG":t=4:w=50
```

Cela revient à catégoriser *vrillée* comme un adjectif féminin singulier (*Adj :Fem+SG*). Toutefois, la façon dont le lemme et la catégorie grammaticale sont attribués pose certains problèmes. En effet, lorsque les utilisateurs de JeuxDeMots entrent des mots, il ne leur est pas demandé de préciser leur catégorie grammaticale (une fonction qui le permet apparaît de temps en temps mais elle semble peu utilisée). Les mots sont ainsi lemmatisés et catégorisés *a posteriori* à l'aide du dictionnaire de mots communs de l'ABU[14], le problème étant qu'il est impossible de

[14]http ://abu.cnam.fr/DICO/donner-dico-uncompress.html

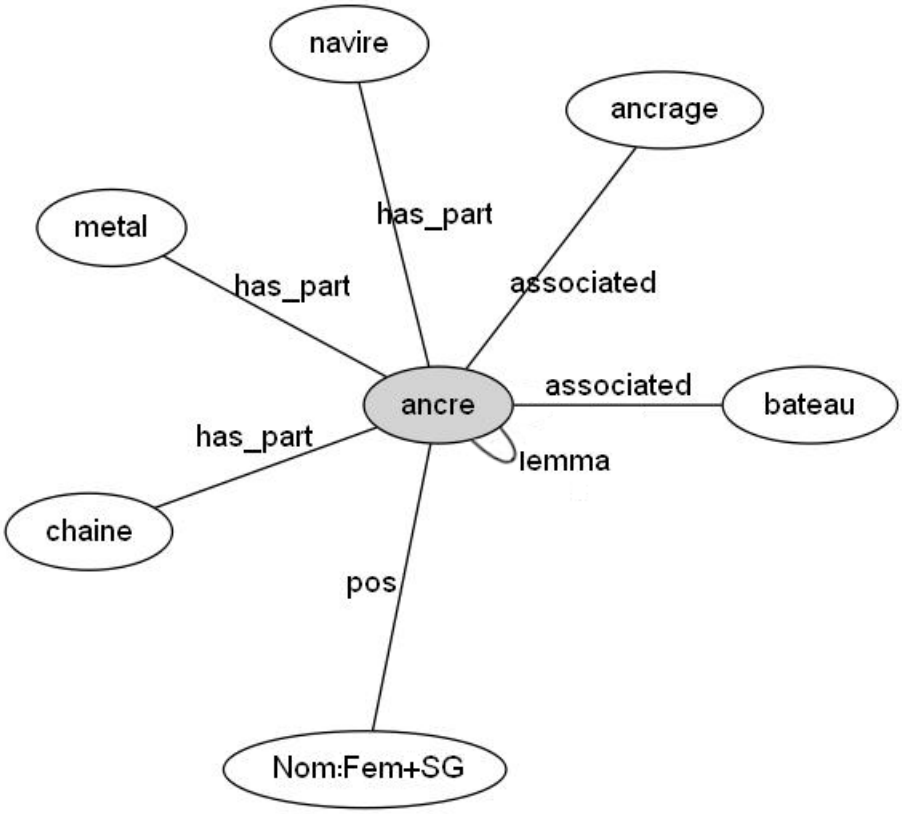

FIG. 3.3 – Représentation du nœud correspondant au mot *ancre* au sein du réseau JeuxDeMots.

déterminer avec certitude le lemme et la catégorie d'un mot de la sorte, puisque cette opération s'effectue *in abstracto*. Ainsi, les cas d'ambiguïté sont légion : par exemple, *restaurant* ou *volant* sont à la fois considérés comme des noms et comme des participes présents (et des adjectifs, pour *volant*). Et cela devient encore plus complexe dans les cas où les deux mots de la relation sont ambigus, comme dans *bus*/*voie* (*bus*/*voie* ? *boire*/*voie* ? *boire*/*voir* ? *boire*/*voie* ?). Nous verrons plus tard quelles ont été les modifications que nous avons dû apporter au réseau pour gérer ce problème.

Ainsi, si l'on s'en tient à ce que l'on a vu jusque-là, un mot du réseau (ou nœud de type *terme*) peut être rattaché à d'autres mots par des liens sémantiques ou un lien *lemme* et peut être rattaché à un nœud représentant une catégorie grammaticale par un lien *pos*[15], comme on peut le voir à la figure 3.3 où nous avons représenté la situation du mot *ancre* dans le réseau[16].

[15]Il existe d'autres types de relations que nous n'avons pas représentées puisqu'elles sont dédiées au fonctionnement du système.

[16]Nous avons rapporté les liens dans leur désignation originale : *associated* correspond à la relation IDÉE ASSOCIÉE, *has_part* à MÉRONYMIE, etc.

Le format de sortie retenu, dans un premier temps, a été le suivant[17] :

mot1	lemme1	cat1	mot2	lemme2	cat2	relation
paix	paix	NOM	violence	violence	NOM	ANTONYMIE
paix	paix	NOM	vivre en paix	vivre en paix	NOM	LOCUTION
palabrer	palabrer	VER	bavarder	bavarder	VER	IDÉE
palabrer	palabrer	VER	discuter	discuter	VER	IDÉE
palabrer	palabrer	VER	parler	parler	NOM	IDÉE
palais	palais	NOM	Angleterre	Angleterre	NOM	LIEU
palais	palais	NOM	avocat	avocat	NOM	LIEU

Ces deux représentations ne font apparaître aucun cas d'ambiguïté. Il est d'ailleurs assez étonnant de voir que *ancre*, en plus d'avoir été reconnu comme un nom, n'ait pas été identifié comme une forme fléchie du verbe *ancrer*, surtout quand on considère la propension qu'a le dictionnaire de l'ABU à considérer que certains noms sont également des flexions de verbes dont les chances d'apparaître dans le réseau sont proches de zéro : *voiturer* pour *voiture*, *musiquer* pour *musique*, *accolader* pour *accolade*, *harmonier* pour *harmonie*, *silhouetter* pour *silhouette*, *épiner* pour *épine*[18]... Comme on peut le constater à la figure 3.4, certains mots ont même été rattachés à trois lemmes différents. Or, dans nos deux autres ressources, les relations relient exclusivement des lemmes, ce qui implique que pour que la base JeuxDeMots puisse être exploitée, il faut impérativement parvenir à rattacher chaque mot à son lemme. Et étant donné que ce type de conflit est évidemment impossible à traiter au cas par cas, nous avons dû modifier le réseau en fonction d'une série de règles nécessairement généralistes : nous avons par exemple considéré que, dans les cas comme *bus*/*voie*, il était plus que probable que les lemmes soient *bus* et *voie*, et non *boire* et *voir*, étant donné que quand un verbe est soumis à JeuxDeMots, il l'est généralement à l'infinitif (on imagine mal, par exemple, dans quel contexte un mot comme *avions* pourrait avoir été ajouté en tant que forme fléchie de *avoir*). Certains cas d'ambiguïté portant sur la catégorie grammaticale sont beaucoup plus complexes (voire impossibles) à gérer. C'est notamment le cas des ambiguïtés nom/adjectif où le lemme est identique comme *clinique*, *désert*, *acide*, *calcaire*, *diésel*, etc. Il est ici impossible de se baser sur des indices formels pour lever l'ambiguïté. Toutefois, on peut se poser la question de la pertinence qu'il y a à essayer de désambiguïser ce genre de cas : étant donné que ces mots ont des sens relativement proches (voire identiques pour certains) qu'ils soient sous leur forme nominale ou adjectivale, pourquoi ne pas se passer simplement du critère de la catégorie grammaticale ? Cela pourrait

[17] Les catégories grammaticales ont été ramenées aux catégories *nom*, *verbe* et *adjectif*.

[18] La liste est malheureusement loin d'être exhaustive.

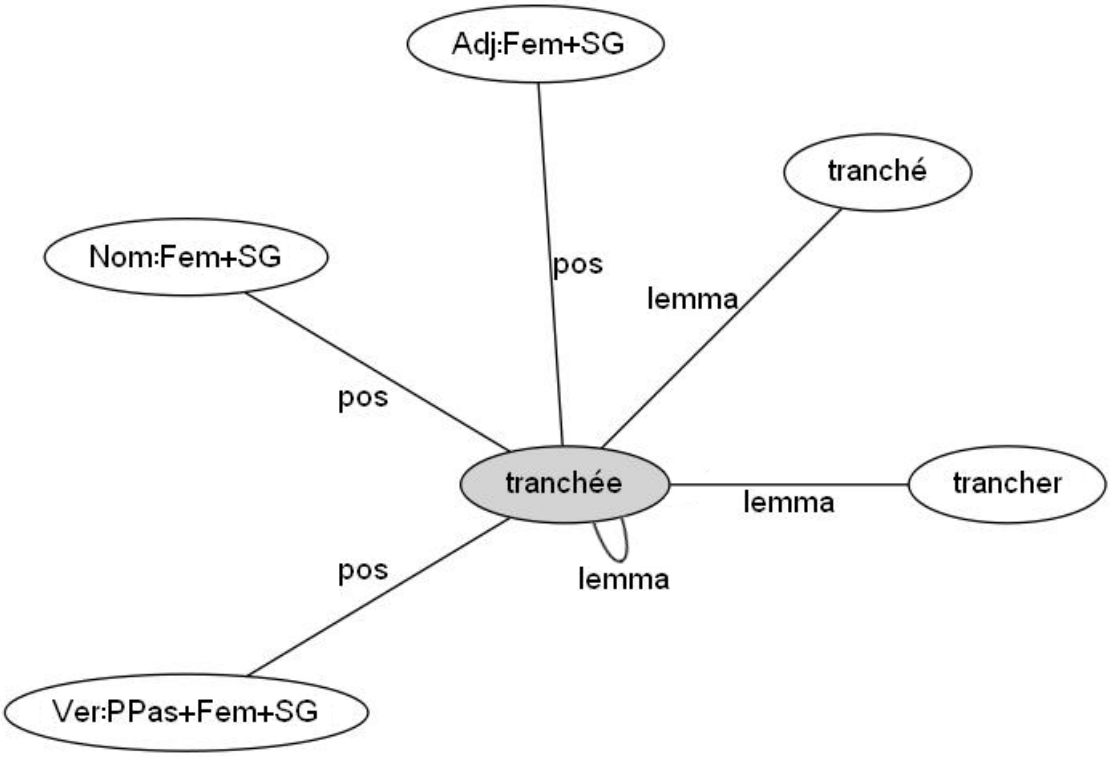

FIG. 3.4 – Représentation des relations de type *lemme* et *pos* liées au nœud correspondant au mot *tranchée* au sein du réseau JeuxDeMots.

permettre une interconnexion entre les mots reliés aux noms et ceux qui sont reliés aux adjectifs et augmenter ainsi le rappel du système qui utilisera la ressource. Cette alternative nous paraît pour l'heure beaucoup plus séduisante qu'une tentative de désambiguïsation qui, dans ce cas, sera forcément imparfaite. Nous reviendrons sur cette idée à la section 3.1.4.

La dernière modification que nous avons apportée au réseau a consisté en l'effacement de certaines relations au motif qu'elles étaient, à notre avis, trop susceptibles de générer du bruit : par exemple, pour la relation AGENT, les mots *personne*, *homme*, *femme*, *enfant*, *fille*, *garçon*, *individu* reviennent pour une très grande partie des verbes, alors que des relations plus pertinentes comme *marin*/*voguer* ou *médecin*/*anesthésier* sont beaucoup plus rares. Cela est dû au fait que *homme*, *femme*, *enfant*, etc. sont des agents potentiels de la grande majorité des verbes d'action et qu'un joueur qui entre ces réponses a de grandes chances de marquer un certain nombre de points (un nombre toutefois peu élevé puisqu'un système de pondération fait baisser la valeur des mots les plus fréquemment soumis). La relation MANIÈRE a également été retirée puisqu'elle relie des verbes à des adverbes, lesquels ne sont pas pris en compte dans notre système de repérage de la cohésion. La relation SENTIMENT s'est avérée trop vague pour pouvoir être conservée.

Au final, les catégories grammaticales n'ont pas été rétablies, et après quelques opérations de dédoublonnage et de désymmétrisation, la base JeuxDeMots ne compte plus que 129 458 relations[19] (alors qu'elle en comptait 180 840 avant le traitement[20]).

3.1.3 Dicosyn

Description

Étant donné que la synonymie est l'une des relations qui permettent de créer de la cohésion, il nous a paru logique de tenter une approche mobilisant un dictionnaire de synonymes *prêt à l'emploi* comme Dicosyn en marge de l'utilisation des voisins et de la base JeuxDeMots.

Développé au CRISCO (Université de Caen), il regroupe les synonymes présents dans sept dictionnaires classiques, à savoir le *Bailly*, le *Benac*, le *Du Chazaud*, le *Guizot*, le *Lafaye*, le *Larousse* et le *Robert*. Il compte environ 49 000 entrées pour 396 000 relations synonymiques et est consultable en ligne [21].

La façon relativement abrupte dont ont été fusionnés les dictionnaires nous pose toutefois quelques problèmes. Il n'a par exemple pas été possible de conserver de distinction entre les différentes acceptions des mots. Ainsi, pour le mot *bureau*, Dicosyn rapporte indifféremment des mots comme *service*, *administration*, *comptoir*, *table*, *secrétariat*, *commission*, *pupitre*, *burlingue* ou encore *écritoire*, dont on voit clairement qu'ils relèvent de plusieurs acceptions différentes (ce problème se pose également avec les autres ressources mais ces dernières ne permettaient pas de faire la distinction à l'origine, contrairement aux dictionnaires qui ont constitué Dicosyn).

De plus, contrairement aux voisins, cette ressource n'est pas spécialement adaptée aux textes sur lesquels nous allons travailler : certains articles de Wikipédia sont relativement spécialisés et font appel à des notions forcément absentes des dictionnaires généralistes sur lesquels est construit Dicosyn (cette critique pourrait également porter sur la base JeuxDeMots). Le repérage des relations qu'entretiennent ces mots est donc exclu d'office.

On pourrait également évoquer le fait que Wikipédia est une entité en constante évolution. Ainsi, il est possible d'y voir apparaître les évolutions de la langue les plus récentes, alors que Dicosyn est basé sur des dictionnaires dont certains commencent à se faire un peu vieillis-

[19] Hors relations système (comme la relation *pos*).

[20] Le réseau qui nous a servi de base de départ est celui du 18 novembre 2008. D'autres versions plus récentes sont disponibles à l'heure actuelle.

[21] http ://www.crisco.unicaen.fr/cgi-bin/cherches.cgi

sants[22]. Cette remarque ne constitue évidemment pas une critique, étant donné que, d'une part, cette démarche d'opérer un tant soit peu en diachronie constitue l'un des atouts de Dicosyn, et d'autre part, cette ressource est également construite sur des dictionnaires plus récents[23]. Et malgré le fait que Dicosyn soit entretenu et régulièrement mis à jour, les moyens mobilisés pour soutenir cette évolution sont évidemment incomparables avec la masse de contributeurs qui font croître Wikipédia. Il faudra ainsi envisager une certaine variation d'ordre diachronique entre le texte contenu dans les articles et les relations recensées dans Dicosyn.

Il est intéressant de noter que les relations de synonymie présentes dans Dicosyn empiètent assez souvent sur d'autres relations comme l'hypéronymie (on retrouve *félin* parmi les synonymes de *chat*, *véhicule* pour *voiture*), la méronymie (*couplet* et *refrain* pour *chanson*), voire la collocation (*disque/guitare*, même s'il s'agit probablement là d'une erreur). Ces irrégularités ne seront en aucun cas gênantes pour l'utilisation que nous comptons faire du dictionnaire, puisqu'ici encore, le fait que deux mots soient reliés nous intéressera plus que la nature précise de la relation qui les rapproche. Ces exemples ne sont là que pour rappeler qu'il faut relativiser l'appellation *dictionnaire de synonymes* en ce qui concerne Dicosyn, puisqu'un grand nombre des relations qu'il recense ne relèvent pas de la synonymie au sens strict.

Traitement

La ressource n'a subi quasiment aucun traitement si ce n'est la désymétrisation et le dédoublonnage de rigueur. Elle ne compte donc plus que 194 576 relations, qui se présentent sous la forme suivante :

```
mot1     cat1      mot2   cat2
------------------------------
bancal    A        éclopé   A
barboté   A        volé     A
barjo     A        cinglé   A
barjo     A        dingo    A
```

3.1.4 Amalgame des trois ressources

Nous disposons à présent de trois ressources lexicales que nous avons traitées à notre convenance et dont le nombre total de relations avoisine les 1,8 millions. Cela est dû au fait que la

[22]Le *Dictionnaire des synonymes de la langue française* de Lafaye date du XIX^{e}siècle.

[23]C'est du moins l'impression qui ressort de la consultation de Dicosyn, étant donné que nous n'avons aucune information quant aux éditions des autres dictionnaires utilisés.

base de voisins est présente dans son intégralité, puisque nous n'avons pas fixé de seuil (cf. section 3.1.1). Et étant donné qu'il serait difficile d'en définir un que l'on pourrait considérer comme satisfaisant du point de vue du rapport entre le taux de recouvrement avec Dicosyn et le nombre de voisins rapporté, nous avons préféré fixer un nombre de relations qui se situe aux alentours de 200 000, soit à peu près autant que pour Dicosyn, qui en contient 194 576. Cette équivalence laisse la porte ouverte à une éventuelle évaluation comparée. Les seuils qui permettent de rapporter un nombre de relations qui est le plus proche de 200 000 sont 0,14 pour le Jaccard (211 882 relations) et 0,2 pour le Lin (201 990). Sachant que pour ces deux seuils, le taux de recouvrement est de 3,85% pour le Jaccard et est de 4,16% pour le Lin, nous avons choisi de filtrer la base en fonction de cette dernière mesure.

Ainsi, pour récapituler, la base dont nous disposons est composée de :

- 201 990 relations provenant des voisins de Wikipédia,
- 129 458 relations du réseau JeuxDeMots,
- 194 576 relations de Dicosyn.

Soit un total de 526 024.

Les premières projections des relations que nous avons effectuées sur des textes étiquetés nous ont incité à opérer une modification assez radicale de la base, que nous avons évoquée précédemment (section 3.1.2). En effet, face à la difficulté qu'il y aurait à reconstituer toutes les catégories grammaticales correctes des entrées de JeuxDeMots, et devant la quantité d'erreurs d'étiquetage que nous avons constatées (nos textes ont été traités par TreeTagger), nous avons envisagé la suppression des champs contenant les catégories grammaticales. De plus, dans de nombreux cas, la catégorie grammaticale des mots restreint l'identification de liens de cohésion, or, nous avons vu avec [Halliday et Hasan, 1976] que les liens cohésifs sont avant tout sémantiques et que, notamment dans le cas des collocations, le critère de la catégorie grammaticale n'entre pas en ligne de compte. La première réaction que l'on peut avoir face à une telle démarche est une appréhension face à l'ambiguïté qu'elle va forcément entraîner. Nous faisons ici le pari que cette ambiguïté peut nous être profitable.

En effet, le critère de la catégorie grammaticale est déterminant quand il s'agit de discriminer des mots dont les lemmes ont la même forme mais appartiennent à des catégories grammaticales différentes. Mais, dans le cas qui nous concerne, est-il vraiment dans notre intérêt de conserver une frontière entre des paires de mots obtenues par conversion comme *manger* en tant que verbe et *manger* en tant que nom, ou entre *aéronautique* en tant qu'adjectif et en tant que nom ? Cette question s'est posée suite aux premières projections de voisins que nous avons effectuées sur un article de Wikipédia (l'article *Narbonne*), le but étant de relier les mots des titres de section avec leur contenu. Parmi les 40 titres de l'article, tous niveaux confondus, seuls

trois mots apparaissent dans plus d'une catégorie grammaticale dans notre base de voisins, à savoir les adjectifs des titres suivants :

- *4.1. Tendances politiques*
- *10.1. Narbonne dans la littérature classique*
- *10.2. Narbonne dans la littérature contemporaine*

L'ambiguïté réside dans le fait qu'en plus d'avoir été annotés -correctement- comme des adjectifs, ces trois mots existent aussi sous la forme de noms. Le fait de s'affranchir des catégories grammaticales en prenant en compte les relations des formes nominales de chacun de ces adjectifs a permis d'améliorer les résultats dans un cas sur trois : pour *classique* et *contemporain* les résultats sont restés les mêmes alors que pour *politique*, le nombre de liens a légèrement augmenté. Les liens tissés entre les deux formes de *politique* et le texte de la section (quelques lignes) se répartissent de la façon suivante :

POLITIQUE (A)	**élection**, français, national, royal, **socialiste**, **ville**
POLITIQUE (N)	député, **élection**, **socialiste**, **ville**, vote

La prise en compte des voisins de la forme nominale de *politique* a permis de créer deux liens supplémentaires avec *député* et *vote* (dont la pertinence est indiscutable), ce qui, au vu du nombre total de liens, n'est pas négligeable. De plus, le fait que les mots *élection*, *socialiste*, et *ville* aient été retrouvés par les deux formes de *politique* vient soutenir l'idée selon laquelle ces deux mots sont tellement proches qu'il pourrait être envisageable de fusionner leurs voisins. La question de savoir dans quelle mesure les cas d'ambiguïté peuvent être ignorés de la sorte sans risquer de générer du bruit est assez délicate, mais étant donné que la conversion est un procédé morphologique courant et que les mots qu'elle permet de générer semblent graviter autour du même noyau de sens que les mots à partir desquels ils sont formés, nous avons bon espoir quant à l'effet positif qu'aura le retrait des catégories grammaticales de la base sur le rappel du système de détection des liens cohésifs que nous mettons en place, puisqu'elle permet d'augmenter le nombre de voisins disponibles pour un pan entier des mots du lexique.

Au final, le fait de retirer les relations grammaticales entraîne l'apparition de doublons que nous avons effacés, ce qui fait baisser la base de voisins de 2543 relations, et Dicosyn de 3831 relations. La base finale, une fois dédoublonnée, ne compte quant à elle plus que 475 798 relations.

3.2 Le texte de travail

Le choix du type de texte à étudier s'est fait de façon assez naturelle : étant donné que les voisins ont été calculés sur Wikipédia, il est apparu logique de travailler sur un article issu de

cette encyclopédie.

> We believe that an automatic cohesion analysis of text, particularly lexical cohesion, is possible if appropriate linguistic knowledge is provided, i.e. the semantics and the grammar of the language in which the text is written.
>
> ([Benbrahim et Ahmad, 1994] p. 43)

Ainsi, l'hétérogénéité de notre ressource finale peut se caractériser par le fait que les relations qui proviennent des voisins auront des chances de se révéler particulièrement pertinentes sur ce type de texte en particulier, alors que Dicosyn et la base JeuxDeMots ont des visées plus généralistes. Le mode de rédaction *collaboratif* des articles issus de Wikipédia présente également un intérêt dans le cadre d'une étude portant sur la cohésion : un texte rédigé par plusieurs auteurs possède-t-il des propriétés cohésives différentes des textes rédigés par un seul et même auteur ? Dans le cadre de notre démarche générale, qui consiste à repérer et analyser les phénomènes cohésifs, nous avons choisi de travailler sur un texte présentant une certaine variété thématique. En effet, une première expérimentation du système a été menée sur un article biographique dédié à William T. Sherman[24], et il s'est avéré que le fait que la quasi-totalité du texte porte sur le même thème, celui de la guerre, rend la détection des liens interphrastiques relativement difficile puisqu'elle exige alors un niveau de grain qui, nous le verrons, n'est pas à la portée de notre ressource. Ainsi, en choisissant un article un tant soit peu hétérogène, nous misons sur le fait que des thématiques différentes et fortement marquées faciliteront le repérage des relations de cohésion, ce qui permettra d'éviter que "l'interprétation des appariemements [ne se révèle] aléatoire, c'est-à-dire fondée sur des "arrière-plans" n'appartenant pas nécessairement au texte analysé" ([Legallois, 2004] p. 5).

Les articles portant sur des villes satisfont à ce critère. Un article de Wikipédia consacré à une ville est en effet construit autour d'un ensemble de sections relativement stable qui englobe les différents aspects d'une ville qu'il est pertinent de décrire dans un article encyclopédique, à savoir sa situation géographique, démographique, son histoire, sa vie culturelle et politique, etc. Ces sections sont divisées en un ensemble de sous-sections dont le nombre dépend de la taille de la ville : la section *Géographie* peut être subdivisée en sous-sections portant sur la topographie, l'hydrographie ou le climat de la ville, alors que la section *Histoire* peut présenter son évolution au cours de différentes périodes comme le Moyen-Âge, la Renaissance, etc. Toutefois, si ces catégories restent relativement similaires d'un article à l'autre, leur agencement peut varier : en effet, pourquoi choisir d'aborder le sujet de la politique du logement d'une ville après une description de son climat et avant la description de ses équipements culturels ? Il n'y a pas ici d'organisation réellement *motivée*, comme c'est le cas par exemple pour les articles

[24]http ://fr.wikipedia.org/wiki/William_Tecumseh_Sherman

biographiques, où les évènement sont décrits dans l'ordre dans lequel il se déroulent (seules les sous-sections présentant l'historique d'une ville sont organisées chronologiquement). De plus, le récit d'une successions d'évènements fait que l'article biographique se rapproche du style de la narration, or, Hoey est assez dubitatif quant à l'efficacité de sa méthode sur des textes narratifs. En choisissant un article de type non-narratif, nous nous assurons un fonctionnement optimal de la théorie développée dans le chapitre suivant.

La taille de l'article à étudier a également guidé notre choix : l'article doit être assez volumineux pour que son analyse présente un intérêt du point de vue discursif (notamment au niveau du repérage des liens de longue portée) mais également assez court pour pouvoir être traité manuellement, ce qui, nous le verrons, va s'avérer indispensable.

Ainsi, le texte sur lequel nous avons choisi de travailler est l'article consacré à la ville de Narbonne[25]. Sous sa forme originale, il comptait 308 lignes, mais au regard de l'approche développée par Hoey, les dernières catégories de l'article (*14. Voir aussi*, *15. Bibliographie*, *16. Références* et *17. Liens externes*) ne présentaient aucun intérêt vu qu'elles correspondent à des listes d'adresses de sites internet et de noms d'œuvres, et ont donc été retirées du texte, qui ne compte plus que 265 lignes (cette dernière version du texte a été reportée à l'annexe A.1).

L'article a enfin été étiqueté par TreeTagger. Les annotations s'étant révélées quelquefois inexactes, il a fallu opérer une vérification manuelle. Cette étape nous a fait prendre conscience du fait que la segmentation opérée par TreeTagger risquait de provoquer l'apparition de liens de cohésion erronés, notamment dans le cas de syntagmes prépositionnels comme *à la place de*, *en raison de* ou *à la tête de*, où les noms ont été étiquetés isolément. La question, plus problématique, des formes figées comme *homme d'affaires*, *station balnéaire* ou *faire faillite* sera débattue dans la partie consacrée à l'expérience : est-il plus avantageux de réannoter ces locutions de façon correcte ou de laisser leurs éléments séparés, sachant que dans ce dernier cas, le nombre de liens de cohésion générés sera plus important ?

[25]http ://fr.wikipedia.org/wiki/Narbonne (version du 29/05/09)

Chapitre 4

Relations lexicales et repérage de liens interphrastiques : la méthode Hoey

Comme nous l'avons indiqué à plusieurs reprises tout au long de ce mémoire, notre étude s'articule principalement autour des concepts de *cohésion lexicale* et d'*organisation discursive*, qui relèvent chacun de deux niveaux différents, à savoir respectivement le niveau textuel et le niveau du discours. Le présent chapitre fait écho au début de ce mémoire où nous avons mis en opposition ces deux dimensions tout en présentant l'objectif de cette étude, qui est de mettre au jour des phénomènes organisationnels en s'appuyant sur les liens de cohésion qui parcourent les textes.

Le point de vue de [Halliday et Hasan, 1976] concernant la relation entre cohésion et structures discursives est particulièrement tranché : "cohesion [. . .] is not a structural relation" (p. 6). Il convient toutefois de développer ces propos : si effectivement la cohésion n'est pas une relation structurante, il serait faux de penser que cohésion et structure relèvent de domaines hermétiquement disjoints. Dans le sens où elles sont des *traces* de la cohérence (qui, comme nous l'avons vu, relève du discours), les marques de cohésion comme les chaînes lexicales peuvent permettre de remonter jusqu'aux structures organisationnelles. Cette approche, que nous pouvons qualifier d'*ascendante* (ou *bottom-up*), prend appui sur la propriété suivante des structures discursives :

> When a chunk of text forms a unity whithin a discourse, there is a tendency for related words to be used. It follows that if lexical chains can be determined, they will tend to indicate the structure of the text. ([Morris et Hirst, 1991] p. 24)

Il apparaît que les éléments des structures discursives ont pour caractéristique de présenter des marques de cohésion, ce qui se manifeste dans le texte par une concentration de mots sémantiquement liés. Ainsi, si la cohésion n'est pas intrinsèquement structurante, elle apparaît

comme une manifestation de surface des phénomènes organisationnels sous-jacents, et à ce titre elle permet de détecter des structures discursives à travers la configuration lexicale d'un texte.

À ce stade de l'étude, il est toutefois un peu prématuré de considérer comme des unités discursives toute concentration de mots sémantiquement liés, dans le sens où cela impliquerait de se placer dans un cadre théorique particulier où serait défini le concept même d'unité discursive, ce que nous n'avons pas fait. On peut tout au plus considérer ces concentrations comme des unités thématiques susceptibles de jouer un rôle au niveau organisationnel.

Ainsi, [Marcu, 2000] mobilise la cohésion lexicale dans le cadre de la RST afin de distinguer certains types de relations : alors que les relations ELABORATION et BACKGROUND relient deux segments qui portent sur le même sujet, JOINT et LIST relient deux ensembles qui portent sur des sujets différents (du point de vue thématique). La notion d'unité thématique se charge ici d'une valeur discursive du fait qu'elle est envisagée dans un cadre défini *a priori* (celui de la RST)[1].

À l'inverse, une approche complètement ascendante a été mise en œuvre dans [Hoey, 1991], où le lexique joue un rôle particulièrement important puisqu'il est vu comme étant au cœur des phénomènes d'organisation discursive[2]. La méthode que M. Hoey développe dans *Patterns of lexis in text* se base sur le fait que l'étude de la répartition des phrases reliées par des liens de cohésion de part et d'autre d'un texte permettrait d'en faire apparaître les dynamiques organisationnelles (notamment au travers des concepts de *topic opening* et *topic closing sentences*). Un corrolaire de cette théorie est que le degré de *centralité* d'une phrase au sein d'un texte est proportionnelle au nombre de phrases auxquelles elle est reliée. En s'appuyant sur la base lexicale présentée au chapitre précédent, nous avons automatisé cette méthode d'analyse et traité le texte décrit en 3.2. Les résultats de cette expérience sont présentés dans la deuxième partie de ce chapitre, la première étant consacrée à la présentation détaillée de la méthode mise en place par Hoey ainsi qu'aux problèmes que nous avons rencontrés lors de son automatisation.

[1] Étant donné que cette approche s'appuie principalement sur des *cue phrases*, nous ne ferons pas ici de description plus poussée de l'algorithme de Marcu.

[2] Nous éviterons de parler de *structure* dans la suite de ce chapitre, étant donné que Hoey réfute cette notion et la remplace par celle de *pattern of organization* (p. 29).

4.1 Méthodologie

4.1.1 Le choix du texte à analyser

Dans *Patterns of lexis in text*, Hoey analyse en détail deux textes. Le premier, que nous reprenons ci-dessous, est un très court texte extrait d'un magazine (il ne compte que cinq phrases[3] plus un titre, on peut le considérer comme une *brève*) qui sert uniquement à illustrer la méthode. Le second est l'introduction d'un ouvrage de philosophie politique de quarante phrases (plus un titre).

Le choix de ces deux textes a été principalement motivé par deux critères, à savoir celui de la taille et celui du genre. Les raisons qui font que ces deux textes sont relativement courts diffèrent. Le fait que le premier soit extrêmement bref est simplement dû aux besoins de la démonstration, alors que la limite de quarante lignes que s'est fixée Hoey en ce qui concerne le deuxième est dûe à la nature même de la procédure : cette dernière nécessite en effet de comparer les phrases une à une, ce qui rend l'analyse à grande échelle difficilement envisageable sans l'aide de l'outil informatique (Hoey, qui a procédé à une analyse entièrement manuelle, a d'ailleurs divisé ce dernier texte en deux). Or, si l'on considère que les liens de cohésion permettent de rendre compte de l'organisation de textes aussi volumineux que des livres entiers (p. 24), l'automatisation de la méthode de Hoey apparaît comme inévitable. Malgré cela, et bien que le recours à l'outil informatique soit mentionné à plusieurs reprises comme une alternative crédible au repérage manuel des relations dans [Hoey, 1991] (p. 74, 77) ainsi que dans d'autres travaux reprenant son analyse comme [Legallois, 2004] (p. 9), Hoey reste assez évasif sur les conditions dans lesquelles pourrait s'envisager l'automatisation de sa méthode (ce qui est parfaitement assumé, cf. p. 271). [Benbrahim et Ahmad, 1994], qui se sont appuyés sur ses travaux pour mettre au point un système de résumé automatique, lui reprochent notamment de ne pas aborder le sujet du type de ressource lexicale (dictionnaire de synonyme, thésaurus, etc.) que l'automatisation amènerait immanquablement à mobiliser (p. 42).

Le deuxième critère qui a motivé le choix de ces deux textes a été celui du genre. Avant de se lancer dans une analyse *à la Hoey*, il est important d'avoir à l'esprit que cette méthode n'est pas aussi efficace sur tous les types de textes. En effet, selon Hoey, le nombre de phrases reliées par des relations de cohésion serait plus faible dans les textes narratifs que dans les textes non narratifs, et leur rapprochement ne serait que peu pertinent au niveau discursif (p. 188). La raison en est que chaque phrase d'un texte narratif s'inscrit dans un cadre spatio-temporel particulier, ce qui empêche son rapprochement avec des phrases situées dans d'autres parties du récit.

[3] La phrase est ici l'unité d'analyse, ce qui rend le critère du nombre de mots non pertinent dans ce contexte.

Afin d'illustrer la démarche de M. Hoey, nous reprenons le texte donné dans le chapitre 2 de *Patterns of lexis in text* :

DRUG-CRAZED GRIZZLIES

1. A drug known to produce violent reactions in humans has been used for sedating grizzly bears Ursus arctos in Montana, USA, according to a report in The New York Times.
2. After one bear, known to be a peaceable animal, killed and ate a camper in an unprovoked attack, scientists discovered it had been tranquilized 11 times with phencyclidine, or "angel dust", which causes hallucinations and sometimes gives the user an irrational feeling of destructive power.
3. Many wild bears have become "garbage junkies", feeding from dumps around human developments.
4. To avoid potentially dangerous clashes between them and humans, scientists are trying to rehabilitate the animals by drugging them and releasing them in uninhabited areas.
5. Although some biologists deny that the mind-altering drug was responsible for uncharacteristic behaviour of this particular bear, no research has been done into the effects of giving grizzly bears or other mammals repeated doses of phencyclidine.

4.1.2 Le repérage des relations lexicales

La première étape consiste à essayer de relier chaque nom, verbe et adjectif aux autres noms, verbes et adjectifs du texte par des relations de *répétition*. Dans la terminologie qu'adopte Hoey, la répétition englobe la *répétition simple* (mot identique ou fléchi), la *répétition complexe* (les dérivés comme *drug/drugging*), la *paraphrase simple* (synonymie), la *paraphrase complexe*, qui comprend aussi bien l'antonymie que certains types de relations indirectes (dans le sens où si il y a une relation entre *writer* et *writings* et entre *writer* et *author*, alors il y a une relation entre *author* et *writings*), la relation d'hypéronymie et les relations de substitutions faisant appel à des éléments grammaticaux comme l'anaphore. Il est à noter que Hoey rejette les relations de collocation (au sens de Halliday & Hasan) comme *sickness/doctor* et *carol/Christmas* au motif qu'elles ne sont pas facilement maîtrisables ("readily controllable", p. 64). En ce qui nous concerne, et comme nous l'avons vu dans le chapitre précédent, le repérage de ce type

de relation est justement l'un des buts que nous visons par l'utilisation d'une ressource aussi hétérogène que les voisins de Wikipédia.

Lors de cette étape, qui consiste, basiquement, à relier des mots, se pose une question aussi ancienne que la linguistique : qu'est-ce qu'un mot ? Hoey mentionne ce problème (p. 271) lorsqu'il se retrouve confronté aux deux occurrences de *grizzly bears*, aux lignes 1 et 5 : faut-il considérer que l'on a affaire à une seule unité polylexicale ou à deux unités qu'il convient de traiter séparément ? Hoey laisse cette question en suspens (p. 245) et choisit la deuxième solution. Dans le cadre de notre expérience, nous nous en sommes, dans un premier temps, remis à la segmentation opérée par TreeTagger, qui considère basiquement comme un mot toute chaîne de caractères limitée par une espace, une virgule, un point, etc. Nous verrons que les conséquences de ce choix sur notre système de repérage automatique des relations de cohésion varient selon le degré d'opacité des séquences figées que l'on peut recontrer au fil du texte : pour un syntagme comme *grizzly bears*, le fait de considérer une ou deux unités ne risquerait pas de générer trop de bruit puisque *grizzly* et *bear* sont très proches sémantiquement (ils entretiennent une relation d'hypéronymie) et les relations qui les relient au reste des mots du texte ont de grandes chances d'être semblables, ce qui n'est évidemment pas le cas pour des syntagmes nominaux comme *bateau à vapeur*, ou verbaux comme *semer la confusion*, *être en mauvaise posture* ou *prendre en compte*.

Un autre problème lié à l'automatisation est celui de la polysémie : par exemple, le mot *théâtre*, qui revient à trois reprises dans le premier texte que nous avions étudié (l'article *Sherman*), se retrouve relié à des mots comme *rôle*, *personnage*, *art* et *texte*, alors qu'il est dans la plupart des cas utilisé pour désigner le champ de bataille (le *théâtre d'opérations*) et non l'art du théâtre ou le bâtiment dans lequel il se pratique. Ainsi, alors que notre système ne fait pas la différence entre ces différentes acceptions, un humain n'aurait aucune difficulté à les discerner parmi les exemples suivants :

- <s=6> En 1864, Sherman succède à Grant à la tête de l'Armée de l'Union sur le **théâtre** occidental de la guerre de Sécession.
- <s=177> [...] Grant nomme Sherman pour lui succéder à la tête de la Division militaire du Mississippi, qui comprend le commandement des forces du **théâtre** occidental de la guerre.
- <s=320> Il se consacre au **théâtre** et à la peinture en amateur [...]

[Benbrahim et Ahmad, 1994] suggèrent d'opérer cette désambiguïsation à l'aide d'un calcul de collocation (au sens premier) qui leur permettrait de rattacher chaque acception d'un mot polysémique à son entrée dans le thésaurus qu'ils utilisent. Cela suppose évidemment que la distinction soit présente dans la ressource utilisée, ce qui n'est pas le cas avec la nôtre.

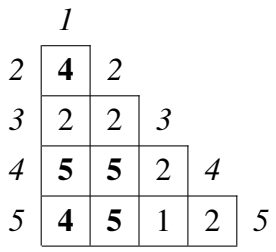

FIG. 4.1 – Matrice obtenue après analyse du texte *Drug crazed grizzlies*.

Comme nous l'avons indiqué précédemment, l'unité d'analyse de la méthode de Hoey est la phrase : le repérage des relations de cohésion n'est qu'une étape de la démarche générale qui consiste à créer des liens[4] entre elles (cela pose plusieurs problèmes sur lesquels nous reviendrons plus tard, notamment le fait que les phrases les plus courtes ont moins de chance d'être reliées aux autres). Ainsi, pour la première phrase, on obtient les rattachements suivants (nous reprenons pour cet exemple la terminologie de Hoey) :

- drug : drugging(phr. 4, répétition complexe), drug(phr. 5, répétition simple),
- produce : causes(phr. 2, paraphrase), responsible for(phr. 5, paraphrase)),
- humans : human(phr. 3 ; répétition complexe[5]), humans(phr. 4 ; répétition simple),
- used : user(phr. 2, répétition complexe),
- sedating : tranquilized(phr. 2, paraphrase), drugging(phr. 4, paraphrase),
- grizzly : grizzly(phr. 5, répétition simple),
- bears : bear(phr. 2, répétition simple), bears(phr. 3, répétition simple), animals(phr. 4, hypéronymie), them(phr. 4, anaphore), bears(phr. 5, répétition simple).

N.B. Il est à noter que Hoey ne marque pas les relations *human*/*scientists*, et *human*/*biologists* au motif que *scientists* et *biologists* ne font pas que répéter *human* mais y apportent une information supplémentaire.

4.1.3 La mise en relation des phrases

Ainsi, la phrase 1 est reliée à la phrase 2 par 4 relations, à la phrase 3 par 2 relations, à la phrase 4 par 5 relations et à la phrase 5 par 4 relations. La démarche est répétée pour chacune des phrases. Les résultats pour le texte peuvent se représenter sous la forme d'une matrice (se reporter à la figure 4.1 pour la matrice obtenue après analyse du texte *Drug-crazed grizzlies*).

[4]Nous verrons qu'un lien (*bond*) entre deux phrases ne se forme que si elles sont reliées par un certain nombre de relations de cohésion (*links*).

[5]Le nom *human* est relié à sa dérivation adjectivale.

Pour qu'un lien puisse se former entre deux phrases, il faut que ces deux phrases soient reliées par au moins trois relations (sachant que même si un mot d'une phrase A est relié à plusieurs mots d'une phrase B, ou inversement, on ne compte qu'une seule relation). Ce nombre, défini de façon intuitive, correspond au nombre minimal de relations à partir duquel on peut considérer que l'association de deux phrases est pertinente du point de vue discursif. Ce seuil pose toutefois plusieurs problèmes, à commencer par le fait qu'il est fixé de façon complètement subjective. De plus, il exclut de l'analyse tout un ensemble de phrases : un seuil de 3 ignore toute phrase simple constituée d'un nom et d'un verbe, par exemple. Hoey reconnaît que cette limite peut poser problème pour certains genres qui privilégient les phrases courtes comme les éditoriaux, mais considère que les liens obtenus par moins de 3 relations ne peuvent pas être pris en compte à cause de la probabilité que ces relations ne soient pas vraiment signifactives au plan organisationnel (p. 190). Toutefois, si le nombre minimum de relations ne peut pas descendre en dessous de 3, il peut être augmenté dans le cas où l'analyse porte sur des textes particulièrement redondants comme les textes juridiques, pour lesquels le seuil peut monter jusqu'à 12 (p. 92).

Ainsi, la phrase 1 sera reliée aux phrases 2, 4 et 5, mais pas à la 3 puisque le seuil de pertinence n'est pas atteint. Une fois ce seuil appliqué à toutes les phrases du texte, les liens peuvent être modélisés sous la forme d'un réseau, la première phrase du texte se trouvant tout en haut, la dernière tout en bas (cf. figure 4.2).

4.1.4 Interprétation du réseau obtenu

La centralité

Comme on a pu le voir avec le texte *Drug crazed grizzlies*, et comme on peut le constater sur le réseau qu'a permis de produire son analyse, certaines phrases possèdent plus de liens avec le reste du texte que d'autres. Partant de ce constat, Hoey émet l'hypothèse que le nombre de liens auxquels est rattachée une phrase peut servir d'indicateur de son importance (sa *centralité*) au sein du texte, et que cette propriété peut être exploitée pour obtenir un résumé du texte analysé. Il distingue ainsi deux types de phrases : celles qui ne sont rattachées à aucune autre, les phrases *marginales*, et celles qui présentent un nombre de liens remarquablement élevé par rapport aux autres, qui sont dites *centrales*.

La première méthode de résumé s'appuie sur les phrases marginales comme la phrase 3 du texte *Drug crazed grizzlies*, “Many wild bears have become "garbage junkies", feeding from dumps around human developments”. Selon Hoey, cette phrase ne participe pas directement à

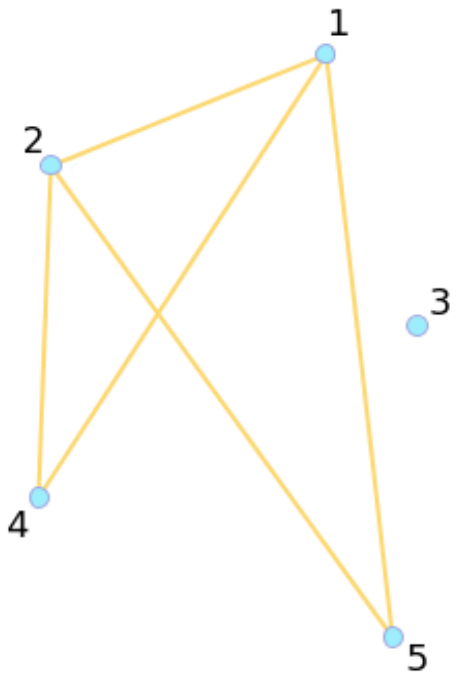

FIG. 4.2 – Réseau obtenu après analyse du texte *Drug crazed grizzlies*.

la construction du thème principal, et l'on pourrait dire qu'elle a plus un but illustratif. Pour essayer de démontrer que les phrases marginales ont peu d'importance par rapport au reste du texte, Hoey pratique le test de l'effacement : si la phrase peut être effacée sans que cela n'affecte l'interprétation du texte, son contenu est considéré comme secondaire et par conséquent elle n'a pas sa place dans un résumé. Et il s'avère en effet que la phrase 3 du texte *Drug crazed grizzlies* pourrait être supprimée. Ce même test a été effectué sur les huit phrases marginales repérées après analyse de l'introduction de l'ouvrage de philosophie politique dont il était question plus haut, et sept d'entre elles ont pu être effacées sans que cela ne gêne l'interprétation générale. Cette suppression a permis d'obtenir un résumé équivalent aux 5/6^{e} du texte original, ce qui reste très peu satisfaisant (le taux de phrases marginales peut toutefois atteindre plus de 50% dans certains types de textes).

À cette approche *négative* s'oppose une approche *positive*, qui consiste non plus à effacer les phrases les moins pertinentes mais au contraire à ne conserver que les phrases centrales, qui sont censées avoir le plus de signification au vu de l'ensemble du texte. Sur l'ensemble du texte d'introduction de l'ouvrage de philosophie politique, cela correspond à quinze phrases, lesquelles constituent, de l'avis de l'auteur, un bon résumé dont le volume est d'un tiers du texte original.

Dans les deux cas, la théorie selon laquelle l'importance d'une phrase au sein d'un texte dé-

pend directement du nombre de liens auxquels elle est rattachée semble être confirmée puisque les résumés produits par l'effacement des phrases marginales et l'extraction des phrases centrales semblent corects. Toutefois, l'évaluation de la qualité des résumés obtenus reste éminemment subjective puisqu'elle ne dépend ici que du jugement du linguiste. Il est à noter que pour juger de la qualité de résumés d'articles scientifiques fournis par un système reprenant la méthode de Hoey, [Benbrahim, 1996] a mis au point un protocole d'évaluation impliquant des spécialistes des domaines desquels les textes résumés dépendaient, ces derniers ayant pour tâche de remplir un questionnaire portant sur la compréhensibilité du résumé et la pertinence des informations qui y étaient rapportées (p. 107).

Les ouvertures et clôtures thématiques

Après avoir opéré une distinction entre les phrases *marginales* et les phrases *centrales* basée sur le nombre total de liens auxquels une phrase est rattachée, Hoey distingue deux autres types de phrases : les *topic-opening* et les *topic-closing sentences* (p.118), que nous désignerons par la suite *ouvertures* et *clôtures thématiques*. Les premières se distinguent par le fait qu'elles sont fortement rattachées à des phrases qui les suivent, les secondes par le fait qu'elles sont fortement rattachées à des phrases qui les précèdent. Ainsi, une phrase qui est très liée aux phrases qui la suivent est considérée comme une ouverture vers un nouveau sujet, alors qu'une phrase liée à celles qui la précèdent a un rôle de clôture.

Cette nouvelle façon d'envisager les liens entre phrases se modélise de la façon suivante : une phrase x se définit par y le nombre de phrases auxquelles elle est reliée parmi les phrases qui la précèdent, et z le nombre de phrases auxquelles elle est reliée parmi les phrases qui la suivent. Soit la représentation suivante : x(y,z). Par exemple, la première phrase du texte *Drug crazed grizzlies* n'est, forcément, précédée d'aucune autre, et est reliée à trois phrases parmi l'ensemble des phrases qui la suivent. Elle a donc la représentation suivante : 1(-,3). Les coordonnées de l'ensemble des phrases de ce texte sont représentées dans le tableau 4.1. Toutefois, comme le nombre de phrases que contient ce court texte est bien trop faible pour que l'on puisse espérer y repérer des phénomènes pertinents, nous avons rapporté au tableau 4.2 les coordonnées de chacune des quarante phrases du texte d'introduction de l'ouvrage de philosophie politique qu'à analysé Hoey (p. 104).

Si effectivement les phrases qui sont particulièrement liées avec celles qui les suivent ou qui les précèdent ont des rôles d'ouverture et clôture thématiques, alors il serait logique de rencontrer ces phrases en début et en fin de texte, puisqu'elles correspondent typiquement à des phrases d'introduction et de conclusion. Et il s'avère en effet que l'on retrouve ces deux profils

1(-,3)
2(1,2)
3(0,0)
4(2,0)
5(2,-)

TAB. 4.1 – Coordonnées des phrases du texte *Drug crazed grizzlies*.

1(-,15)	11(2,0)	21(4,4)	31(0,0)
2(1,0)	12(2,4)	22(1,1)	32(0,1)
3(0,2)	13(0,0)	23(5,6)	33(0,1)
4(1,6)	14(4,1)	24(5,0)	34(5,0)
5(0,0)	15(0,1)	25(1,2)	35(1,0)
6(1,0)	16(2,1)	26(6,4)	36(2,0)
7(2,1)	17(1,6)	27(0,0)	37(1,1)
8(2,0)	18(0,0)	28(8,3)	38(11,1)
9(0,0)	19(0,5)	29(0,1)	39(5,0)
10(0,0)	20(1,9)	30(0,0)	40(2,-)

TAB. 4.2 – Coordonnées des phrases du texte d'introduction de l'ouvrage de philosophie politique analysé par Hoey.

au début et à la fin du texte modélisé dans le tableau 4.2 :

– parmi toutes les phrases du texte, la première (1(-,15)) est celle qui possède le plus de liens (ces derniers étant forcément orientés dans la même direction) :

> What is attempted in the following volume is to present to the reader a series of actual excerpts from the writings of the greatest political theorists of the past ; selected and arranged so as to show the mutual coherence of various parts of an author's thought and his historical relation to his predecessors or successors ; and accompagnied by introductory notes and intervening comments designed to assist the understanding of the meaning and importance of the doctrine quoted.

Il apparaît clairement qu'effectivement, le rôle de cette phrase est de présenter le contenu général de l'ouvrage et il est donc logique qu'elle contienne un certain nombre d'éléments auxquels viendront se référer les phrases suivantes. On peut noter que ce rôle introductif est ici explicité dans le premier segment de la phrase : "What is attempted in the following volume is to [...]" (ce genre d'indice n'est toutefois pas exploité dans la méthode de Hoey).

– les phrases 38(11,1) et 39(5,0), soit l'antépénultième et la pénultième, ont clairement le profil de phrases conclusives. Elles correspondent à une synthèse et reprennent donc un certain nombre d'éléments dont il a été question tout le long du texte :

> He will not expect the works which follow to provide the same kind of instruction for the political student as his medical text-books provide for the medical student. If he did he would certainly confine his attention to the most up-to-date, and would no more think of learning his art from Aristotle than the medical student thinks of learning his from the pages of Galen.

En marge de ces zones immédiatement identifiables, on s'aperçoit que ces profils se retrouvent également au cœur du texte. Il serait en effet normal pour un texte introductif d'amener de façon consécutive les différentes problématiques que va aborder l'ouvrage, et par conséquent de voir cette succession se manifester par une alternance de clôtures et d'ouvertures thématiques. Prenons la phrase 19(0,5) : "A work of politics, it would have been said, is the handbook of an art, the art of governing". Selon Hoey, cette affirmation implique un développement, lequel serait appuyé par la phrase suivante, 20(1,9), qui présente également un très fort profil introductif. La phrase 24(5,0) présente les traits d'une clôture thématique et est de plus assez explicite quant à son caractère conclusif : "But this is certainly not the advantage which a modern reader can be promised from a study of their works". Entre les deux se dégage un profil mixte, comme celui de la phrase 26(6,4).

> If it were correct, the writers of political theory would need to be themselves past masters in the art of governing, and statesmen would need to apprentice themselves to them in order to learn their job.

Ce genre de phrase pourrait être interprété comme une transition entre deux unités thématiques. Hoey considère que la dualité de cette phrase peut s'observer à travers le "If it were correct" introducteur qui renvoie vers une partie antérieure du texte, et à travers l'intention implicite de l'auteur de s'appuyer sur cette réfutation d'une thèse précédemment énoncée pour présenter sa propre thèse, ce qui implique une ouverture vers la suite du texte. Ces indices ne font cependant que guider l'interprétation des profils cohésifs et restent beaucoup plus sujets à discussion que les liens de cohésion.

Ainsi, sans avoir la prétention de mettre au jour l'ensemble des structures discursives, cette approche montre qu'il est bel et bien possible de faire émerger des phénomènes d'ordre organisationnel en se basant exclusivement sur des indices lexicaux. Cette façon d'aborder le problème se distingue clairement des démarches qu'ont adoptées des auteurs comme [Marcu, 2000] ou [Morley, 2006], qui ne mobilisent d'indices de niveau lexical qu'en parallèle à une autre approche, à savoir respectivement un repérage des *cue phrases* et une prise en compte du formatage typo-dispositionnel des textes. Les applications de cette méthode sont multiples : en plus du résumé automatique ([Benbrahim et Ahmad, 1994], [Benbrahim, 1996]), elle a déjà été utilisée pour une tâche de segmentation thématique ([Berber Sardinha, 2001]), et Hoey a bon espoir de voir sa méthode appliquée à la didactique des langues. En effet, le repérage des phrases possédant le plus de liens de cohésion dans un texte en langue étrangère en faciliterait considérablement l'accès pour les apprenants, dans le sens où la redondance facilite l'interprétation. De plus, cette méthode est parfaitement compatible avec des langues autres que l'anglais, puisqu'elle a été appliquée avec succès au français ([Legallois, 2004]), au portugais ([Hoey, 1991] p. 232) et même au gallois ([Benbrahim et Ahmad, 1994]).

4.2 Application à un article encyclopédique

Avant de passer au repérage des liens de cohésion et à l'analyse du réseau de phrases obtenu, il nous faut revenir une dernière fois sur la question de l'annotation de l'article. Il pourrait paraître exagéré d'autant insister sur cet aspect du traitement des données qui peut sembler trivial, mais c'est en fait un point crucial du processus d'automatisation : si la correspondance entre le format des mots du texte et celui des mots de la base lexicale n'est pas optimale, alors la ressource sera sous-exploitée et l'on court le risque de brider le potentiel de la méthode de Hoey. Ce problème est mentionné dans quelques-unes des études que nous avons rencontrées :

> Wherever possible, we try to identify in the input text any compound words and phrases that are included in WordNet, because they are much better indicators of the meaning of the text than the words that compose them, taken separately. For instance, *private school*, which is listed in the noun index as *private_school*, is more indicative than *private* and *school* taken separately. ([Hirst et St. Onge, 1997] p. 11)

> Nominal compounds can be crucial in building correct lexical chains [...] ; considering the words *crystal* and *ball* independently is not at all the same thing as considering the phrase *crystal ball*. Roget's [Thesaurus] has a very large number of phrases, but we do not take advantage of this, as we do not have a way of tagging phrases in a text. ([Jarmasz et Szpakowicz, 2003] p. 2)

Nous avons déjà évoqué la question de la définition du *mot* et de la segmentation opérée par TreeTagger. L'analyse opérée par cet étiqueteur étant assez limitée, seuls les mots constitués d'une chaîne de caractères continue sont pris en compte, ce qui revient à traiter un pan entier des mots du texte d'une façon qui pourra poser problème lors de l'analyse.

Nous avons par exemple vu le cas des syntagmes prépositionnels : si le syntagme contient un mot plein, nous ne l'avons pas pris en compte (l'effacement a été effectué manuellement). Le problème des locutions nominales, verbales et adjectivales est plus complexe, puisqu'il faudrait, dans l'idéal, décider de la façon de traiter ces syntagmes au cas par cas, selon leur degré d'opacité. Une première idée serait d'essayer d'apparier les mots du texte avec les syntagmes de longueur maximale présents dans la base de relations. Cela permettrait une meilleure précision puisqu'un syntagme comme *donner lieu* pourra être identifié comme tel, ce qui permettra au système d'activer les 58 relations dans lesquelles il prend place. Dans le cas où le syntagme est traité comme deux mots différents, le système mobilisera les relations de *donner* et de *lieu*, qui sont certes beaucoup plus nombreuses (816 au total) mais qui seront forcément non pertinentes puisque le sens de *donner* et *lieu* séparés ne correspond pas du tout à celui de la locution. Cette méthode poserait toutefois deux problèmes, le premier étant que la reconstitution des syntagmes peut s'avérer assez complexe dans le sens où il ne suffit pas de concaténer des lemmes mais il faut également gérer les problèmes d'accord et d'élision : *homme d'affaires*, par exemple, se lemmatise *homme de affaire*, qui n'est évidemment pas présent dans notre base. Le deuxième problème se situe dans le fait que concaténer des mots pour en faire des syntagmes réduit forcément le nombre de liens potentiels d'une phrase. Or, dans le cas de locutions comme *frère aîné* ou *église catholique* (et contrairement à *homme d'affaires*), il peut être préférable de traiter les deux mots séparément puisque le degré d'opacité de ces locutions est nul et que la somme des relations des mots pris séparément est bien plus importante que s'ils avaient été

considérés comme un syntagme : on compte 11 relations pour *frère aîné*, 181 pour *frère*, 41 pour *aîné*, et 3 pour *église catholique*, 358 pour *église*, 90 pour *catholique*. Étant donné que ce problème reste assez complexe et qu'il serait hors-sujet de l'approfondir ici, nous préférons pour l'instant en rester à ces quelques remarques et réviser manuellement l'annotation du texte afin de lier les mots dont la prise en compte séparée risquerait de générer du bruit.

La première projection de la ressource que nous avons effectuée sur l'article *Narbonne*, avec un seuil de trois relations, a permis de mettre au jour pas moins de 6748 liens entre phrases. Le premier constat a été qu'un certain nombre de mots apparaissait à de très nombreuses reprises parmi les relations repérées, créant ainsi des liens sans grande pertinence : il suffisait que deux phrases contiennent des mots comme *Narbonne*, *ville* et *région* pour qu'elles soient reliées. Deux solutions se sont alors présentées : l'augmentation du seuil et l'effacement des mots les plus *envahissants*. Nous avons dans un premier temps préféré en rester à un seuil de 3 et effacer les 10 mots ayant le *tfidf* le plus important[6], soit, dans l'ordre, *narbonne*, *ville*, *commune*, *france*, *habitant*, *aude*, *siècle*, *national*, *région* et *posséder*[7]. Le nombre de liens tombe ainsi à 4070, ce qui reste très important. De plus, une rapide observation des résultats suffit pour se rendre compte que la grande majorité des liens interphrastiques sont dus à des relations lexicales qui, pour la plupart, sont loin d'être pertinentes. Par exemple, parmi les premiers résultats obtenus, le lien le plus fort a été repéré entre les phrases 187 et 198 :

> **<s=187>** Depuis que l'Aude a quitté son ancien cours, et que son lit accueille le canal de la Robine, classé par l'UNESCO au Patrimoine mondial de l'humanité, une seule arche suffit au passage de l'eau, les autres servant de caves aux maisons bâties des deux côtés du pont.
>
> **<s=198>** À proximité de la statue d'Ernest Ferroul, érigée en 1933 par la Confédération générale des vignerons par souscription nationale, est édifié, à partir de 1938, sous la direction de Joachim Genard, le palais des Arts, des Sports et du Travail, ensemble architectural d'influence néo-classique abritant une piscine, une salle de sport, une salle des fêtes, achevée en 1967, des salles d'assemblées et des sièges syndicaux.

Ces phrases sont reliées par les douze couples de mots suivants : *ancien/général, passage/travail, classer/abriter, servir/direction, pont/statue, accueillir/ériger, lit/siège, quitter/influence, cours/salle,*

[6]Le tfidf est un calcul permettant de définir l'importance d'un terme par rapport à un document en se basant sur la fréquence à laquelle il y apparaît et sur le nombre de documents de la collection dans lesquels il apparaît également. Le calcul a ici été effectué sur le corpus Wikipédia dont il était question à la section 3.1.1.

[7]Sans compter certains verbes comme *être*, *avoir* ou *faire*, qui ont été supprimés plus tôt dans la chaîne de traitement.

eau/proximité, bâtir/édifier, maison/fête. Étant donné le contexte, seules les paires *pont/statue* et *bâtir/édifier* sont ici pertinentes. Les relations *lit/siège* et *cours/salle* ne sont pas fondamentalement incorrectes, mais la polysémie de *lit* et de *cours* fait qu'elles sont ici inappropriées. En revanche, les huit paires restantes ne sont clairement pas pertinentes, et le contexte dans lequel elles pourraient éventuellement l'être n'est pas immédiatement perceptible. Ainsi, le fait que la majorité des relations lexicales qui relient ces deux phrases sont erronées rend leur rapprochement ininterprétable au niveau discursif puisqu'elles ne respectent aucune des deux propriétés des liens définies par Hoey :

> *The weak claim* : each bond marks a pair of sentences that is semantically related in a manner not entirely accounted for in terms of its shared lexis.
>
> *The strong claim* : because of the semantic relation refered to in the weak claim, each bond forms an intellegible pair in its context. ([Hoey, 1991] p. 125-126)

Ce premier constat nous a poussé à retravailler la ressource qui a servi à repérer ces relations. Étant donnée la surabondance de liens non pertinents que nous avons pu constater lors de cette première expérience, nous avons décidé de réduire le nombre de relations présentes dans la ressource en nous basant sur un critère de cooccurrence : nous n'avons gardé que les paires qui apparaissent ensemble au moins une fois dans une même phrase du corpus Wikipédia. Nous misons sur le fait que ce critère permettra d'éliminer une partie des couples les moins pertinents, ce qui permettra de repérer des liens à la fois moins nombreux et plus susceptibles de présenter des propriétés discursives analysables.

Une fois le filtrage effectué, la ressource passe de 475 798 relations à 260 392, ce qui correspond à une réduction d'environ 45%. Toutefois, le nombre de liens repérés grâce à la ressource filtrée est loin de baisser dans les mêmes proportions puisqu'il passe de 4070 à 3846. Cela peut s'expliquer par le fait qu'une grande partie des relations qui ont été retirées de la base grâce au filtrage l'ont été à cause du fait que les mots reliés n'étaient tout simplement pas repérables dans le corpus : la façon dont est étiqueté le texte du corpus ne permet pas de repérer certaines structures comme *donner un coup de main*, par exemple. Au niveau qualitatif, les résultats ne semblent pas plus convaincants. Si l'on reprend le couple de phrases vu précédemment, les relations lexicales repérées restent strictement les mêmes avant et après le filtrage de la ressource.

Nous avons ainsi décidé de repenser l'algorithme de repérage des liens de cohésion en introduisant une pondération des couples présents dans notre base lexicale : à chaque paire est attribué un score correspondant au nombre d'apparitions des deux mots dans la même phrase divisé par la somme des fréquences de chacun des deux mots. Ce calcul, qui se rapproche d'un calcul de collocation (au sens de Firth), permet d'installer une hiérarchie parmi les paires de la base lexicale : les couples ayant le score le plus élevé seront considérés commes les plus

0.01159 pierre	0.00030 monter
0.00984 édifice	0.00029 voir
0.00819 demeure	0.00029 comporter
0.00787 terrain	0.00028 retrouver
0.00661 emplacement	0.00028 organiser
0.00634 mur	0.00026 pousser
0.00610 construire	0.00026 assister
0.00597 architecte	0.00015 composer
0.00585 édifier	0.00013 peinture
0.00573 enceinte	0.00009 modifier

TAB. 4.3 – Comparaison des dix mots les plus fortement reliés au mot *bâtir* et des dix mots avec lesquels le score est le plus faible.

susceptibles de présenter une association sémantiquement pertinente. Nous avons rapporté au tableau 4.3, dans la colonne de gauche, les dix mots avec lesquels le mot *bâtir* a le score le plus élevé, et, dans la colonne de droite, ceux avec lesquels le score est le plus bas. Il apparaît très nettement que les mots de la colonne de gauche sont sémantiquement plus proches du verbe *bâtir* que ceux de la colonne de droite, ce qui implique que notre indice de collocation peut être considéré comme fiable et peut par conséquent être utilisé dans notre système.

Dès lors, l'appariement des mots des phrases à relier se fera en fonction de cet indice : un mot de la phrase 1 sera automatiquement relié au mot de la phrase 2 avec lequel il a le score le plus élevé. Les effets positifs de cette pondération sur les résultats obtenus peuvent se constater sur le couple de phrases que nous avons pris comme exemple précédemment. Pour rappel, les douze relations à l'origine du lien entre les phrases 187 et 198 étaient les suivantes : *ancien/général, passage/travail, classer/abriter, servir/direction, pont/statue, accueillir/ériger, lit/siège, quitter/influence, cours/salle, eau/proximité, bâtir/édifier, maison/fête*. En prenant en compte la pondération des relations, notre système relie le même couple de phrases par les relations suivantes : *ancien/général, maison/palais, cours/travail, eau/proximité, quitter/direction, accueillir/abriter, pont/statue, bâtir/édifier, servir/art, lit/salle, côté/influence, patrimoine/fête*. On peut constater que ces deux ensembles comptent quatre paires en commun (*ancien/général, pont/statue, eau/proximité, bâtir/édifier*), parmi lesquelles les deux paires que nous avions considérées comme les plus pertinentes lors de notre première analyse. Et si certaines paires comme *servir/art* ou *lit/salle* restent peu voire pas du tout pertinentes, l'on constate que certains couples se sont ici formés de façon plus judicieuse : ainsi, *maison* n'est plus relié à *fête* mais à *palais*,

et *accueillir* est désormais relié à *abriter* et non plus à *ériger*.

Cette modification du système entraine une nouvelle baisse du nombre de liens repéré sur l'article *Narbonne* : celui-ci passe de 3846 à 3134. Étant donné qu'il nous est impossible d'analyser en détail un tel nombre de liens, plusieurs perspectives se sont offertes à nous, dont une qui aurait consisté à tronquer le texte pour arriver à un nombre de liens moins important. Cette possibilité à toutefois été rapidement écartée étant donné qu'elle exclurait l'analyse des liens de longue portée, qui font partie intégrante de la théorie de Hoey. Nous avons préféré centrer notre attention sur un sous-ensemble bien particulier des phrases du texte, à savoir les sept premières. Ces phrases constituent ce que la terminologie de Wikipédia appelle un *résumé introductif*[8]. Le rôle du résumé introductif par rapport au reste de l'article est de poser son contexte et de reprendre les principaux points qui y sont développés : il se distingue de l'introduction classique par le fait qu'il ne fait pas que présenter d'éventuelles problématiques qui seront discutées dans le reste de l'article, mais il rapporte également les réponses. Il se doit avant tout d'être un condensé "autonome et auto-suffisant", une sorte de "mini-article" qui serait une réduction de l'article original composée de ses points-clés. L'intérêt que présente un tel objet du point de vue de l'étude de la cohésion est indéniable. Nous avons vu que la méthode développée par Hoey permettait de générer des résumés par la concaténation des phrases d'un texte identifiées comme centrales (pour rappel, ces phrases sont celles qui possèdent le nombre maximal de liens avec les autres phrases). Par conséquent, la première supposition que l'on peut faire sur la nature des phrases que contient un résumé introductif, est qu'elles auront des chances de posséder un nombre de liens plus important que la moyenne[9].

Ainsi, en réduisant notre champ d'investigation aux sept premières phrases du texte, nous augmentons nos chances d'observer des phénomènes comme des liens de longue distance entre les phrases du résumé introductif et leur développement dans le corps du texte. Les sept phrases en question sont les suivantes[10] :

> **<s=2>** Narbonne (en occitan Narbona) est une commune française, située dans le département de l'Aude et la région Languedoc-Roussillon.
>
> **<s=3>** La commune est traversée par le canal de la Robine, classé au Patrimoine mondial de l'humanité par l'UNESCO depuis 1996.
>
> **<s=4>** Elle est classée 120e ville de France en termes de population.

[8] http ://fr.wikipedia.org/wiki/Wikipédia :Résumé_introductif

[9] Dans le même ordre d'idée, l'importance des sections introductives commes les chapeaux des textes journalistiques pour l'analyse des liens interphrastiques est brièvement évoquée dans [Legallois, 2004].

[10] Les titres ayant été balisés comme des phrases, la première phrase du texte porte le numéro 2 puisqu'elle est précédée par le titre de l'article. De plus, comme nous reviendrons assez en détail sur ces quelques phrases, nous avons rapporté en annexe A.2 leur version étiquetée.

<s=5> Située au cœur du 51e parc naturel de France «parc naturel régional de la Narbonnaise en Méditerranée», Narbonne possède également d'autres sites naturels classés, comme le massif de la Clape et celui de l'abbaye Sainte-Marie de Fontfroide, ainsi que les étangs de Bages-Sigean.

<s=6> Fondée par les Romains en 118 av. J. C., elle était leur plus ancienne colonie en Gaule et son centre urbain garde trace de nombreux siècles d'histoire (cathédrale Saint-Just-et-Saint-Pasteur, palais des Archevêques, restes de la voie Domitienne).

<s=7> La ville est entourée d'un environnement exceptionnel fait de garrigues et de vignes ; proche du littoral d'une région très touristique, elle possède une plage de cinq kilomètres de sable fin à Narbonne-Plage.

<s=8> Les habitants de Narbonne sont appelés les Narbonnais.

Ces phrases abordent différents aspects de la ville de Narbonne comme sa situation géographique (phrase 2), son patrimoine environnemental (phrases 3, 5 et 7), sa démographie (phrase 4) et son histoire (phrase 6). La phrase 8 correspond à l'indication du gentilé, qui conclut un grand nombre de résumés introductifs d'articles portant sur des villes. Ces thèmes ne sont pas les seuls à être traités dans l'article, mais l'on peut considérer que ce sont ceux qui y sont les plus développés. Le nombre total de liens rattachés à ces sept phrases est de 280. On peut voir à la figure 4.3 que ces liens sont répartis de façon assez inégale. Cela est principalement dû au fait que toutes les phrases ne possèdent pas le même nombre de mots :

- les phrases 4 et 8 ne sont reliées à aucune autre phrase du texte étant donné qu'il leur est impossible d'atteindre le seuil des trois relations : les mots *France*, *habitant*, *ville* et *Narbonne* faisant partie des dix mots ayant le tfidf le plus élevé, ils ont été précédemment exclus de l'analyse, ce qui fait que les phrases 4 et 8 ne comptent plus que 2 mots susceptibles de créer des liens de cohésion,
- les phrases 5 et 6, qui sont celles qui comptent le plus de mots, sont les plus fortement reliées au reste du texte.

Nous allons maintenant procéder à une analyse détaillée des phrases reliés aux phrases 2, 3, 5, 6 et 7.

4.2.1 Étude des phrases du résumé introductif

Phrase 2

<s=2> Narbonne (en occitan Narbona) est une commune française, située dans le département de l'Aude et la région Languedoc-Roussillon.

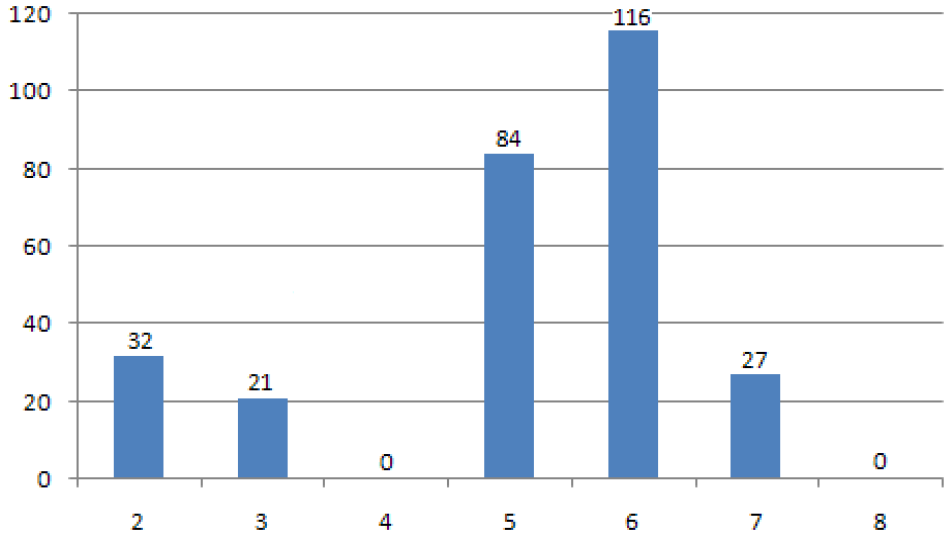

FIG. 4.3 – Nombre de liens rattachés à chacune des sept premières phrases de l'article *Narbonne*.

Cette phrase est la première du résumé introductif, donc de l'article, et en tant que telle, elle constitue le premier contact du lecteur avec le texte de l'article. Les deux informations qu'elle fournit sont la nature de l'objet dont il est question et sa localisation (puisqu'il est ici question d'un lieu). Une fois retirés les mots filtrés par le tfidf (ici *Narbonne* et *commune*), les mots susceptibles de créer des liens de cohésion sont *occitan*, *Narbona*, *français*, *situer*, *département*, *Aude*, *région* et *Languedoc-Roussillon*. Parmi ces huit mots, trois sont des noms propres et auront donc peu de chances d'établir des relations autres que la répétition simple. Le mot qui nous apparaît le plus représentatif de la phrase est le verbe *situer*, puisque sa répétition permet de repérer des phrases qui, à l'instar de la phrase 2, ont trait à la localisation (aucun des mots reliés à *situer* ne s'est avéré assez pertinent pour permettre de repérer des phrases ayant un rapport avec une idée de localisation) :

> **<s=5>** Située au cœur du 51e parc naturel de France « parc naturel régional de la Narbonnaise en Méditerranée », Narbonne possède également d'autres sites naturels classés, comme le massif de la Clape et celui de l'abbaye Sainte-Marie de Fontfroide, ainsi que les étangs de Bages-Sigean.
>
> **Relations :** situer/situer, français/régional, département/massif

Cette phrase est tout à fait compatible avec la phrase 2 et vient développer son propos en apportant des précisions sur la situation géographique de la ville de Narbonne (“au cœur du 51e parc naturel de France”). Toutefois, on ne peut évidemment pas parler ici de lien de longue distance puisque cette phrase se trouve dans la même section que la phrase de départ. La section de l'article la plus à même de voir ses phrases reliées à la phrase 2 est la section *1.1.1. Localisation*,

qui englobe les phrases 12 à 17. Parmi celles-là, seule la phrase 13 a un lien avec la phrase 2 :

> **<s=13>** Centrée sur la basse plaine de l'Aude, elle s'étend, d'est en ouest, de la mer Méditerranée aux contreforts des Corbières maritimes, englobant le massif calcaire de la Clape, et, du nord au sud, du cours actuel de l'Aude aux abords de l'île de Sainte-Lucie, intégrant une grande partie des étangs de Bages-Sigean.
>
> **Relations :** situer/ouest, département/nord, français/actuel

Si l'on ne peut que rester dubitatif quant à la pertinence des relations qui permettent de créer ce lien, force est de constater que celui-ci est tout à fait pertinent puisque cette phrase joue le même rôle que la phrase 5, à savoir apporter un développement de la phrase 2[11]. Toutefois, ce développement se fait de manière indirecte puisque la phrase 13 réfère directement à la phrase qui la précède (notamment via le pronom personnel anaphorique *elle*), qui est à nos yeux la plus proche de la phrase 2 (à laquelle elle n'est toutefois pas liée) :

> **<s=12>** Narbonne est située dans le sud-est de la France, sur la côte méditerranéenne, entre Montpellier au nord et Perpignan au sud.

La section *1.2. Site* (phrases 43 à 48) contient également des phrases qu'il pourrait être pertinent de relier avec la phrase 2, ce qui n'est le cas que pour l'une d'entre elles, la phrase 47 :

> **<s=47>** Le territoire de la commune englobe aussi la station balnéaire Narbonne-Plage située à plus de dix kilomètres du centre ancien de Narbonne.
>
> **Relations :** situer/situer, français/ancien, département/territoire

L'on peut également remarquer quelques liens pointant vers la partie de l'article dédiée à l'histoire de Narbonne. Cela est dû au fait que certaines phrases évoquent l'évolution géopolitique de la ville :

> **<s=89>** En 759 la ville est reprise par Pépin le Bref et fait définitivement partie du royaume franc.
>
> **Relations :** français/franc, situer/faire partie, département/royaume
>
> **<s=82>** Elle était située sur la via Domitia, la première route romaine en Gaule, qui permettait de relier l'Italie et l'Espagne.
>
> **Relations :** français/franc, situer/faire partie, département/royaume

Phrase 3

<s=3> La commune est traversée par le canal de la Robine, classé au Patrimoine mondial de l'humanité par l'UNESCO depuis 1996.

[11] Si l'on s'était placé dans le cadre de la RST, il est probable que nous aurions parlé de relation d'*élaboration*.

La phrase avec laquelle la phrase 3 est la plus fortement rattachée est la 187^{e}, puisque le lien qui les relie repose sur pas moins de 8 relations. Toutefois, même si ce rapprochement peut être considéré comme pertinent, l'apparente force de ce lien est à relativiser puisque sur ces 8 relations, cinq sont des répétitions d'éléments d'une même formule qui apparaît relativement figée ("classé au Patrimoine mondial de l'humanité par l'UNESCO").

> **<s=187>** Depuis que l'Aude a quitté son ancien cours, et que son lit accueille le canal de la Robine, classé par l'UNESCO au Patrimoine mondial de l'humanité, une seule arche suffit au passage de l'eau, les autres servant de caves aux maisons bâties des deux côtés du pont.
>
> **Relations :** humanité/humanité, mondial/mondial, unesco/unesco, classer/classer, patrimoine/patrimoine, robine/robine, canal/canal, traverser/pont

La phrase de l'article qui nous apparaît comme la plus proche de la phrase 3 est la phrase 31. On peut en effet considérer que ces deux phrases entretiennent une relation de paraphrase puisque l'information véhiculée par la phrase 3 ("La commune est traversée par le canal de la Robine") se retrouve reformulée dans la phrase 31 via une inversion passif/actif :

> **<s=31>** Au niveau fluvial, le canal de la Robine traverse la ville en suivant l'ancien lit de l'Aude.
>
> **Relations :** traverser/traverser, robine/robine, canal/canal, classer/suivre

Dans ces deux cas, c'est la répétition seule qui a permis de créer ces liens, mais certaines relations lexicales pertinentes permettent également de faire des rapprochements intéressants. Par exemple dans les deux phrases suivantes, le verbe *traverser* est relié à *rejoindre* et à *relier*. Combinés, dans le premier cas avec une répétition de *canal* et, dans le deuxième cas, avec une relation *canal/pont*, ces relations permettent de rapprocher des phrases qui expriment la localisation géographique d'un élément par rapport à un autre, l'un de ses éléments relevant du domaine des cours d'eau :

> **<s=32>** Il permet de rejoindre le fleuve Aude puis le canal du Midi, qui passe plus au nord de la ville, via le canal de Jonction.
>
> **Relations :** canal/canal, traverser/rejoindre, classer/passer
>
> **<s=185>** Le pont des Marchands, reliant le Bourg à la Cité, ce pont permettait à l'origine le franchissement de l'Aude par la voie Domitienne.
>
> **Relations :** canal/pont, traverser/relier, classer/origine

Un autre type de rapprochements est dû au mot *patrimoine*, qui se retrouve lié à *monument* et *édifice* dans les phrases suivantes :

<s=170> Unique en Europe, il est situé sous un monument disparu qui aurait pu servir d'entrepôt public (horreum).

Relations : patrimoine/monument, mondial/unique, classer/pouvoir, traverser/situer

<s=191> L'église Notre-Dame de Lamourguier, édifice typique du gothique languedocien et catalan, est l'unique reste d'un prieuré bénédictin.

Relations : patrimoine/édifice, mondial/unique, traverser/reste, classer/gothique

Ces deux phrases se situent respectivement dans les sections *8.1.1. Patrimoine antique* et *8.1.2. Patrimoine médiéval*, incluses dans la section *8.1. Patrimoine culturel*, elle-même subordonnée à la section *8. Patrimoine*. Une mention du canal de la Robine comme élément du patrimoine peut également se retrouver dans la section *8.2. Patrimoine environnemental*, à la phrase 210, qui est aussi liée à la phrase 3 :

<s=210> Plusieurs espaces verts sont à la disposition des Narbonnais, comme le jardin de la Révolution, le jardin du Plan Saint Paul, le jardin des Archevêques, récemment rénové, le parc de la Campane, les berges du canal de la Robine et le massif de l'abbaye de Fontfroide.

Relations : robine/robine, canal/canal, traverser/massif

Un dernier rapprochement qui mérite notre attention est celui que l'on observe entre la phrase 3 et la phrase 262 :

<s=262> - Classée " Ports propres " (pour la Nautique et Narbonne-Plage)

Relations : classer/classer, canal/port, mondial/propre

En effet, comme peuvent le laisser deviner le tiret et l'absence de sujet, cette phrase se distingue de la plupart des phrases du texte par le fait qu'elle s'inscrit dans un contexte de liste. Elle est en effet un élément de l'énumération que constitue la section *13. Distinctions* (nous reviendrons plus tard sur l'intérêt que peuvent présenter les structures énumératives dans le cadre d'une étude sur la cohésion). Ce lien peut être considéré comme pertinent étant donné que les deux phrases expriment le fait que la ville de Narbonne se distingue favorablement par ses aspects fluviaux/maritimes.

Phrase 5

<s=5> Située au cœur du 51e parc naturel de France «parc naturel régional de la Narbonnaise en Méditerranée», Narbonne possède également d'autres sites naturels classés, comme le massif de la Clape et celui de l'abbaye Sainte-Marie de Fontfroide, ainsi que les étangs de Bages-Sigean.

Étant donné qu'elle est l'une des deux phrases les plus volumineuses du résumé introductif, la phrase 5 possède un nombre assez conséquent de liens (84) dont la plupart ne sont que peu pertinents. Comme c'était le cas pour la phrase 2, le verbe *situer* est à l'origine de certains rapprochements avec des phrases exprimant une idée de localisation :

> **<s=14>** Située dans les Corbières maritimes, Narbonne est un carrefour connu de l'homme depuis le néolithique.
>
> **Relations :** situer/situer, naturel/connaître, régional/maritime

Toutefois, la principale information contenue dans la phrase 5 est qu'elle possède des sites naturels. Certains de ces sites sont mentionnés à plusieurs endroits du texte, ce qui entraîne la création de liens, comme avec les phrases 13, 193 et 210 :

> **<s=13>** Centrée sur la basse plaine de l'Aude, elle s'étend, d'est en ouest, de la mer Méditerranée aux contreforts des Corbières maritimes, englobant le massif calcaire de la Clape, et, du nord au sud, du cours actuel de l'Aude aux abords de l'île de Sainte-Lucie, intégrant une grande partie des étangs de Bages-Sigean.
>
> **Relations :** étang/étang, massif/massif, clape/clape, méditerranée/méditerranée, bages-sigean/bages-sigean, situer/ouest, parc/île, site/partie, régional/maritime, naturel/actuel, classer/intégrer
>
> **<s=193>** Non loin de Narbonne, l'abbaye Sainte-Marie de Fontfroide est un ancien monastère cistercien.
>
> **Relations :** abbaye/abbaye, sainte-marie/sainte-marie, fontfroide/fontfroide
>
> **<s=210>** Plusieurs espaces verts sont à la disposition des Narbonnais, comme le jardin de la Révolution, le jardin du Plan Saint Paul, le jardin des Archevêques, récemment rénové, le parc de la Campane, les berges du canal de la Robine et le massif de l'abbaye de Fontfroide.
>
> **Relations :** parc/parc, abbaye/abbaye, massif/massif, fontfroide/fontfroide, site/espace, situer/plan

Il est toutefois possible de repérer des liens vers des phrases que l'on peut interpréter comme des développements de la phrase 5 sans que n'intervienne la répétition simple (ou dans une moindre mesure) :

> **<s=7>** La ville est entourée d'un environnement exceptionnel fait de garrigues et de vignes ; proche du littoral d'une région très touristique, elle possède une plage de cinq kilomètres de sable fin à Narbonne-Plage.
>
> **Relations :** situer/kilomètre, étang/plage, naturel/proche, massif/entourer

<s=46> La commune possède au sud une partie de l'étang de Sigean.

Relations : étang/étang, situer/sud, parc/partie

<s=47> Le territoire de la commune englobe aussi la station balnéaire Narbonne-Plage située à plus de dix kilomètres du centre ancien de Narbonne.

Relations : situer/situer, parc/centre, site/station

La phrase 7 fait partie du résumé introductif et peut être considérée comme un prolongement de la phrase 5 dans le sens où *garrigue*, *vigne*, *littoral* et *plage* sont des hyponymes de *site naturel*. Les phrases 46 et 47 représentent le tiers des phrases qui composent la section *1.2. Site*. Directement subordonnée à la section *1. Géographie*, cette section contient des considérations assez générales sur l'implantation de la ville de Narbonne, et il n'est donc pas étonnant de voir certaines de ses phrases liées avec la phrase 5.

Parmi la masse de relations incorrectes, il est possible de trouver certains liens basés sur des relations sémantiques pertinentes comme c'est le cas avec la phrase 30 :

<s=30> Enfin, la route départementale 613 permet de rejoindre au sud-ouest Couiza et la haute vallée de l'Aude en traversant la région des Corbières et la départementale 607 rejoint au nord-ouest Lacaune dans la montagne Noire.

Relations : massif/vallée, régional/départemental, situer/nord-ouest, parc/montagne, site/route

On ne peut pas réellement considérer que cette phrase forme une paire du point de vue discursif avec la phrase 5 puisqu'elles ne font que partager un même champ sémantique sans que l'une ne renvoie à l'autre, même de façon implicite.

Phrase 6

<s=6> Fondée par les Romains en 118 av. J. C., elle était leur plus ancienne colonie en Gaule et son centre urbain garde trace de nombreux siècles d'histoire (cathédrale Saint-Just-et-Saint-Pasteur, palais des Archevêques, restes de la voie Domitienne).

Parmi les sept phrases du résumé introductif, la phrase 6 est celle qui possède le plus de liens (116). Elle porte sur l'histoire de la ville de Narbonne et, à l'instar de la phrase 5, contient une énumération (les éléments entre parenthèses sont des exemples de vestiges romains).

Une fois écartés les liens générés par des relations non pertinentes, l'on obtient deux ensembles de liens qui pointent clairement vers deux endroits distincts du texte. Le premier ensemble regroupe les phrases 81, 82, 83, 84 et 85 :

<s=81> Les Romains fondèrent en 118 av. J.-C. une colonie romaine du nom de Colonia Narbo Martius d'après le nom du consul romain Quintus Marcius Rex.

Relations : colonie/colonie, romain/romain, histoire/nom, ancien/fonder, archevêque/consul

<s=82> Elle était située sur la via Domitia, la première route romaine en Gaule, qui permettait de relier l'Italie et l'Espagne.

Relations : gaule/gaule, romain/romain, voie/route, reste/situer

<s=83> Avant cette période, Narbonne était un oppidum préromain (site de Montlaurès, à quatre kilomètres au nord de la ville actuelle), la capitale des Elisyques, un des peuples de la Celtique méditerranéenne.

Relations : romain/peuple, ancien/actuel, centre/nord, histoire/période, palais/capitale

<s=84> Narbonne était la capitale de la Province de la Gaule narbonnaise, province romaine créée en 27 avant J.-C.

Relations : gaule/gaule, romain/romain, centre/capitale, histoire/créer, colonie/province

<s=85> Elle fut jusqu'à la fin de l'Antiquité romaine l'une des villes les plus importantes de la Gaule avec une superficie de plus de 2 kilomètres carrés.

Relations : gaule/gaule, romain/romain, nombreux/important, histoire/antiquité

Le bloc de texte que constituent ces cinq phrases se situe dans la section *2. Histoire*. Il se caractérise par la récurrence des mots *romain* et *gaule*, qui sont également présents dans la phrase 6. Toutefois, la répétition de ces mots n'est ici clairement pas le seul facteur qui a permis de relier toutes ces phrases à la phrase 6 puisque l'on peut observer que ces rapprochements sont également motivés par un certain nombre de relations tout à fait pertinentes comme *archevêque/consul*, *voie/route*, *romain/peuple*, *histoire/période*, *histoire/antiquité*, etc.

Le deuxième ensemble, moins dense que le premier, regroupe les phrases 171, 172, 175, 183, 188, 189 et 193 :

<s=171> Au centre de la place de l'Hôtel-de-Ville, l'antique voie Domitienne (Via Domitia)est visible dans son état de la fin du IVe siècle.

Relations : voie/voie, domitien/domitien, centre/centre, romain/antique, palais/place, histoire/état

<s=172> C'est un vestige de la première grande route romaine, tracée en Gaule à partir de 120 av. J.-C., par le proconsul Cneus Domitius Ahenobarbus, deux ans avant la fondation de la Colonia Narbo Martius, première colonie romaine en Gaule.

Relations : colonie/colonie, gaule/gaule, romain/romain, voie/route, trace/vestige, reste/partir

<s=175> Le vestige découvert le 7 février 1997 présente une portion de voie dallée de calcaire dur, marquée par de profondes ornières.

Relations : voie/voie, trace/vestige, reste/découvrir, garder/présenter, histoire/dur

<s=183> Le palais des archevêques est composé du palais Vieux d'origine romane et du palais Neuf de style gothique.

Relations : palais/palais, archevêque/archevêque, cathédrale/gothique, histoire/roman, ancien/vieux, reste/composer, centre/style

<s=188> La basilique Saint-Paul, l'une des plus anciennes églises gothiques du Midi de la France, est construite sur les vestiges d'un ancien cimetière paléochrétien (IIIe-IVe siècles).

Relations : ancien/ancien, cathédrale/gothique, palais/église, trace/vestige, garder/construire, colonie/cimetière

<s=189> Cet édifice a la particularité de mêler art roman et art gothique.

Relations : cathédrale/gothique, histoire/art, palais/édifice, trace/particularité

<s=193> Non loin de Narbonne, l'abbaye Sainte-Marie de Fontfroide est un ancien monastère cistercien.

Relations : ancien/ancien, cathédrale/abbaye, palais/monastère

Ces phrases s'étalent sur les sections *8.1.1. Patrimoine antique* et *8.1.2. Patrimoine médiéval*. Elles recoupent largement la thématique historique mais la section dans laquelle elles prennent place, *8. Patrimoine*, se situe à une certaine distance de la section *2. Histoire*. Cela peut éventuellement s'expliquer par le fait qu'il n'est ici pas tant question de l'histoire de la ville de Narbonne mais de celle de ses vestiges. En effet, cet ensemble se distingue du premier par le fait qu'il est parcouru par le champ sémantique du bâtiment (souvent religieux), notamment à travers les relations *cathédrale/gothique*, *cathédrale/abbaye*, *palais/édifice*, *palais/monastère*, *palais/église*. La répétition simple semble également y tenir une part moins importante et l'on peut y remarquer des relations particulièrement pertinentes comme *trace/vestige* qu'il serait intéressant de distinguer automatiquement : il se trouve que le mot *trace* est particulièrement polysémique et qu'il est la cause d'un certain nombre de relations erronées, du moins dans ce contexte, comme *trace/passage* ou *trace/suivre*.

Phrase 7

<s=7> La ville est entourée d'un environnement exceptionnel fait de garrigues et de vignes ; proche du littoral d'une région très touristique, elle possède une plage de cinq kilomètres de sable fin à Narbonne-Plage.

Ici encore le thème abordé est celui de la situation géographique de la ville et de ses atouts touristiques. Les phrases que l'on retrouve liées à la phrase 7 sont du même type que celles qui sont associées avec la phrase 5 (nous avons d'ailleurs vu que ces deux phrases étaient liées entre elles), si bien que les deux seuls liens que l'on peut considérer comme pertinents pour la phrase 7 pointent vers des phrases que nous avions déjà rapprochées lors de l'analyse de la phrase 5 :

> **<s=13>** Centrée sur la basse plaine de l'Aude, elle s'étend, d'est en ouest, de la mer Méditerranée aux contreforts des Corbières maritimes, englobant le massif calcaire de la Clape, et, du nord au sud, du cours actuel de l'Aude aux abords de l'île de Sainte-Lucie, intégrant une grande partie des étangs de Bages-Sigean.
>
> **Relations :** plage/mer, littoral/sud, sable/île, proche/actuel, kilomètre/abord, entourer/massif

> **<s=47>** Le territoire de la commune englobe aussi la station balnéaire Narbonne-Plage située à plus de dix kilomètres du centre ancien de Narbonne.
>
> **Relations :** narbonne-plage/narbonne-plage, kilomètre/kilomètre, littoral/situer, entourer/englober

4.2.2 Analyse globale

Le premier constat que l'on peut faire au vu des résultats obtenus est que certaines paires que le système a permis de former se révèlent tout à fait conformes à ce que l'on pouvait attendre d'une application de la méthode de Hoey. En effet, nous avons montré que certaines phrases, parfois assez éloignées du début du texte, faisaient logiquement écho à des phrases contenues dans le résumé introductif. Lors de notre interprétation de ces liens, nous avons évoqué une relation de *développement* entre ces phrases. Parmi les indices qui ont motivé une telle interprétation, on peut citer la présence de structures de forme *hypéronyme+hyponyme 1, hyponyme 2, hyponyme3* comme dans les phrases 5, 6 voire 7. Ces dernières nous sont en effet particulièrement précieuses puisque l'énumération des hyponymes permet de générer des liens que la seule mention de l'hypéronyme n'aurait pas pu permettre de repérer. Par exemple, dans la phrase 5, le fait que *les étangs (de Bages-Sigean)* soient évoqués comme des sites naturels a permis de relier cette phrase avec la phrase 7, où *étang* est relié à *plage*. Par conséquent, *plage* peut lui aussi être considéré comme un *site naturel*, et la phrase 7 comme une extension de la phrase 5. La deuxième raison qui appuie cette idée est liée à la situation des phrases étudiées au sein du texte. Nous avons en effet analysé des phrases qui constituent une entité discursive, le résumé introductif, dont le rôle est clairement défini : puisqu'un résumé consiste à compacter

l'information contenue dans un texte, il est tout à fait logique de voir les phrases qu'il contient reliées à d'autres phrases qui développent les points évoqués.

Toutefois, il faut garder à l'esprit que les paires pertinentes comme celles que nous avons analysées sont souvent noyées parmi un ensemble de liens inanalysables dont voici deux exemples :

> **<s=3>** La commune est traversée par le canal de la Robine, classé au Patrimoine mondial de l'humanité par l'UNESCO depuis 1996.
>
> **<s=225>** Michel Sidobre, dans Et pourquoi pas Narbonne, décrit l'ombre des platanes et la douceur de vivre.
>
> **Relations :** classer/décrire, traverser/vivre, humanité/douceur

> **<s=7>** La ville est entourée d'un environnement exceptionnel fait de garrigues et de vignes ; proche du littoral d'une région très touristique, elle possède une plage de cinq kilomètres de sable fin à Narbonne-Plage.
>
> **<s=223>** C'est alors qu'un jeune homme inconnu de vingt ans, qui est pris « pour une fille habillée en garçon », et nommé Aymerillot, s'avance vers l'empereur et affirme : « J'entrerai dans Narbonne et je serai vainqueur. »
>
> **Relations :** plage/garçon, proche/nommer, entourer/prendre

Ainsi, nous avons certes repéré pour la phrase 6 au moins une dizaine de phrases pour lesquelles le rattachement faisait sens, mais il est difficile de considérer ces rapprochements comme une preuve de l'efficacité du système puisque la phrase 6 est reliée à presque 44% des phrases du texte. De plus, parmi tous les liens que nous avons précédemment analysés, beaucoup sont en partie construits sur des relations incorrectes comme *naturel/connaître, classer/suivre*, ou *traverser/reste*. Dans ce genre de cas, c'est la qualité de la ressource qui est à remettre en cause. Il est évident qu'en l'état, et malgré les traitements que nous lui avons appliqués, notre base lexicale est beaucoup trop bruitée. Il ne fait aucun doute qu'une amélioration du système ne pourrait s'envisager sans une évaluation des ressources et la mise en place de nouvelles méthodes de filtrage. Dans un deuxième temps, un second niveau de filtrage contextuel pourrait détecter que les relations comme *lit/siège* et *cours/salle*, que nous avons évoquées plus haut, ne sont pas pertinentes quand elles relient les phrases 187 et 198.

Au vu de ce constat plutôt pessimiste, il serait tentant de se rabattre sur une relation de cohésion qui laisse beaucoup moins de place à l'ambiguïté, à savoir la répétition simple. En effet, nous avons vu lors de nos analyses qu'un certain nombre de rapprochements pertinents pouvaient être en partie dus au fait que les deux phrases reliées avaient des mots en commun. Le rôle positif que joue la répétition simple dans le repérage de liens est indéniable, mais la seule prise en compte de cette relation est loin d'être une alternative crédible. En effet, sur l'ensemble

des phrases du résumé introductif, la projection des seuls liens de répétition simple n'a permis de repérer que huit liens (entre les phrases 2 et 231, 3 et 31, 3 et 187, 5 et 13, 5 et 193, 5 et 210, 6 et 171, 6 et 172), dont certains, comme celui qui relie les phrases 2 et 231, peuvent ne pas être considérés comme porteurs d'une relation discursive :

> **<s=2>** Narbonne (en occitan Narbona) est une commune française, située dans le département de l'Aude et la région Languedoc-Roussillon.
>
> **<s=231>** Les studios de la première radio occitane du Languedoc-Roussillon, Ràdio Lenga d'òc, sont situés à Narbonne.
>
> **Relations :** languedoc-roussillon/languedoc-roussillon, occitan/occitan, situer/situer

Ainsi, si la répétition simple reste la relation de cohésion *par excellence* puisqu'elle relie dans la plupart des cas (on n'est jamais à l'abri de la polysémie…) deux mots qui ont des contenus sémantiques identiques, il apparaît que sa seule prise en compte pour le repérage de liens interphrastiques limite considérablement la portée de l'analyse. Son utilisation en plus d'une ressource externe nous semble actuellement la meilleure alternative.

À ce stade de l'étude, nous devons écarter la possibilité de nous livrer à des analyses en termes de phrases centrales, d'ouvertures et de fermetures thématiques puisque la pertinence des liens repérés automatiquement par notre système doit être évaluée au cas par cas. Nous avons en effet vu que ce type d'analyse se basait sur le calcul du nombre de liens d'une phrase et de leur orientation. Or, il nous est impossible d'entreprendre une telle démarche avec les liens que nous avons repérés, puisque la grande majorité d'entre eux ne sont pas fiables (car basés sur des relations lexicales non pertinentes). Ainsi, alors que nous nous trouvons obligé de constater qu'un pan entier de l'analyse *à la Hoey* se retrouve hors de notre portée, il est légitime de se poser la question suivante : la méthode de Hoey est-elle réellement automatisable ?

Comme nous l'avons fait remarquer plus haut, il est étonnant de voir à quel point l'aspect *automatisation* est absent de *Patterns of lexis in text*. La méthode d'analyse qui y est développée repose en effet sur l'idée que les liens générés par le repérage des relations de cohésion peuvent jouer un rôle au niveau organisationnel, et ce quelle que soit la distance qui sépare les deux phrases. Or, la première étape de cette analyse implique de comparer toutes les phrases du texte analysé deux à deux, ce qui, dans le cas où l'analyse est effectuée *à la main*, limite considérablement la taille des textes qu'il est possible de traiter. Et il est difficile de considérer que deux phrases rapprochées dans un texte qui en compte vingt peuvent constituer un lien *de longue distance*. La théorie de Hoey ne peut donc pas, à notre avis, être explorée comme elle le mériterait tant qu'elle n'a pas été automatisée. Au vu d'études comme [Benbrahim et Ahmad, 1994] et [Benbrahim, 1996], il serait difficile de répondre à la question qui conclut le paragraphe

précédent par la négative. Ces travaux, que nous avons déjà évoqués, ont en effet consisté en l'automatisation de la méthode de Hoey pour la génération de résumés d'articles scientifiques, et ont permis de fournir des résultats validés par des experts du domaine concerné (la physique nucléaire). Le type de ressource mobilisé (un thésaurus) et le genre de texte analysé font que les auteurs n'ont, semble-t-il, quasiment jamais rencontré de problème de relations lexicales non pertinentes. Le seul cas qu'ils mentionnent est celui d'une relation *make/shape* qui était simplement due à une erreur d'annotation morpho-syntaxique. Cela les amène à une conclusion qui est à l'opposé de la nôtre :

> Note also that the automatic use of the thesaurus can sometimes lead to the extraction of links that are rather ambiguous and misleading. We refer in this respect to the simple mutual paraphrase links *make-shape* and *shapes-make*. The relationships between these pairs of words are very suspect. The word *shapes* is treated as a noun in the text whereas the word *make* is actually used as a verb. [...] Nevertheless, **suspect links and the like are not harmful**. (p. 89) [nous soulignons]

Nous analysons notre semi-échec comme étant dû, d'une part, à une mauvaise maîtrise de la ressource mobilisée (qui s'est révélée bien trop permissive), et d'autre part, au fait qu'il était peut-être un peu ambitieux de se lancer directement dans une projection des relations sur l'ensemble du texte. Nous avons justifié le fait de nous servir d'un article portant sur une ville comme texte de travail par le fait que les thématiques y sont relativement séparées. Nous comptions sur le fait que cela aurait pu permettre de *guider* la formation de liens, dans le sens où la différence de thème, et donc de lexique, entre les sections les aurait rendues un tant soit peu imperméables les unes aux autres. Force est de constater que ce n'est pas le cas : la figure 4.4 montre que même si l'on observe quelques agglomérats (nous en avons décrit certains plus haut), la répartition des liens est loin d'être aussi ciblée que ce que l'on pouvait espérer, notamment pour les phrases les plus volumineuses comme la 5 et la 6. Il est toutefois intéressant de noter que certaines zones du texte ne sont pas (ou quasiment pas) couvertes par ces deux phrases, ni par les autres. Ces zones, qui sont au nombre de trois, peuvent être clairement délimitées car elles correspondent aux sections *1.5. Structures* (phrases 53 à 63), *7. Personnages célèbres* (phrases 146 à 163), et *13. Distinctions* (phrases 254 à 256). Ces sections présentent la particularité de ne pas être composées de phrases reliées les unes aux autres de façon linéaire à l'aide, éventuellement, de connecteurs linguistiques, mais d'*items* qui sont pour la plupart des syntagmes nominaux. En effet, ces sections sont des structures énumératives et la raison pour laquelle elles se retrouvent *marginalisées* (au sens de Hoey) est que, comme on peut le voir ci dessous, la plupart des éléments qui les composent sont de taille très réduite et ont donc moins de chance d'être reliés avec le reste du texte :

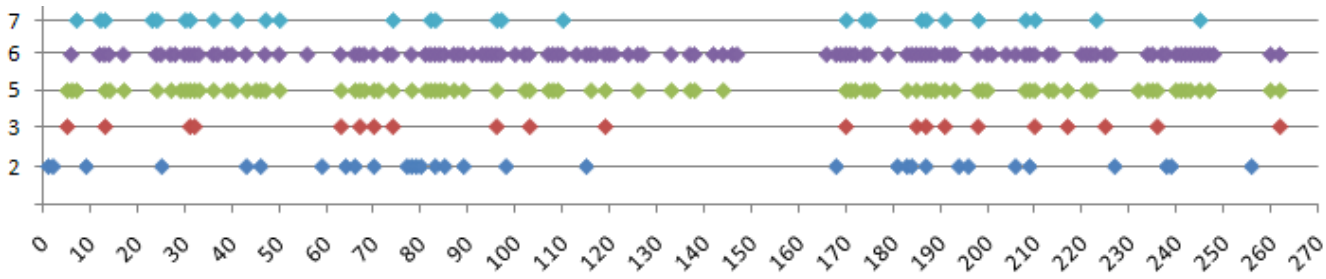

FIG. 4.4 – Répartition des phrases liées à chaque phrase du résumé introductif.

> **<s=52>**1.5. Structures
>
> **<s=53>** - Espace de Liberté : Piscine olympique, patinoire, bowling, cafétéria.
>
> **<s=54>** - Théâtre Scène Nationale
>
> **<s=55>** - Médiathèque - 9 salles de cinéma
>
> **<s=56>** - Palais des Sports, des Arts et du Travail
>
> [...]
>
> **<s=63>** - Pole universitaire de l'Université de Perpignan Via Domitia, comprenant de nombreuses formations juridiques : licence en Droit, IUT carrieres juridiques, IUP et Master Urbanisme-Immobilier, licence vini-viticole, capacité en Droit...

Certains items sont toutefois suffisamment volumineux pour être reliés à d'autres phrases du texte. La phrase 63, par exemple, se retrouve rapprochée de la phrase 214 :

> **<s=214>** Elle est également dotée d'une antenne Université de Perpignan (faculté de Droit et de Sciences économiques)et d'un IUT carrières juridiques.
>
> **Relations :** droit/droit, juridique/juridique, perpignan/perpignan, iut/iut, université/université, formation/science, nombreux/économique, licence/faculté, comprendre/doter

On ne peut saisir pleinement la pertinence de ce rapprochement que si l'on considère la structure dans laquelle apparaît la phrase 63. En effet, prise *hors cotexte*, cette phrase fait peu de sens puisqu'elle est dépourvue de verbe. Or, on peut considérer qu'un verbe est virtuellement présent dans cette phrase, et qu'il est fourni par la structure énumérative : étant donné que l'énumération est chapeautée par le titre *Structures*, on peut dans un premier temps paraphraser la phrase 63 par : "Le pôle universitaire de l'Université de Perpignan Via Domitia, qui comprend de nombreuses formations juridiques [...] **est** une structure". Et étant donné que le titre *1.5. Structures* fait lui-même partie d'une autre énumération qu'est l'ensemble de la titraille, subsumée par le titre de l'article, *Narbonne*, on peut encore développer la phrase 63 en : "La ville

de Narbonne possède des structures, dont le pôle universitaire de l'université de Perpignan Via Domitia qui comprend de nombreuses formations juridiques". Or, il est clair que cette phrase est une paraphrase parfaite de la phrase 214 (surtout si l'on développe le pronom anaphorique *elle* en *Narbonne*). Cette démarche d'expliciter au maximum les éléments d'une phrase pour qu'ils puissent être pris en compte lors du repérage des liens s'appelle, dans [Hoey, 1991] (p. 170, 172-173), une *expansion*. Et même s'il est indéniable qu'une opération comme l'expansion discursive, qui consiste à remplacer les pronoms anaphoriques par le nom ou le syntagme nominal qu'ils reprennent, pourrait améliorer le repérage des liens, cette démarche nous paraît, à l'heure actuelle, assez difficile à automatiser.

4.2.3 Conclusion

Le bilan que l'on peut faire au terme de ce chapitre est plutôt mitigé. Nous avons vu que l'automatisation de la méthode de Hoey nous avait permis de repérer des liens de discours entre des phrases situées à plus de 200 phrases d'intervalle, et en cela on peut considérer que notre expérience aborde un versant inexploré de la théorie de Hoey, celui des liens de longue distance. D'un autre côté, ces liens, et les liens pertinents en général, ne constituent qu'une minorité des liens repérés puisque la plupart ne peuvent pas être considérés comme porteurs d'une relation discursive analysable. Cela est dû au fait que ces liens s'appuient sur des relations lexicales trop souvent inappropriées. De plus, il est probable que le fait de projeter l'intégralité de la base lexicale sur le texte entier ait facilité l'apparition de ce type de relations.

Afin de vérifier cette dernière hypothèse, nous avons élaboré, dans le chapitre suivant, une expérience où les relations lexicales ont cette fois été projetées sur des sous-parties de l'article que nous venons d'étudier. Avec cette étude, nous sortons du cadre de la théorie des liens de Hoey et nous abordons la cohésion à travers le concept de *chaîne lexicale*. Cette étude s'inscrit également dans une problématique relative au discours : [Ho-Dac *et al.*, 2004], [Rebeyrolle et Jacques, 2006] et [Rebeyrolle *et al.*, 2009] ont en effet développé une typologie des titres de section en fonction de leurs rôles discursifs, et nous avons tenté de déterminer si le degré de cohésion des sections analysées était un facteur permettant de discriminer les différentes catégories de titres.

Chapitre 5

La détection de chaînes lexicales pour l'étude du rôle discursif des titres de section

Cette seconde étude découle du bilan en demi-teinte que nous avons dressé dans le chapitre précédent suite à l'analyse des résultats produits par notre système. Nous avons effectivement vu que l'automatisation de la méthode de Hoey avait posé plusieurs problèmes, dont celui, le plus important, de la ressource mobilisée. Cela nous a notamment conduit à devoir renoncer à tout un pan de la méthode d'analyse proposée par Hoey. L'approche que nous adoptons dans ce chapitre présente un terrain plus favorable à l'analyse des relations de cohésion. En effet, les relations stockées dans notre base lexicales ne sont pas ici projetées sur l'ensemble du texte mais contenues dans des sections, dont le degré de cohésion est mesuré à l'aide de chaînes lexicales. Le concept même de chaîne lexicale est intimement lié à une approche TAL, et il ne nécessite donc aucune adaptation de notre part.

Il est formalisé pour la première fois dans [Morris et Hirst, 1991][1] et désigne un ensemble de mots reliés par des relations qui engendrent de la cohésion lexicale. Les mots reliés les uns aux autres forment ainsi des réseaux plus ou moins étalés sur le texte qui permettent de faire émerger des zones textuelles particulièrement cohésives suceptibles de jouer un rôle au niveau organisationnel. C'est cette dimension organisationnelle qui nous intéresse le plus ici, puisque nous faisons l'hypothèse que le repérage des chaînes qui parcourent un texte (ici, une section) nous permettra d'inférer le rôle discursif du titre qui le chapeaute.

[1] Il est à noter que sans en faire d'analyse approfondie, [Brown et Yule, 1983] qualifient déjà, à la suite de [Halliday et Hasan, 1976] (p. 287,) de *chain of lexical collocation* des suites de mots comme {birdsong, woodlark, nightingale, blackbirds, rook} (p. 194).

5.1 Le concept de chaîne lexicale

5.1.1 L'algorithme de base

Dans l'étude de Morris & Hirst, la ressource utilisée pour repérer ces relations est le Roget's Thesaurus, qui a l'avantage de permettre la prise en compte des collocations (que les auteurs qualifient de "nonsystematic semantic relations"). Il est à noter que malgré le fait que le concept même de chaîne lexicale ait été défini en fonction des contraintes liées au traitement automatique, l'étude de Morris & Hirst a été réalisée *à la main*, pour la simple raison qu'aucune version électronique du thésaurus n'était alors disponible[2].

Le repérage de ces relations se fait à partir de *mots candidats*, à savoir les mots pleins, à l'exception de ceux qui ont une fréquence trop élevée comme *do* et *good*[3]. La démarche est la suivante : pour chaque mot candidat, l'algorithme cherche une chaîne à laquelle pourrait se rattacher le mot parmi les trois phrases le précédant. Si aucun lien entre le mot et les chaînes déjà présentes n'est trouvé, alors le mot devient le premier maillon d'une nouvelle chaîne. Dans cette approche, le type de relation qui relie les éléments d'une chaîne importe peu[4], mais le rattachement se fait sous certaines conditions.

La première, dite *de transitivité*, porte sur le nombre de liens dans le thésaurus qu'il est permis d'avoir entre un mot et les autres mots d'une chaîne. Le fait de définir un seuil de transitivité permet d'éviter que, dans la chaîne {vache, mouton, laine, écharpe, bottes, chapeau, neige}, *vache* ne se retrouve lié à *neige*. Dans leur expérience, Morris & Hirst ont intuitivement défini le seuil à une seule relation : si a est relié à b, et que b est relié à c, alors a, b, et c feront partie de la même chaîne, même si a et c n'ont aucun lien. Toutefois, dans le cas où c est seulement relié à d, d sera exclu.

Le deuxième paramètre à prendre en compte est celui de la distance permise entre deux mots d'une même chaîne. Chez Morris & Hirst, une chaîne s'interrompt si elle n'est plus alimentée pendant trois phrases consécutives. Si un élément de la chaîne apparaît au-delà, on a alors affaire à un *retour de chaîne* (*chain return*), qui peut apparaître après une digression, par exemple. Morris & Hirst avancent le fait que le potentiel de chaque chaîne à réapparaître dépend de sa

[2]Le Roget's Thesaurus a été numérisé par le Projet Gutenberg seulement neuf mois après la publication de l'article. Il est disponible à l'adresse suivante : http ://www.gutenberg.org/dirs/etext91/roget15a.txt Nous verrons que cette version se serait de toute façon révélée inexploitable en l'état.

[3]"[...] in general, the higher the frequency of a lexical item (its overall frequency in the system) the smaller the part it plays in lexical cohesion in texts." ([Halliday et Hasan, 1976] p. 290)

[4]La répétition semble toutefois avoir un statut particulier chez [Morris et Hirst, 1991]. L'indication "Reiterations – the more repetitions, the stronger the chain" reste cependant ambiguë étant donné que même selon leur classification (p. 21-22), la réitération englobe également la reprise par un hypéronyme ou un synonyme.

solidité (*strength*), qui varie selon trois critères :

- sa densité (le nombre de maillons qui la composent),
- le nombre de répétitions qu'elle contient,
- sa longueur (la quantité de texte sur laquelle elle s'étend).

Cette étude a été le point de départ d'un nombre important de travaux qui se sont appuyés sur le même algorithme en proposant d'autres critères pour la solidité des chaînes, en modifiant les valeurs des variables que sont la transitivité ou la distance maximale autorisée entre deux maillons, etc. La principale différence entre cette étude et la majorité des travaux qu'elle a engendrés réside dans le type de ressource mobilisé.

5.1.2 La question des ressources

Dans ce type d'approche, l'appariement des mots d'une chaîne se fait en fonction des relations présentes *a priori* dans une ressource de référence externe. Cela a pour effet de placer la-dite ressource au cœur du problème, puisque l'efficacité de la détection des chaînes est limitée par la couverture et la précision des données de référence. Ainsi, étant donné que la seule version numérique libre de droit du Roget's Thesaurus est celle de l'édition de 1911, les études qui se sont basées sur l'algorithme de Morris & Hirst se sont tournées vers WordNet pour éviter que la variation diachronique n'altère les résultats. De plus, il s'est avéré que la version électronique du thésaurus était dépourvue d'index, ce qui semble être un frein de taille à son exploitation par une machine (dans [Ellman, 2000], un index est généré grâce à un algorithme développé par l'auteur, ce qui a permis l'exploitation de la ressource pour le repérage de chaînes lexicales).

À l'opposé, l'utilisation de WordNet s'est révélée moins rebutante puisque le projet était alors en plein développement (et il l'est toujours) et que tout était fait pour permettre aux chercheurs de l'utiliser, notamment dans des applications de TAL. De plus, nous verrons que la structure en synsets du réseau fait qu'il est possible d'implémenter de façon relativement simple des modules de désambiguïsation. Sans oublier le fait que WordNet est libre de droits. Au final, il semblerait que le principal avantage du Roget's Thesaurus par rapport à WordNet réside dans le fait que le thésaurus permette la prise en compte de la relation de collocation : nous avons vu dans la section 2.2.2 que dans un thésaurus, un mot comme *chat* se trouvait relié à *miauler*, *ronronner*, etc. par une relation assez vague intitulée *relatif à*. Ainsi, concernant l'évaluation de leur algorithme (que nous décrivons ci-après), Hirst & St. Onge désignent WordNet comme une limite à la performance du système en prenant pour exemple une phrase où *child care* n'est pas relié à *school*, comme il aurait dû l'être :

> In fact, relations such as antonymy, holonymy, or meronymy are not appropriate to

link *child care* to *school*. Rather, the relation that exists between these two words is a **situational relation**. Many such relations are hard to classify (*e.g.*, the situational relation between *physician* and *hospital*), and it was a strength of Morris and Hirst's *Roget*-based algorithm that it was able to make such connections nonetheless. (*op. cit.* p. 13)

Ici encore, il semble que la question des ressources soit loin d'être résolue.

5.1.3 Évolutions du modèle

Hirst et St. Onge (1997)

En 1995, l'algorithme développé dans les travaux de Morris & Hirst a été adapté dans [Hirst et St. Onge, 1997] pour fonctionner à partir de données fournies par le réseau WordNet dans le but de corriger les *fautes de langue* (*malapropisms*). Le système visait à détecter les confusions entre mots dont la graphie est proche (des paronymes), qui ne sont pas repérées par les systèmes de correction traditionnels puisque l'erreur se situe au niveau sémantique. On peut prendre pour exemple en français la confusion entre *conjoncture* et *conjecture*. Dans une phrase contenant *conjecture économique*, l'algorithme repérera que *conjecture* n'appartient à aucune chaîne du co-texte immédiat, mais qu'un mot dont la graphie est proche, à savoir *conjoncture*, pourrait s'intégrer à une chaîne déjà formée, si tant est que le reste du texte ait un rapport avec le domaine de l'économie. Il proposera ainsi la correction à l'utilisateur. L'évaluation du système, basée sur un corpus d'articles de journaux de 322 645 mots parmi lesquels 1409 paronymes (insérés par les auteurs), a montré que seulement 24,8% des erreurs étaient détectées et correctement corrigées[5].

La principale évolution qu'a apporté le système de Hirst & St. Onge à l'algorithme de Morris & Hirst est la pondération des relations entre mots : la répétition est considérée comme *extra-forte* (*extra-strong*), les relations de synonymie, d'antonymie et d'hypéronymie sont considérées comme *fortes* (*strong*), et les relations qui relient deux mots qui ont entre deux et cinq liens d'écart dans le réseau (en suivant certains chemins autorisés) sont qualifiées de *semi-fortes* (*medium-strong*). Cette distinction se veut représentative de la proximité sémantique entre les deux mots reliés. Elle a une influence sur la taille de la fenêtre dans laquelle les mots vont se rattacher les uns aux autres :

1. une relation extra-forte peut être établie entre deux mots quelle que soit la distance qui les sépare,

[5] Se référer à [Hirst et Budanitsky, 2001] pour une étude visant à perfectionner le système de [Hirst et St. Onge, 1997].

2. une relation forte (synonymie, antonymie ou hypéronymie) peut porter sur deux mots séparés par moins de sept phrases,

3. une relation semi-forte n'a qu'une portée de trois phrases avant et après le mot.

Barzilay et Elhadad (1997)

Ce modèle a, à son tour, été amélioré dans [Barzilay et Elhadad, 1997] pour une application au résumé automatique, notamment grâce à l'implémentation d'un module de désambiguïsation s'appuyant sur un filtrage mutuel des synsets (l'acception d'un mot est définie selon les autres mots de la chaîne). Ces travaux présentent toutefois l'intérêt de proposer un second point de vue sur la mesure de la solidité des chaînes. Alors que chez [Morris et Hirst, 1991] elle se mesurait selon le nombre de répétitions et de lemmes qu'elle contenait ainsi que selon sa longueur, Barzilay & Elhadad ne retiennent que le critère de la longueur ainsi qu'une autre variable, l'*index d'homogénéité*, obtenu par le nombre de mots différents de la chaîne divisé par sa longueur. La multiplication des deux valeurs obtenues permet de distinguer les chaînes les plus solides, qui seront considérées comme les plus représentatives du contenu des textes : ne sont retenues que les chaînes dont le score est supérieur au score moyen auquel est ajouté le double de l'écart-type, soit la formule `Score(Chaîne) > Moyenne(Scores) + 2*Ecart-Type(Scores)`.

Une fois les chaînes *fortes* sélectionnées, l'étape suivante vise à extraire les phrases correspondant à ces chaînes. Parmi les trois méthodes envisagées, celle qui fonctionne le mieux consiste à sélectionner, pour chaque chaîne retenue, la phrase qui contient les premières occurrences des mots *représentatifs* de la chaîne (ce dernier critère étant fixé selon la fréquence des mots au sein de la chaîne).

L'évaluation a été menée à partir d'un ensemble de textes (des articles de journaux) résumés par des humains : pour un résumé à 10% du volume du texte original, la précision est de 61% et le rappel de 67%, alors que pour un résumé à 20% la précision chute à 47% et le rappel à 64%. Les auteurs comptent sur le recours à d'autres types de ressources pour améliorer ces scores, qui restent toutefois largement supérieurs à ceux obtenus par un système de résumé automatique commercial (la version de Microsoft Summarizer disponible sur Word 97), dont les scores de précision et de rappel oscillent entre 32% et 39%. Les autres limites du système résident principalement dans le fait que l'extraction de phrases entières plutôt que de segments plus ciblés peut se révéler préjudiciable, et dans le fait que les anaphores présentes dans les phrases extraites se retrouvent déconnectées.

Brunn, Chali et Pinchak (2001)

Cette étude, ainsi que la suivante, porte sur le résumé automatique et reprend dans leurs grandes lignes les algorithmes vus précédemment. Les travaux de [Brunn *et al.*, 2001] tirent leur originalité du fait qu'ils incluent une phase de pré-traitement dans laquelle le texte à résumer est traité par un segmenteur thématique et un analyseur syntaxique.

Le but de l'analyse syntaxique est double. Le premier est de repérer les syntagmes nominaux pour améliorer la couverture[6]. Elle permet ensuite de discerner les propositions principales des subordonnées. Cette distinction s'inscrit dans une démarche de filtrage des mots candidats qui s'appuie sur la thèse selon laquelle seuls les mots contenus dans les propositions principales méritent de figurer dans des chaînes[7].

La segmentation thématique est effectuée par le segmenteur dont le fonctionnement est décrit dans [Choi, 2000], et dont le principe repose sur la répétition lexicale. Cette segmentation trouve son utilité lors de la phase d'extraction des phrases pour le résumé : une fois les chaînes construites, les segments sont classés selon le nombre de membres de chaînes qu'ils contiennent, et les phrases sont extraites, toujours par un système d'addition assez basique, parmi ceux qui ont le score le plus élevé.

Silber et McCoy (2002)

Les améliorations qu'apporte le système de [Silber et McCoy, 2002] à celui qu'ont développé [Barzilay et Elhadad, 1997] consistent principalement à optimiser le stockage des données contenues dans WordNet ainsi que la façon dont sont constituées les chaînes afin de gagner en temps d'exécution et de permettre le traitement de grands documents. En effet, le système de Barzilay & Elhadad génère toutes les combinaisons de synsets possibles avant d'éliminer celles qui ne donnent pas lieu à une chaîne dont le nombre de maillons dépasse un certain seuil. Or, cette méthode est très coûteuse en espace de stockage et en temps d'exécution. De plus, l'exponentialité du nombre de combinaisons possibles implique une limite dans la taille des textes à traiter. Ainsi, l'idée de Silber & McCoy est de ne pas générer toutes les combinaisons possibles mais de construire une structure de données dans laquelle elles seront représentées virtuellement, les plus longues étant effectivement générées. Cette méthode permet de traiter un petit document en 4 secondes au lieu de 300 pour le système de [Barzilay et Elhadad, 1997].

De plus, cette étude introduit un système de calcul de la solidité des chaînes basé sur une pondération des relations qui lient leurs maillons : la répétition et la synonymie valent un point, ainsi que l'hypéronymie et la co-hyponymie, à l'exception près que ces deux dernières relations

[6]Cette fonction avait déjà été implémentée dans [Barzilay et Elhadad, 1997].

[7]Comme le fait remarquer [Stokes, 2004], l'efficacité de cette méthode n'a pas été démontrée.

voient leur score faiblir si le lien outrepasse les limites de la phrase.

Une autre spécificité des travaux de Silber et McCoy tient dans leur méthode d'évaluation. Comme l'indique le titre de l'article, leur méthode vise à évaluer les chaînes lexicales comme une méthode de représentation *intermédiaire* du texte, c'est-à-dire que l'extraction des phrases pour le résumé n'est ici pas prise en compte. Ainsi, l'évaluation ne va pas se faire entre deux résumés, le premier réalisé par un humain et le deuxième généré par le système, mais entre les chaînes détectées dans le document et l'ensemble des noms présents dans son résumé : les chaînes les plus solides sont sélectionnées (le calcul est celui de [Barzilay et Elhadad, 1997]), et l'on calcule le pourcentage de noms présents dans le résumé qui se retrouvent dans ses chaînes. En moyenne, c'est le cas pour 80,83% d'entre eux, ce qui donne une réponse assez claire à la question de savoir si les chaînes lexicales constituent de bonnes représentations du contenu d'un texte[8].

5.2 Titres de section et organisation discursive

Dans l'étude développée au chapitre précédent, l'une des conclusions à laquelle nous étions parvenu était que la projection des relations lexicales aurait probablement gagné à être un tant soit peu cadrée. Nous avions en effet choisi de coller au mieux à la méthode de Hoey afin d'obtenir des résultats similaires et nous avions ainsi projeté notre ressource sur l'ensemble du texte, mais force est de constater qu'au vu des données dont nous disposons, l'étude des phénomènes de cohésion ne peut se faire que dans des environnements beaucoup mieux maîtrisables. Ainsi, au lieu de prendre le texte comme unité d'analyse, la présente étude se focalise sur l'entité *section*. De plus, l'analyse de la cohésion des sections se fait ici vis-à-vis du titre qui les chapeaute

Cette dernière restriction est doublement motivée. Alors que dans l'expérience précédente, la profusion de liens nous avait poussé à centrer notre analyse sur l'objet *résumé introductif*, la démarche qui consiste à analyser exclusivement les relations de cohésion en lien avec les titres de section peut être vue comme une volonté de garder le contrôle sur tous les éléments du protocole. Cette démarche est d'ailleurs évoquée dans [Benbrahim, 1996] :

> However, relying fully on the thesaurus and limiting the distance between words of a chain is not free of difficulties particularly in cases where lexical chains span across paragraphs and therefore creating overlaps with other chains. One can think of solutions to these problems by assuming that a chain will only span one para-

[8]Cette notion de représentativité a été utilisée en recherche d'information dans [Ngai et Holliman, 2002] pour rapprocher les documents présentant des configurations de chaînes similaires.

> graph at a time, or alternatively, **one can take into account the words that appear in headers of paragraphs or sections as pilots for buildings chains**. (p. 24) [nous soulignons]

Mais il ne s'agit toutefois pas ici simplement de se rabattre sur un élément du texte pour réguler la masse de phénomènes à analyser. Le choix des mots du titre comme *pilotes* pour le repérage des chaînes lexicales n'est pas anodin : il est motivé par le fait que les titres de section sont porteurs d'une valeur discursive et que la mise en relation de cette dimension organisationnelle avec les propriétés lexicales des sections est, à notre avis, susceptible de laisser apparaître des phénomènes mesurables de co-dépendance.

Parmi les rares travaux qui se sont penchés sur la question, on peut citer [Morley, 2006], qui a analysé les liens de cohésion lexicale entre les titres d'articles de journaux et le texte qu'ils chapeautent afin de faire émerger des structures discursives. Une des conclusions de cette étude est que :

- les mots contenus dans les titres d'articles de journaux sont de bons indicateurs de la (ou *les)* thématique qui sera développée dans l'article,
- certains mots qui apparaissent dans les titres peuvent être porteurs d'une valeur rhétorique représentative de l'organisation du texte chapeauté : par exemple, un titre comme "Choice moment" implique qu'il est probable que le texte contienne des structures de mise en opposition (puisqu'il est question de choix).

En ce qui nous concerne, nous nous situons dans le cadre théorique fourni par les travaux de [Ho-Dac *et al.*, 2004], [Rebeyrolle et Jacques, 2006] et [Rebeyrolle *et al.*, 2009], qui gravitent autour de la question du rôle discursif des titres de section (ou *intertitres*). La démarche adoptée dans ces études a été d'analyser la façon dont les titres de section étaient repris dans les textes chapeautés afin de faire émerger des profils et voir s'ils correspondent à un certain type de fonctionnement discursif. Chaque reprise se définit selon un certain nombre de traits qui ont été repérés et annotés manuellement sur un corpus composé de textes divers (des articles scientifiques et des documents de travail). Ces traits peuvent porter sur la forme du titre (type de syntagme, présence de coordination...), sur les conditions de reprise du titre (dans son intégralité, en partie, à travers un pronom...), sur l'endroit du texte où il est repris (dans la première phrase ou plus loin), etc. Une analyse statistique (dite *multifactorielle*) a permis de dégager deux regroupements de traits majeurs :

> A. des intertitres ayant la forme d'un syntagme nominal, donnant lieu à une reprise unique et immédiate (dans la première phrase), en position sujet, soit sous la forme d'une répétition strictement identique du titre, soit par anaphore pronominale – phénomène "extrême", rappelons-le ;

> B. des intertitres ayant la forme de syntagmes nominaux présentant une partition interne (avec, par exemple, une coordination, deux-points. . .), de syntagmes prépositionnels ou verbaux ou encore de phrases complètes, donnant lieu à une répétition multiple, dans la section mais pas dans la première phrase, d'une partie ou de la totalité du lexique du titre, en le déconstruisant, avec une fonction syntaxique autre que la fonction sujet. ([Rebeyrolle *et al.*, 2009] p. 10)

Ces faisceaux de traits correspondent à deux *types fonctionnels* distincts. Dans le premier cas, on parle d'*implication référentielle* : le rôle de l'intertitre est d'introduire un nouveau référent ou d'en reprendre un qui a déjà été évoqué plus tôt dans le texte. Dans le deuxième cas, on parle d'*implication thématique* : "l'intertitre ouvre un espace thématique qui est ensuite développé dans la section" (*id.*)

Il existe un troisième type fonctionnel qui se distingue de ceux que nous venons de décrire. Les titres à implication zéro ne sont en effet quasiment jamais repris dans le texte chapeauté. La raison en est que leur fonction se limite à porter des éléments *méta* qui opèrent au niveau "textuo-organisationnel" ([Ho-Dac *et al.*, 2004] p. 11), comme *Introduction*, *Conclusion*, *Partie 1*, etc.

[Rebeyrolle et Jacques, 2006] (p. 5) évoquent les phénomènes de reprise de titre comme relevant du domaine de la cohésion. En effet, dans cette étude, ne sont prises en compte que les reprises qui se font soit sous la forme d'une répétition simple (partielle ou intégrale), soit à l'aide de mots qui sont des conversions des mots du texte (*représentation/représentée*, *analyse/analyser*), soit à l'aide d'une reprise pronominale. Les relations sémantiques (et en particulier la collocation) sont exclues du champ de la répétition.

Or, nous avons vu que les titres à implication référentielle se définissaient par le fait qu'ils développaient tout au long de la section un cadre thématique. Nous pensons qu'à l'aide des ressources dont nous disposons, cette caractéristique peut être implémentée dans un système de repérage de chaînes lexicales : la mesure de la couverture des chaînes dans une section pourrait en effet être un bon révélateur de la nature du titre puisque, si l'on en croit cette typologie, des chaînes qui s'étendent d'un bout à l'autre de la section auraient de bonnes chances d'être révélatrices d'un titre de type *thématique*.

5.3 Projection des mots reliés aux titres de section

5.3.1 Sélection des mots candidats

Nous avons indiqué plus haut qu'afin de garder un certain contrôle sur la façon dont allaient être dégagées les chaînes, nous allions nous servir des mots contenus dans le titre de la section

analysée. Étant donné que nous avons conclu dans la dernière expérience que l'utilisation de la répétition simple était insuffisante pour générer un nombre satisfaisant de relations, nous nous sommes directement lancé dans une projection qui, en plus de mobiliser la répétition simple, inclut l'ensemble des mots de notre base lexicale qui sont reliés à chaque mot du titre. De plus, puisque les textes à analyser sont d'une taille très réduite et qu'il y a peu de chance qu'ils présentent une trop grande variété thématique, nous avons choisi de nous servir de la ressource dans son état original, c'est-à-dire avant qu'elle n'ait subi les étapes de filtrage et de pondération que nous avons opérées à la section 4.2.

Le premier texte que nous avons traité est la section *1.5.1. Logement* (lignes 64 à 75). La projection des répétitions du mots *logement* ainsi que des mots qui lui sont reliés dans la base donne les résultats visible à la figure 5.1. On constate que sur les 20 relations repérées (dont quand même 10 de répétition simple), toutes sont pertinentes. On constate toutefois que certains mots comme *construction* ou *habitant* qui auraient pu être rattachés à *logement* ne l'ont pas été. On peut ainsi tenter d'accroître la couverture de notre système en opérant une projection de deuxième niveau : alors que, dans un premier temps, nous nous sommes servis du titre comme pilote pour repérer les mots candidats à la formation de chaînes, nous allons maintenant nous servir de ces mots candidats comme pilotes en projetant sur le texte l'ensemble des mots qui sont reliés à ces mots candidats dans notre base. En effet, il s'avère que la prise en compte des mots reliés aux mots reliés aux mots du titre permet de récupérer un certain nombre de mots tout à fait pertinents, ce qui aurait été inconcevable si l'analyse avait porté sur l'ensemble du texte. La figure 5.2 illustre ce phénomène. Nous y avons représenté les différents types de mots candidats que nous envisageons d'utiliser :

- le nœud bleuté correspond au pilote originel (le mot-titre *logement*),
- les nœuds grisés correspondent aux mots directement reliés au pilote originel (ils ont été mobilisés lors de la projection illustrée à la figure 5.1),
- les nœuds incolores sont les mots du texte qui ne sont pas directement reliés à *logement*, mais qui le sont via un mot qui lui est relié.

Nous avons vu que le fait d'utiliser les mots directement reliés aux mots du titre donnait des résultats très satisfaisants, malgré le fait que certains mots pertinents n'étaient pas détectés. Or, il s'avère que les mots indirectement reliés à *logement*, que nous avons représentés par des nœuds incolores, permettent de récupérer une certaine partie des ces mots non-retrouvés. Cependant, on peut constater que tous ces mots ne sont pas pertinents : *espace*, *petit* et *article*, par exemple, ne devraient pas être pris en compte alors que *pièce*, *construction* et *propriétaire* font clairement partie du même champ lexical que *logement*. Un critère permet toutefois de distinguer ces deux types de mots. Nous avons en effet constaté qu'il était possible de conserver

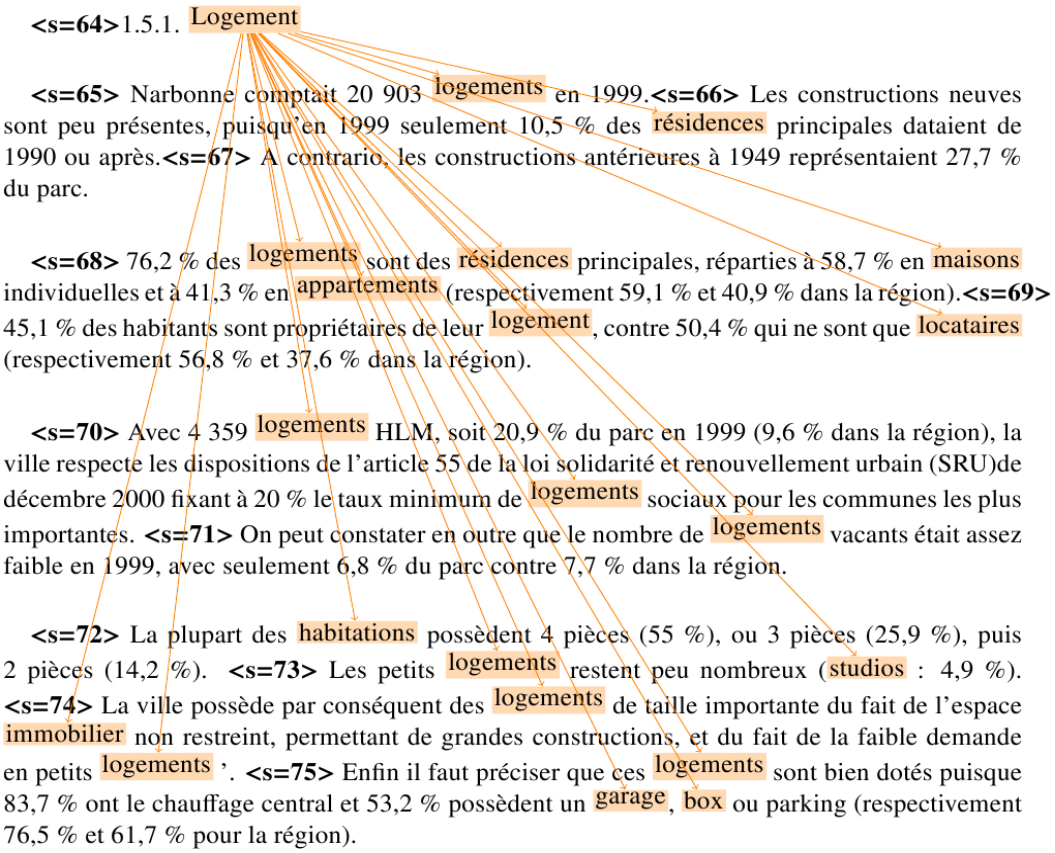

FIG. 5.1 – Mots reliés au titre *Logement* dans la section éponyme de l'article *Narbonne*.

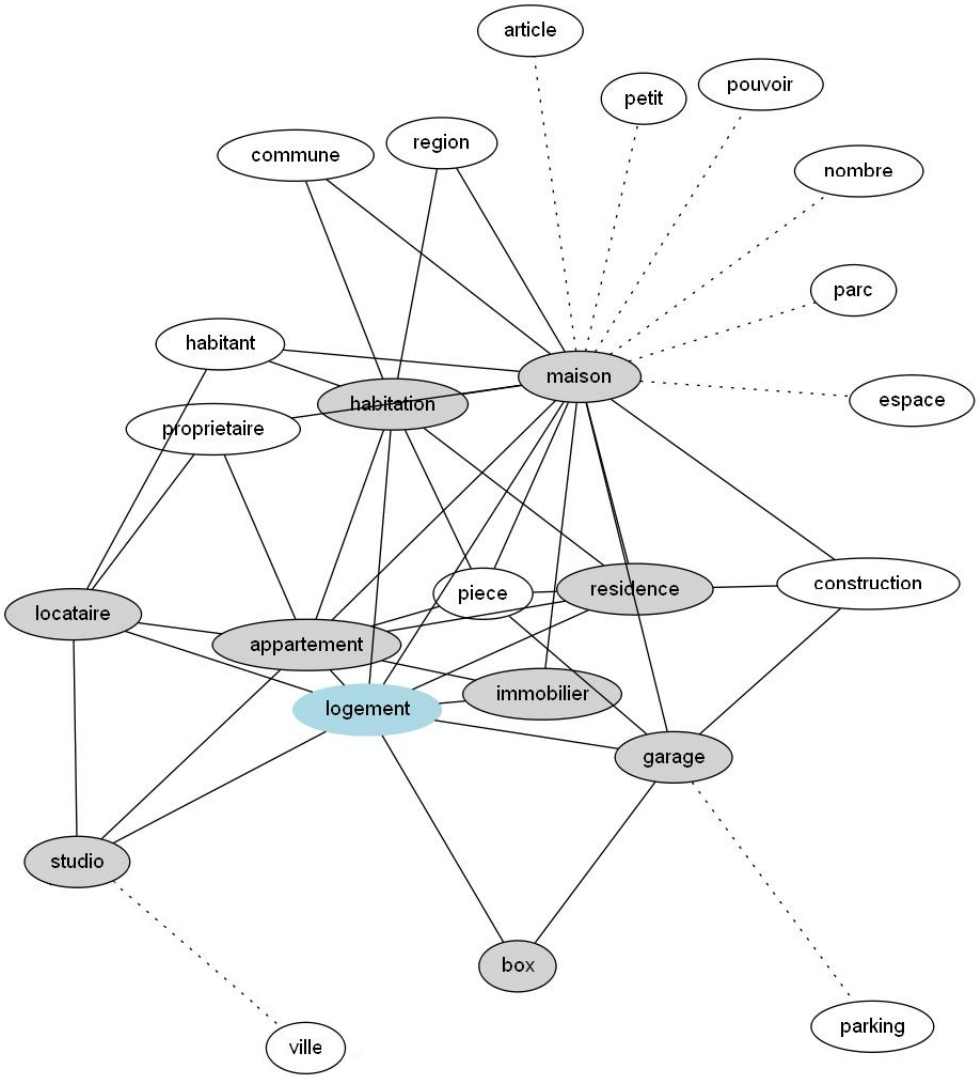

FIG. 5.2 – Réseau des mots directement ou indirectement reliés au mot *logement* dans la section éponyme.

la plupart des mots pertinents en excluant les mots qui n'étaient reliés à leur pilote que par un seul lien. On peut constater que parmi ces huit mots, qui, dans la figure 5.2 sont reliés au reste du réseau par des arêtes en pointillés, cinq peuvent être clairement qualifiés de non pertinents (*article*, *petit*, *pouvoir*, *nombre* et *espace*). Les trois autres, *parc*, *parking* et *ville* auraient été pertinents, mais ne sont pas pris en compte. Toutefois, il est clair que sur un texte de cette taille, le fait de mobiliser les mots reliés à au moins deux mots qui sont directement reliés au titre ne peut que rendre le système plus efficace : parmi les six mots supplémentaires qu'a permis de rapporter cette méthode, tous peuvent être directement reliés à *logement* (seul *région*, éventuellement, pourrait être écarté). Ces liens ont été représentés en bleu à la figure 5.3. Leur prise en compte fait passer le nombre de mots candidats de 20 à 34, ce qui correspond à une augmentation de 59%. Cet effet bénéfique peut également s'observer sur un texte à peine plus petit (dix lignes), mais aux phrases plus longues, à savoir la section *4.1. Tendances politiques*. On peut constater à la figure 5.4 que des mots comme *voter*, *candidat* ou *élire* pourraient être clairement rattachés à *politique*, mais la proportion de mots non pertinents parmi les mots relevant d'une relation de deuxième niveau (et reliés à au moins deux mots) est plus élevée que dans l'exemple précédent. De plus, force est de constater que cette opération se révèle beaucoup moins efficace (voire néfaste) dès que la taille du texte augmente et qu'interviennent des mots ayant un très grand nombre de relations. Nous avons représenté à la figure 5.5 le réseau de mots généré sur la section *1.1.3. Voies de communication et transport* (19 lignes) à partir du mot *transport*. On peut constater que les mots représentés par des nœuds incolores et qui sont reliés à au moins deux autres mots sont, pour la plupart, non pertinents (comme c'est le cas pour l'intégralité des mots qui ne sont liés au réseau que par une seule relation). De plus, le mot *ville*, de par son grand nombre de liens, permet l'activation de huit mots, qui, sans ça, n'auraient été reliés au réseau que par un seul lien et n'auraient donc pas été pris en compte (ce qui aurait été plutôt positif, puisqu'un seul de ces mots, *destination*, peut éventuellement être considéré comme sémantiquement proche de *transport*).

Les sections d'article ayant généralement une taille supérieure à une dizaine de lignes, nous jouons la carte de la prudence et décidons d'en rester aux relations de premier niveau, c'est-à-dire aux mots qui sont directement reliés aux mots contenus dans les titres. Toutefois, il nous semble que cette propriété des sections très courtes de permettre, dans certains cas, la création de liens indirects (comme on peut en trouver dans WordNet, par exemple), peut se révéler une piste intéressante pour l'analyse de certains types de textes.

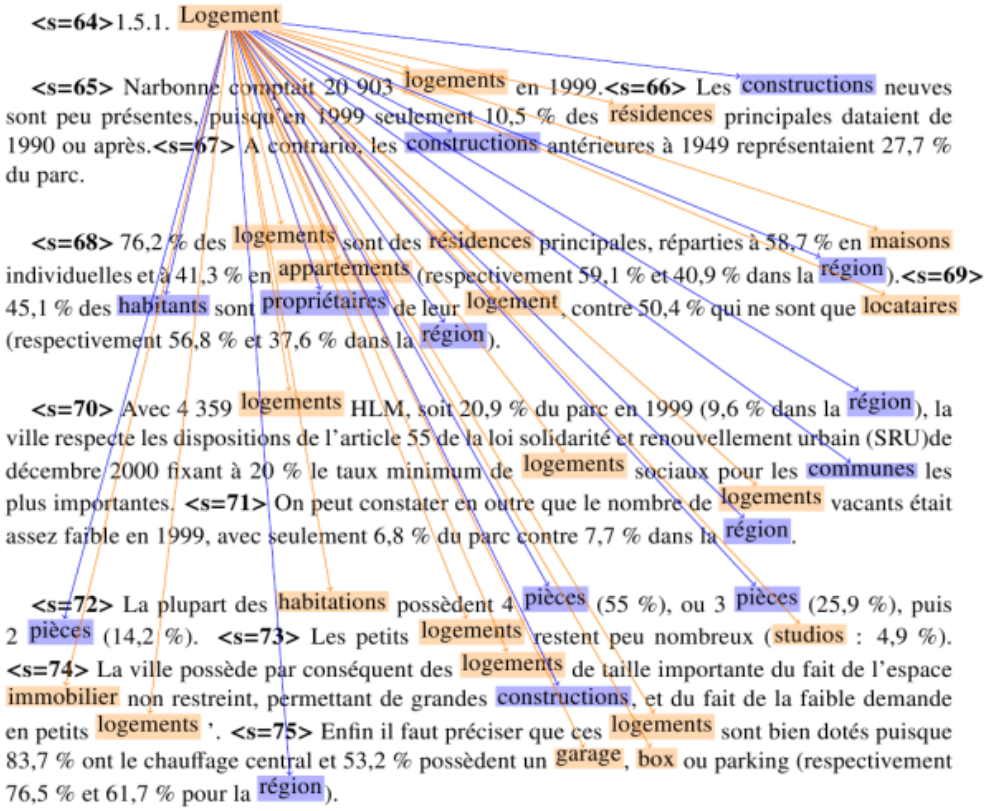

FIG. 5.3 – Mots reliés directement (en orange) et indirectement (en bleu) au titre *Logement* dans la section éponyme de l'article *Narbonne*.

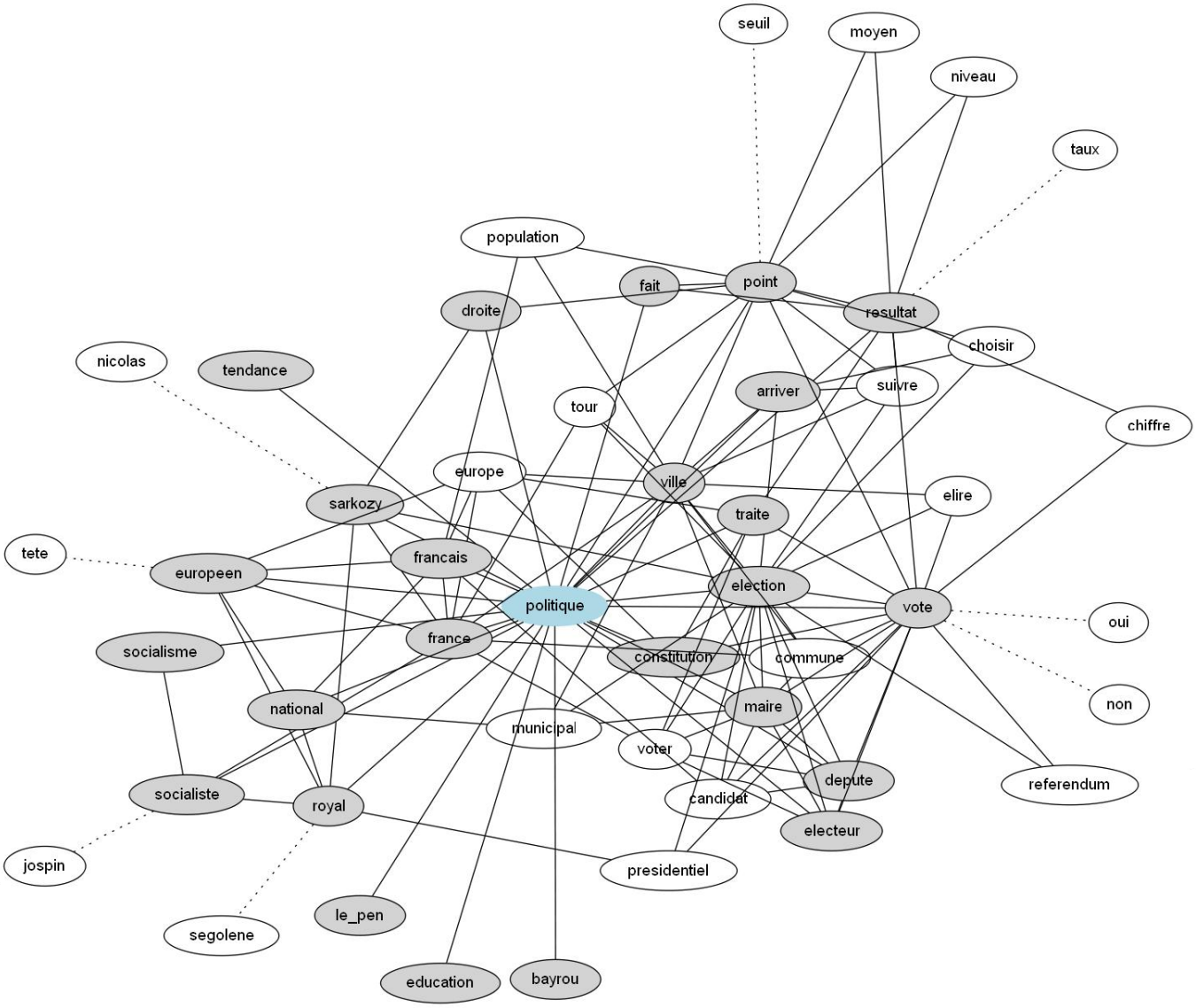

FIG. 5.4 – Réseau des mots directement ou indirectement reliés au mot *politique* dans la section *4.1. Tendances politiques*.

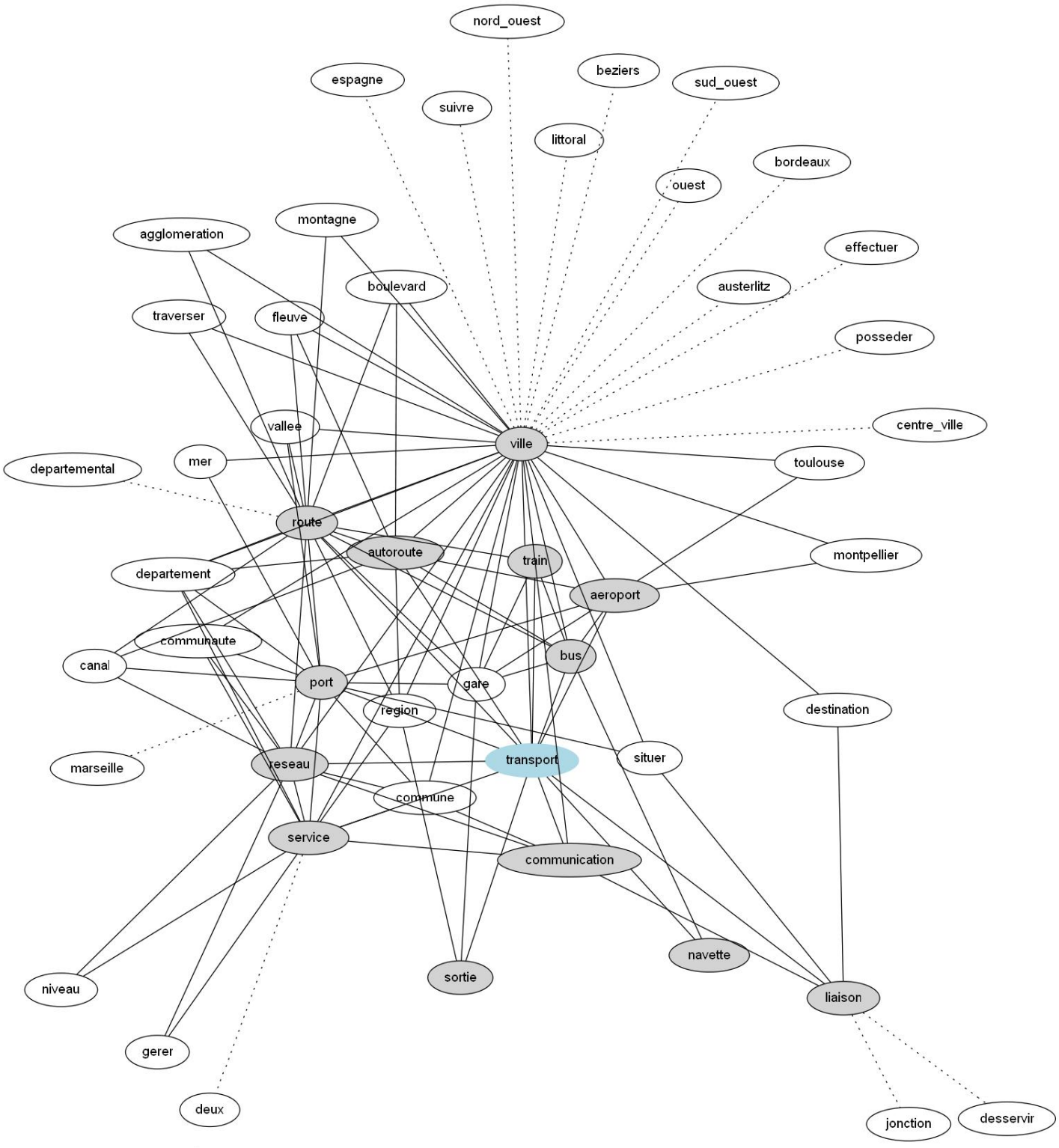

FIG. 5.5 – Réseau des mots directement ou indirectement reliés au mot *transport* dans la section *1.1.3. Voies de communication et transport*.

5.3.2 Mesure de l'étendue des chaînes

L'algorithme que nous avons développé reprend en partie les mécanismes du modèle de [Morris et Hirst, 1991]. Il est relativement simple : dès qu'il repère un des mots candidats, que nous avons définis comme étant les mots directement liés aux mots du texte, le programme va balayer les quinze mots[9] qui le suivent à la recherche de l'un des mots avec lesquels il est lié dans la base lexicale. Si un mot est trouvé, ce dernier est intégré à la chaîne et la recherche se poursuit sur les quinze prochains mots. Si aucun mot n'est trouvé, alors la chaîne meurt.

Voici par exemple les chaînes que retrouve notre algorithme sur la section *9. Éducation* (lignes 211 à 217) :

19 lycée	33 science	43 étudiant
29 université	36 carrière	**Longueur : 36 mots**
31 faculté	39 étudiant	
Longueur : 12 mots	41 campus	5 école
	43 étudiant	7 collège
17 institution	48 cafétéria	13 ferry
19 lycée	49 universitaire	16 cité
29 université	53 universitaire	17 institution
32 droit	**Longueur : 24 mots**	19 lycée
33 science		29 université
36 carrière	7 collège	31 faculté
50 projet	16 cité	32 droit
52 installation	17 institution	33 science
Longueur : 35 mots	19 lycée	39 étudiant
	29 université	43 étudiant
29 université	31 faculté	50 projet
31 faculté	33 science	51 développement
32 droit	39 étudiant	**Longueur : 46 mots**

Pour finir, le programme assigne deux scores à la section analysée qui correspondent :

- au rapport entre le nombre de mots couverts par l'ensemble des chaînes (éventuellement divisé par le nombre de mots du titre) et l'ensemble des mots de la section (ici 2,83),
- au calcul de la longueur moyenne des chaînes (elle est ici de 30,6 mots).

[9] Vue la nature des textes étudiés, il était en effet inconcevable de garder la fenêtre de trois phrases utilisée chez Morris et Hirst. La fenêtre de quinze mots a été fixée de façon empirique.

5.3.3 Mesure du rapport entre longueur des chaînes et types fonctionnels

Nous avons ainsi établi deux indices nous apportant des informations sur la nature et la longueur des chaînes lexicales au sein d'une section. Nous allons maintenant les mettre en parallèle avec l'un des indices qui a permis d'établir la typologie présentée en 5.2, à savoir la reprise du titre en position sujet dans la première phrase.

Nous avons dans un premier temps constitué deux corpus de cinquante titres/sections chacun extrait d'un ensemble d'articles Wikipédia portant sur des sujets variés comme l'astronomie, l'histoire, la politique, l'économie, etc. (ces articles ont pour seul point commun de figurer dans la rubrique des articles dits "de qualité") :

– dans le premier cas, le titre est repris en position sujet dans la première phrase :

> **9. Science de l' Égypte antique**
>
> La science de l' Égypte antique jouissait d' un grand prestige dans les temps anciens.
>
> **3. Le mouvement minimaliste**
>
> Le mouvement minimaliste apparaît dans les années 1960 avec les travaux de La Monte Young et Terry Riley, considérés comme les précurseurs.

– dans le deuxième cas, le titre n'est pas immédiatement repris (il peut éventuellement apparaître au-delà de la première phrase).

Ces données ont été recueillies de façon semi-automatique. Étant donné que nous n'avons pas mobilisé d'analyseur syntaxique, nous avons dans un premier temps récupéré un ensemble de sections dont la première phrase contenait tous les mots du titre avant le premier verbe, et un ensemble de sections où le titre n'apparaissait pas dans la première phrase. Il nous a ensuite fallu procéder à un nettoyage manuel. Étant donné que l'expérience implique une mesure de la portée des chaînes lexicales, nous avons exclu les couples titre/section dont le titre ne pouvait pas donner lieu à un repérage des chaînes, soit parce qu'il était constitué d'un ou plusieurs noms propres, soit parce que les mots qu'il contenait n'avaient aucune chance de figurer dans notre ressource et donc de permettre le repérage de relations (*8. Mégatsunamis*, *4.6. La baryogénèse*...). Nous avons également procédé à une normalisation de la taille des sections : nous avons fait en sorte que la taille de chaque section se situe aux alentours de 80 mots (quitte à amputer certaines sections d'un paragraphe ou deux).

Nous avons précédemment émis l'hypothèse selon laquelle la couverture des chaînes se révèlerait probablement plus importante dans le deuxième type de sections, puisque ces dernières ne présentent pas de reprise immédiate du titre, ce qui est un trait caractéristique des titres à implication thématique. Or, les résultats montrent que ce n'est pas le cas : le rapport entre le

nombre de mots couverts par l'ensemble des chaînes et le nombre total de mots de la section est en moyenne de 3,43 pour les sections où le titre est immédiatemment repris, et de 2,41 pour les autres. De même, la longueur moyenne des chaînes est de 21,75 mots pour le premier type de section et de 17,01 pour le second. Ces résultats sont toutefois à prendre avec précaution : il serait un peu prématuré de tirer des conclusions sur l'importance de la cohésion lexicale comme trait pour la définition des types fonctionnels. En effet, nous avons vu qu'un type se définissait par un ensemble de traits, et nous n'avons ici comparé notre taux de couverture des chaînes qu'avec un seul de ces traits, à savoir la reprise en position sujet. De plus, il y a tout un ensemble de variables qu'il reste à prendre en compte comme le type de texte : la typologie décrite précédemment a été établie à la suite de l'étude d'un corpus d'articles scientifiques et de documents de travail, qui sont des types de texte complètement différents de l'article encyclopédique (en particulier dans le sens où ils sont destinés à un public relativement restreint et spécialisé). On peut également remettre en question certains points de la méthodologie que nous avons adoptée :

- les corpus de titres/sections sont plutôt restreints et il est difficile d'établir des statistiques fiables sur une cinquantaine de textes. Toutefois, la mobilisation d'outils de TAL comme un analyseur syntaxique pourrait facilement permettre de constituer des corpus fiables sans avoir systématiquement recours à la vérification manuelle,
- les mesures de couverture des chaînes lexicales pourraient également être améliorées et diversifiées, dans le sens où nous avons vu que les chaînes pouvaient également être envisagées sous l'angle du type de relations qui les construisent (on peut éventuellement envisager que si les sections chapeautées par un titre à implication référentielle laissent apparaître des chaînes lexicales, alors leur concentration en répétitions simples pourra être plus élevée que celle des chaînes qui apparaissent dans les sections thématiques).

Concernant les titres à implication zéro, nous avons trouvé, dans notre corpus de cinquante sections où le titre n'était pas repris, six couples titres/section où le titre n'était à l'origine d'aucun lien avec les mots du texte. On peut trouver plusieurs raisons à cela :

- le titre ne sert qu'à indiquer la structure du document (et non celle du discours qui y est développé). Ces titres, ici *8. Notes*, sont les titres à implication zéro définis dans [Ho-Dac *et al.*, 2004],
- les mots du titre sont absents de la base lexicale, soit parce que ce sont des noms propres[10], soit parce qu'ils relèvent d'un domaine de spécialité, comme dans *4.3. La nucléosynthèse primordiale (+ 3 minutes)*,
- le titre est trop lié au titre auquel il est subordonné pour être traité de façon isolée. Ici, le

[10]Notre base en contient cependant un certain nombre.

titre *3.1. Préparatifs* ne donne lieu à aucune relation dans le texte puisque son sens est extrêmement vague. Il faut remonter au niveau supérieur, *3. L'assassinat* pour avoir une idée des thèmes qui seront abordés dans la section *3.1. Préparatifs*,

– le titre est une métaphore, comme dans la section *6.2. La rupture*, où il est question d'un désaccord entre un premier ministre et son gouvernement.

5.3.4 Conclusion

Nous avons présenté dans ce chapitre une deuxième façon de se livrer au repérage des relations lexicales dans un but qui reste le même : accéder à des fonctionnements discurifs. Le fonctionnement auquel nous nous sommes ici intéressé est celui des intertitres, ou *titres de section*. Nous avons proposé l'hypothèse selon laquelle l'étude des propriétés cohésives des sections, via le repérage de chaînes, pouvait apporter des informations quant à la fonction discursive des titres. Nous avons essayé de trouver une correlation entre le critère de la cohésion lexicale et celui de la reprise du titre dans la première phrase du texte. Les résultats se sont avérés être un peu en décalage par rapport à ce que l'on pouvait envisager au vu de la typologie développée dans [Ho-Dac *et al.*, 2004], mais il faut garder à l'esprit que nous n'avons fait que comparer l'indice qu'est la cohésion au critère de la reprise en position sujet, qui n'est que l'un des traits qui permet de caractériser les types fonctionnels (nous n'avons pas ici pris en compte la forme des titres).

L'expérience que nous avons menée a été l'occasion d'observer la façon dont notre ressource s'adaptait au changement d'échelle : nous avons en effet constaté, sans grande surprise, que son efficacité augmentait quand la taille des textes sur laquelle elle est projetée diminuait.

En guise de perspective, il nous semble intéressant d'envisager cette démarche de typologisation des titres en fonction avec leur place dans la hiérarchie (la *titraille*). En effet, un certain nombre d'intertitres n'ont pas d'autonomie et ne peuvent être interprétés que s'ils sont mis en rapport avec le titre auquel ils sont subordonnés. C'est notamment le cas dans l'exemple suivant :

14. Le jansénisme hors de France
    14.1. À Louvain
    14.2. En Hollande
    14.3. En Italie

Comme nous l'avons constaté lors de notre repérage des titres à inmplication zéro, ces titres sont indissociables du titre qui leur est immédiatement supérieur, dans le sens où ils en constituent des élipses : les titres 14.1., 14.2. et 14.3. sont à interpréter comme *Le jansénisme à*

Louvain, *Le jansénisme en Hollande*, etc., Louvain, Hollande et Italie étant des hyponymes de la catégorie créée par le syntagme *hors de France*. Il est noter que nous avions rencontré un phénomène similaire lors de notre expérience précédente (p. 80).

Conclusion générale

Alors que cette étude touche à sa fin, on ne peut s'empêcher de penser qu'au vu des résultats que nous avons obtenus aux chapitres 4 et 5, elle pose probablement plus de questions qu'elle ne donne de réponses. En effet, nous avons, dans ces deux chapitres, mené des expériences qui, dans les deux cas, ont donné des résultats assez contrastés. C'est en particulier le cas pour celle que nous avons menée au chapitre 4, où l'automatisation de la méthode de M. Hoey a certes permis de faire apparaître des liens de longue distance à partir du repérage de relations lexicales, mais a surtout généré une quantité de bruit qui nous laisse dubitatif quant à l'efficacité du système. De même, au chapitre suivant, la thèse selon laquelle les chaînes lexicales les plus longues se mesureraient sur des textes où le titre ne serait pas repris dans la première phrase en position sujet n'a pas pu être prouvée. On peut se demander ici si le postulat de départ était faussé, si la méthodologie employée contenait des biais, etc.

Dans tous les cas, ces expériences ont été l'occasion de faire des observations sur la ressource employée et les relations lexicales en général : nous avions déduit, au terme du chapitre 4, qu'une des raisons pour lesquelles une grande partie des relations repérées n'étaient pas pertinentes était que la projection de la ressource avait été effectuée sur un texte entier, ce qui laissait le champ libre à la formation d'un certain nombre de relations fantaisistes. Nous avons donc, dans le chapitre suivant, réduit la taille des textes analysés, et il s'est effectivement avéré que dans des contextes extrêmement restreints comme les sections d'article d'une dizaine de lignes, la plupart des relations repérées étaient pertinentes. Il est toutefois évident que nous ne pouvons pas nous satisfaire de tels résultats, puisque nous nous situons ici dans une approche discursive de la cohésion et que le nombre de phénomènes discursifs observables sur des textes de dix lignes nous paraît assez restreint. Ainsi, le premier constat que l'on peut formuler au terme de ce mémoire est qu'il y a encore un énorme travail à faire sur le filtrage des données lexicales avant d'envisager des études en lien avec le discours : nous avons en effet vu au chapitre 4 que l'impossibilité de générer des liens fiables nous obligeait à devoir renoncer aux analyses du texte développées dans [Hoey, 1991] comme le repérage des ouvertures et fermetures thématiques ou celui des phrase centrales.

Un autre élément qui aurait gagné à être plus exploré est celui des genres discursifs. Plu-

sieurs d'entre eux ont été évoqués tout au long de ces cent pages : la brève, l'essai, l'article de journal, le texte scientifique, etc. Ces genres sont clairement définis, et il est probable que le rapport entre cohésion et structure soit de nature différente chez chacun d'entre eux. Dans notre cas, la principale motivation qui nous a poussé à prendre un article de Wikipédia comme texte de travail a été le fait que les voisins ont été calculés sur un corpus composé d'articles de ce type, et que cela nous aurait permis d'avoir dans notre base un certain nombre de relations dont on aurait pu espérer qu'elles se montreraient plus pertinentes que les autres (et cela a d'ailleurs peut-être été le cas). Toutefois, nous avons fait l'erreur de ne considérer qu'un seul genre *Texte encyclopédique*, ce qui nous a, dans un premier temps, mis sur une piste que nous avons par la suite considérée comme infructueuse (p. 46) : il s'avère en effet que les articles biographiques et les articles de villes présentent des propriétés cohésives totalement différentes et qu'il y a donc des chances pour que leur rapport à la structure varie en conséquence.

Au final, il ressort de ce mémoire qu'au vu des ressources dont nous disposons, l'utilisation de données lexicales pour l'observation de phénomènes discursifs ne pourra se faire que dans des environnements plus cadrés, mieux définis, et avec des ressources beaucoup mieux maîtrisées.

Annexes

A.1 Texte de l'article *Narbonne*

<s=1>0. Narbonne

<s=2> Narbonne (en occitan Narbona) est une commune française, située dans le département de l'Aude et la région Languedoc-Roussillon. **<s=3>** La commune est traversée par le canal de la Robine, classé au Patrimoine mondial de l'humanité par l'UNESCO depuis 1996. **<s=4>** Elle est classée 120e ville de France en termes de population.

<s=5> Située au cœur du 51e parc naturel de France « parc naturel régional de la Narbonnaise en Méditerranée », Narbonne possède également d'autres sites naturels classés, comme le massif de la Clape et celui de l'abbaye Sainte-Marie de Fontfroide, ainsi que les étangs de Bages-Sigean. **<s=6>** Fondée par les Romains en 118 av. J. C., elle était leur plus ancienne colonie en Gaule et son centre urbain garde trace de nombreux siècles d'histoire (cathédrale Saint-Just-et-Saint-Pasteur, palais des Archevêques, restes de la voie Domitienne). **<s=7>** La ville est entourée d'un environnement exceptionnel fait de garrigues et de vignes ; proche du littoral d'une région très touristique, elle possède une plage de cinq kilomètres de sable fin à Narbonne-Plage.

<s=8> Les habitants de Narbonne sont appelés les Narbonnais.

<s=9>1. Géographie

<s=10>1.1. Situation

<s=11>1.1.1. Localisation

<s=12> Narbonne est située dans le sud-est de la France, sur la côte méditerranéenne, entre Montpellier au nord et Perpignan au sud. **<s=13>** Centrée sur la basse plaine de l'Aude, elle s'étend, d'est en ouest, de la mer Méditerranée aux contreforts des Corbières maritimes, englobant le massif calcaire de la Clape, et, du nord au sud, du cours actuel de l'Aude aux abords de l'île de Sainte-Lucie, intégrant une grande partie des étangs de Bages-Sigean.

<s=14> Située dans les Corbières maritimes, Narbonne est un carrefour connu de l'homme

depuis le néolithique. **<s=15>** Aujourd'hui, la ville est un carrefour routier, ferroviaire et autoroutier.

<s=16> Elle est limitrophe avec les communes de Gruissan, Fleury d'Aude, Armissan, Vinassan, Coursan, Cuxac-d'Aude, Moussan, Marcorignan, Montredon-des-Corbières, Bages, Peyriac-de-Mer et Bizanet. **<s=17>** Ces communes constituent, avec Névian, Ouveillan, Raissac-d'Aude, Salles-d'Aude et Villedaigne, la CAN, communauté d'agglomération de la Narbonnaise.

<s=18>1.1.2. Climat

<s=19> Le climat de Narbonne est de type méditerranéen. **<s=20>** Il se caractérise par des hivers doux, des étés secs, une luminosité importante et des vents assez violents. **<s=21>** Narbonne compte 300 jours d'ensoleillement par an.

<s=22>1.1.3. Voies de communication et transport

<s=23> Narbonne est située sur un n$\frac{1}{2}$ud de communication le long du littoral méditerranéen. **<s=24>** L'autoroute des Deux Mers, et plus précisément le tronçon est de l'A61, se termine au sud de la ville en rejoignant l'A9. **<s=25>** L'A61 permet d'accéder à l'ouest à Toulouse puis Bordeaux, via Carcassonne, tandis que l'A9 rejoint au nord Montpellier puis Nîmes et au sud Perpignan puis Barcelone en Espagne. **<s=26>** Deux sorties (37 Narbonne-Est et 38 Narbonne-Sud)de l'A9 desservent la commune. **<s=27>** Sur le réseau secondaire, la route nationale 113 se termine aussi à Narbonne en rejoignant la route nationale. **<s=28>** La RN113 permet de rejoindre Toulouse à l'ouest, tandis que la RN 9 rejoint Montpellier au nord et Perpignan au sud. **<s=29>** Le réseau principal d'autoroute est parallèle au réseau secondaire de nationale sur toute la commune et la région. **<s=30>** Enfin, la route départementale 613 permet de rejoindre au sud-ouest Couiza et la haute vallée de l'Aude en traversant la région des Corbières et la départementale 607 rejoint au nord-ouest Lacaune dans la montagne Noire.

<s=31> Au niveau fluvial, le canal de la Robine traverse la ville en suivant l'ancien lit de l'Aude. **<s=32>** Il permet de rejoindre le fleuve Aude puis le canal du Midi, qui passe plus au nord de la ville, via le canal de Jonction. **<s=33>** Au sud, le canal rejoint la mer Méditerranée. **<s=34>** Narbonne possède aussi son port nautique sur les rives de l'étang de Sigean.

<s=35> La commune est aussi accessible par le train grâce à la liaison Bordeaux-Sète et Marseille via Toulouse et la liaison Tarascon-Port-Bou. **<s=36>** La gare de Narbonne est située à l'ouest de la ville sur le boulevard Frédéric-Mistral, et c'est la première gare ferroviaire du département de l'Aude (11), avant celle de Carcassonne.

<s=37> Plusieurs TGV quotidiens, effectuant les liaisons Perpignan-Paris, Perpignan-Lille,

Toulouse-Lille via Lyon, les Corail Téoz effectuant les liaisons Cerbère-Paris Austerlitz, Bordeaux-Toulouse-Marseille ainsi que beaucoup de TER à destination de Perpignan, Narbonne, etc, desservent cette gare.

<s=38> Narbonne ne possède pas d'aéroport, mais les plus proches sont ceux de Béziers, de Carcassonne situés à 30 min, puis ceux de Perpignan et de Montpellier.

<s=39> La ville possède son réseau de bus, géré par les TAN (Transports de l'Agglomération Narbonnaise)sous la houlette de la communauté d'agglomération de la Narbonnaise et mis en place en juillet. **<s=40>** Trois lignes de bus desservent le centre-ville et plusieurs services de navettes parcourent le reste de la communauté d'agglomération. **<s=41>** Enfin, en été, des navettes permettent d'accéder aux plages de Narbonne-Plage, Saint-Pierre-la-Mer et Gruissan.

<s=42>1.2. Site

<s=43> De très grande superficie, le territoire communal de 17 296 hectares fait de Narbonne la commune la plus vaste de la région Languedoc-Roussillon et du département de l'Aude. **<s=44>** Elle est la 23e commune la plus étendue de la France métropolitaine. **<s=45>** La ville est construite dans une cuvette entre le massif de la Clape à l'est et les collines des Corbières à l'ouest. **<s=46>** La commune possède au sud une partie de l'étang de Sigean. **<s=47>** Le territoire de la commune englobe aussi la station balnéaire Narbonne-Plage située à plus de dix kilomètres du centre ancien de Narbonne. **<s=48>** A peine un quart du territoire communal est urbanisé.

<s=49>1.3. Morphologie urbaine

<s=50> Narbonne est divisée en plusieurs quartiers : Anatole France, Baliste-Razimbaud, Centre Ville, Crabit-Amarats, Horte Neuve, Gazagnepas-Route de Gruissan, Hauts de Narbonne-Montplaisir Roche Grise-La Nautique, Joffre-Voltaire, Mayolle, Narbonne-Plage, Pastouret, Révolution, Saint Jean Saint Pierre-La Campane-La Source et Saint Jean Saint Pierre-Saint Salvayre.

<s=51>1.4. Urbanisme

<s=52>1.5. Structures

<s=53> - Espace de Liberté : Piscine olympique, patinoire, bowling, cafétéria.

<s=54> - Théâtre Scène Nationale

<s=55> - Médiathèque - 9 salles de cinéma

<s=56> - Palais des Sports, des Arts et du Travail

<s=57> - Parc des Sports et de l'Amitié (Egassiairal)

<s=58> - Parc des expositions

<s=59> - Golf

<s=60> - Port de plaisance

<s=61> - Station nautique

<s=62> - Musées - Autoroutes A9 et A61

<s=63> - Pole universitaire de l'Université de Perpignan Via Domitia, comprenant de nombreuses formations juridiques : licence en Droit, IUT carrieres juridiques, IUP et Master Urbanisme-Immobilier, licence vini-viticole, capacité en Droit...

<s=64>1.5.1. Logement

<s=65> Narbonne comptait 20 903 logements en 1999. **<s=66>** Les constructions neuves sont peu présentes, puisqu'en 1999 seulement 10,5 % des résidences principales dataient de 1990 ou après. **<s=67>** A contrario, les constructions antérieures à 1949 représentaient 27,7 % du parc.

<s=68> 76,2 % des logements sont des résidences principales, réparties à 58,7 % en maisons individuelles et à 41,3 % en appartements (respectivement 59,1 % et 40,9 % dans la région). **<s=69>** 45,1 % des habitants sont propriétaires de leur logement, contre 50,4 % qui ne sont que locataires (respectivement 56,8 % et 37,6 % dans la région).

<s=70> Avec 4 359 logements HLM, soit 20,9 % du parc en 1999 (9,6 % dans la région), la ville respecte les dispositions de l'article 55 de la loi solidarité et renouvellement urbain (SRU)de décembre 2000 fixant à 20 % le taux minimum de logements sociaux pour les communes les plus importantes. **<s=71>** On peut constater en outre que le nombre de logements vacants était assez faible en 1999, avec seulement 6,8 % du parc contre 7,7 % dans la région.

<s=72> La plupart des habitations possèdent 4 pièces (55 %), ou 3 pièces (25,9 %), puis 2 pièces (14,2 %). **<s=73>** Les petits logements restent peu nombreux (studios : 4,9 %). **<s=74>** La ville possède par conséquent des logements de taille importante du fait de l'espace immobilier non restreint, permettant de grandes constructions, et du fait de la faible demande en petits logements. **<s=75>** Enfin il faut préciser que ces logements sont bien dotés puisque 83,7 % ont le chauffage central et 53,2 % possèdent un garage, box ou parking (respectivement 76,5 % et 61,7 % pour la région).

<s=76>1.5.2. Projets

<s=77> Narbonne tente de restructurer l'habitat assez ancien du centre-ville. **<s=78>** Pour cela, la communauté d'agglomération de la Narbonnaise propose des aides supplémentaires pour améliorer le confort des logements, les façades, la création de garages et l'accès aux

étages. **<s=79>** De plus, une Opération Programmée d'Amélioration de l'Habitat a été menée en 1979 et en 2006.

<s=80>2. Histoire

<s=81> Les Romains fondèrent en 118 av. J.-C. une colonie romaine du nom de Colonia Narbo Martius d'après le nom du consul romain Quintus Marcius Rex. **<s=82>** Elle était située sur la via Domitia, la première route romaine en Gaule, qui permettait de relier l'Italie et l'Espagne. **<s=83>** Avant cette période, Narbonne était un oppidum préromain (site de Montlaurès, à quatre kilomètres au nord de la ville actuelle), la capitale des Elisyques, un des peuples de la Celtique méditerranéenne.

<s=84> Narbonne était la capitale de la Province de la Gaule narbonnaise, province romaine créée en 27 avant J.-C. **<s=85>** Elle fut jusqu'à la fin de l'Antiquité romaine l'une des villes les plus importantes de la Gaule avec une superficie de plus de 2 kilomètres carrés.

<s=86> En 462, Narbonne fut intégrée au royaume wisigoth de Toulouse.

<s=87> En 719 la ville est conquise par les troupes berbéro-musulmanes venus d'Espagne et devient Arbûna, le siège d'un wâli pendant quarante ans, la capitale d'une des cinq provinces d'al-Andalus, aux cotés de Cordoue, Tolède, Mérida et Saragosse. **<s=88>** Les musulmans laissèrent aux anciens habitants, chrétiens et juifs, la liberté de professer leur religion moyennant tribut.

<s=89> En 759 la ville est reprise par Pépin le Bref et fait définitivement partie du royaume franc.

<s=90> En 859, Narbonne est pillée par les Vikings du chef Hasting, qui viennent de Nantes et avaient hiverné en Camargue.

<s=91> Jusqu'à la fin du Moyen Âge, Narbonne est gouvernée par deux seigneurs : l'archevêque et le vicomte.

<s=92> À la Renaissance, les protestants sont chassés de la ville en 1562. **<s=93>** Charles IX est reçu en grande pompe dans la ville lors de son tour de France royal (1564-1566), accompagné de la Cour et des Grands du royaume : son frère le duc d'Anjou, Henri de Navarre, les cardinaux de Bourbon et de Lorraine.

<s=94> L'arrivée du canal du Midi et la présence de l'archevêché marquent la période prérévolutionnaire. **<s=95>** Après la Révolution, privée du siège épiscopal, la commune n'est plus qu'une sous-préfecture rurale.

<s=96> Devenue capitale d'un espace viticole à partir du développement de la vigne vers 1850/1870, et profitant de sa situation de nœud ferroviaire, Narbonne se démarque politiquement : les vignerons et commerçants sont Républicains. **<s=97>** La municipalité s'oppose à

Napoléon III avant la chute de l'Empire en 1870 (le maire est Marcelin Courral), se soulève contre les Versaillais de Thiers en une commune en mars 1871, puis élit à la fin du XIXe siècle un maire félibrige et socialiste, Ernest Ferroul, dit le docteur " des pauvres ", qui soutient la grande Révolte des vignerons du Languedoc en 1907. **<s=98>** Léon Blum en devient député en 1929. **<s=99>** Le maire socialiste, révoqué par Vichy, Achille Lacroix, meurt en déportation. **<s=100>** Jusqu'à l'arrivée du tourisme dans les années 1960, la commune reste très liée aux crises de la viticulture.

<s=101>3. Héraldique

<s=102> Narbonne - blason actuel « Parti, au premier de gueules, à une clef d'or forée en pal ; au deuxième aussi de gueules, à une double croix d'or posée d'argent, au chef d'azur chargé de trois fleurs de lis d'or ».

<s=103> La clef symbolise les portes de la Cité représentée par les Consuls, la croix archiépiscopale représente le siège des archevêques de la province et les trois fleurs de lys représentent son attachement au royaume de France. **<s=104>** La couleur bleue du blason est celle des vicomtes de Narbonne.

<s=105>4. Administration

<s=106> Narbonne est chef-lieu de trois cantons :

<s=107> - Le canton de Narbonne-Est est formé d'une partie de Narbonne (16 851 habitants) ;

<s=108> - Le canton de Narbonne-Ouest est formé d'une partie de Narbonne et des communes de Bizanet, Canet, Marcorignan, Montredon-des-Corbières, Moussan, Névian, Raissac-d'Aude et Villedaigne (21 459 habitants) ;

<s=109> - Le canton de Narbonne-Sud est formé d'une partie de Narbonne et de la commune de Bages (16 043 habitants).

<s=110> Narbonne fait partie de la juridiction d'instance, de grande instance et de commerce de Narbonne, ainsi que de la cour d'appel de Montpellier.

<s=111>4.1. Tendances politiques

<s=112> Politiquement, après une longue période de socialisme municipal, Narbonne à été pendant 35 ans une ville de droite, les électeurs ayant voté majoritairement à droite pour les élections municipales depuis 1971. **<s=113>** Le maire de la commune élu de 1999 à 2008, Michel Moynier, est classé Divers Droite. **<s=114>** Il a succédé à Hubert Mouly, lui-même Divers Droite. **<s=115>** Enfin, en 2008, c'est le député socialiste Jacques Bascou qui a été élu maire de Narbonne.

<s=116> À l'élection présidentielle de 2002, le premier tour a vu arriver en tête Jean-Marie Le Pen avec 20,87 %, suivi de Lionel Jospin avec 18,30 %, puis de Jacques Chirac avec 16,85 %, puis Arlette Laguillier avec 5,80 %, et enfin Jean-Pierre Chevènement avec 5,77 %, aucun autre candidat ne dépassant le seuil des 5 %. **<s=117>** Au second tour, les électeurs ont voté à 76,54 % pour Jacques Chirac contre 23,46 % pour Jean-Marie Le Pen, avec un taux d'abstention de 22,24 %, résultat différant des tendances nationales (respectivement 82,21 % et 17,79 % ; abstention 20,29 %)avec cinq points supplémentaires pour Jean-Marie Le Pen.

<s=118> Au référendum sur le traité constitutionnel pour l'Europe du 29 mai 2005, les Narbonnais ont largement voté contre la Constitution Européenne, avec 62,47 % de Non contre 37,53 % de Oui, avec un taux d'abstention de 32,61 % (France entière : Non à 54,67 % ; Oui à 45,33 %). **<s=119>** Ces chiffres sont assez conformes à la tendance départementale de l'Aude (Non à 64,62 % ; Oui à 35,38 %), l'électorat ayant choisi le vote positif étant, selon les analystes politiques, le fait d'une population plus privilégiée économiquement et d'un plus haut niveau d'éducation que la moyenne des Français.

<s=120> À l'élection présidentielle de 2007, le premier tour a vu se démarquer en tête Nicolas Sarkozy avec 29,26 %, suivi par Ségolène Royal avec 28,07 %, François Bayrou avec 15,30 %, Jean-Marie Le Pen avec 12,97 %, puis Olivier Besancenot avec 4,18 %, et enfin Marie-George Buffet avec 2,68 %, aucun autre candidat ne dépassant le seuil des 2 %. **<s=121>** Le second tour a vu arriver en tête Nicolas Sarkozy, avec 50,93 % (national : 53,06 %)contre 49,07 % pour Ségolène Royal (résultat national : 46,94 %).

<s=122>4.2. Liste des maires

<s=123>4.3. Sécurité

<s=124> Le taux de criminalité de la circonscription de police de Narbonne est de 85,86 actes pour 1000 habitants (crimes et délits, chiffres 2005), ce qui en fait un taux assez élevé dans l'Aude, mais presque équivalent à la moyenne nationale (83/1000). **<s=125>** Ce taux est néanmoins inférieur au taux de criminalité de la région Languedoc-Roussillon (109,31). **<s=126>** Le taux de résolution des affaires par les services de police est de 33,45 %, un taux moyen pour le département, mais plus proche des moyennes régionale (26,79 %)et nationale (28,76 %).

<s=127> La mairie de Narbonne gère environ une vingtaine d'agents de la police municipale.

<s=128>4.4. Jumelages

<s=129> Quatre jumelages sont actifs :

<s=130> - Weilheim (de)(Allemagne)- Grosseto (Italie)

<s=131> - Salford (Royaume-Uni)

<s=132> - Aoste (Italie)

<s=133> - Niederalteich (de)(Allemagne), voir St. Gotthard-Gymnasium, Niederalteich/Bavière, jumelage avec l'Institution Sévigné et avec le lycée Beauséjour

<s=134>5. Démographie

<s=135> La ville de Narbonne compte 50 776 habitants en 2006. **<s=136>** Elle est la ville la plus peuplée de l'Aude. **<s=137>** La densité, de 269 hab/km^{e}, est largement plus forte que celle de la moyenne nationale, de 111 hab/km^{e}, mais est assez faible par rapport à d'autres villes importantes de la région, notamment en raison de l'étendue de la commune. **<s=138>** L'aire urbaine de Narbonne est aussi la deuxième la plus peuplée du département avec 70 750 habitants et couvre 14 communes, venant après celle de Carcassonne (82 577 hab)et devant celles de Castelnaudary (19 079 hab)et de Limoux (15 160 hab).

<s=139> L'évolution démographique de la ville est en régulière augmentation depuis le XIXe siècle, étant passée de 9 086 habitants en 1800 à 50 776 habitants en 2006.

<s=140>5.1. Tableau démographique du XXe siècle

<s=141>6. Économie

<s=142> Narbonne est le siège de la Chambre de commerce et d'industrie de Narbonne - Lézignan-Corbières et Port-la-Nouvelle. **<s=143>** Elle gère Port-la-Nouvelle et le Port-la-Robine.

<s=144> Au nord-ouest de la commune, à Malvesi, se trouve une usine de raffinage et de conversion d'uranium (Comurhex, faisant partie du groupe Areva)ainsi qu'une entreprise de fabrication de produits chimiques.

<s=145>7. Personnages célèbres

<s=146> - Félix Barthe (1795-1863), ministre de l'Instruction publique et des Cultes, puis ministre de la Justice et Premier président de la Cour des comptes.

<s=147> - Léon Blum, homme politique élu député puis président du Conseil en 1936.

<s=148> - Joë Bousquet, écrivain surréaliste

<s=149> - Ermengarde de Narbonne

<s=150> - Ernest Ferroul - René Iché, sculpteur moderne

<s=151> - Jean-Joseph Cassanéa de Mondonville, compositeur et violoniste

<s=152> - Meshoullam ben Nathan (XIIe siècle)

<s=153> - Moshe haDarshan (XIe siècle)

<s=154> - Saint Sébastien

<s=155> - Saint Prudent

<s=156> - Saint Théodard

<s=157> - Pierre Reverdy, écrivain précurseur du surréalisme

<s=158> - Charles Trenet, (18 mai 1913, Narbonne - 19 février 2001, Créteil)chanteur et poète

<s=159> - Benjamin Crémieux, écrivain, universitaire, chargé de missions diplomatiques...

<s=160> - Varron (Publius Terentius Varro Atacinus), poète épique romain (82-37 av. J.-C.

<s=161>)- Amédée Domenech, (1933 - 2003), joueur de rugby à XV

<s=162> - Dimitri Szarzewski, joueur de rugby à XV en Equipe de France de rugby à XV

<s=163> - René Coll.

<s=164>8. Patrimoine

<s=165>8.1. Patrimoine culturel

<s=166> Narbonne est classée ville d'art et d'histoire.

<s=167>8.1.1. Patrimoine antique

<s=168> Narbonne a perdu la plupart des monuments qui l'ornèrent durant l'époque romaine. **<s=169>** Les horrea sont composés de galeries souterraines remontant au Ier siècle avant notre ère. **<s=170>** Unique en Europe, il est situé sous un monument disparu qui aurait pu servir d'entrepôt public (horreum).

<s=171> Au centre de la place de l'Hôtel-de-Ville, l'antique voie Domitienne (Via Domitia)est visible dans son état de la fin du IVe siècle. **<s=172>** C'est un vestige de la première grande route romaine, tracée en Gaule à partir de 120 av. J.-C., par le proconsul Cneus Domitius Ahenobarbus, deux ans avant la fondation de la Colonia Narbo Martius, première colonie romaine en Gaule. **<s=173>** La voie Domitienne reliait l'Italie à l'Espagne romanisée. **<s=174>** À Narbonne, elle rencontrait la voie Aquitaine (via Aquitania), ouverte en direction de l'Atlantique par Toulouse et Bordeaux, attestant dès cette époque du rôle de carrefour tenu par la ville. **<s=175>** Le vestige découvert le 7 février 1997 présente une portion de voie dallée de calcaire dur, marquée par de profondes ornières. **<s=176>** Elle est bordée de trottoirs et de la base d'une fontaine.

<s=177>8.1.2. Patrimoine médiéval

<s=178> La cathédrale Saint-Just-et-Saint-Pasteur est une cathédrale dont la première pierre fut posée en 1272. **<s=179>** Sa construction fut arrêtée en 1355, en raison de l'invasion de la ville par le Prince Noir. **<s=180>** Elle ne sera jamais terminée. **<s=181>** Cette cathédrale est

la quatrième plus haute de France, après celles de Beauvais, d'Amiens et de Metz. **<s=182>** Sa hauteur est de 41 mètres.

<s=183> Le palais des archevêques est composé du palais Vieux d'origine romane et du palais Neuf de style gothique. **<s=184>** Il accueille depuis le XIXe siècle la mairie, le musée d'Art et le musée archéologique.

<s=185> Le pont des Marchands, reliant le Bourg à la Cité, ce pont permettait à l'origine le franchissement de l'Aude par la voie Domitienne. **<s=186>** Ce pont bâti, rare en Europe, était constitué de 7 arches. **<s=187>** Depuis que l'Aude a quitté son ancien cours, et que son lit accueille le canal de la Robine, classé par l'UNESCO au Patrimoine mondial de l'humanité, une seule arche suffit au passage de l'eau, les autres servant de caves aux maisons bâties des deux côtés du pont.

<s=188> La basilique Saint-Paul, l'une des plus anciennes églises gothiques du Midi de la France, est construite sur les vestiges d'un ancien cimetière paléochrétien (IIIe-IVe siècles). **<s=189>** Cet édifice a la particularité de mêler art roman et art gothique. **<s=190>** Son bénitier à la grenouille est célèbre.

<s=191> L'église Notre-Dame de Lamourguier, édifice typique du gothique languedocien et catalan, est l'unique reste d'un prieuré bénédictin. **<s=192>** Elle sert désormais de musée lapidaire où ont été déposés les blocs sculptés antiques retirés des remparts de Narbonne où ils avaient été réutilisés.

<s=193> Non loin de Narbonne, l'abbaye Sainte-Marie de Fontfroide est un ancien monastère cistercien.

<s=194>8.1.3. Patrimoine de la Renaissance

<s=195> Il est principalement représenté par la maison des Trois-Nourrices.

<s=196>8.1.4. Patrimoine des XXe et XXIe siècles

<s=197> Les halles de style Baltard ont été inaugurées le 1er janvier 1901.

<s=198> À proximité de la statue d'Ernest Ferroul, érigée en 1933 par la Confédération générale des vignerons par souscription nationale, est édifié, à partir de 1938, sous la direction de Joachim Genard, le palais des Arts, des Sports et du Travail, ensemble architectural d'influence néo-classique abritant une piscine, une salle de sport, une salle des fêtes, achevée en 1967, des salles d'assemblées et des sièges syndicaux. **<s=199>** Le projet d'y loger un théâtre n'aboutira pas. **<s=200>** Les locaux de la proprement dite Bourse du Travail, partie prenante de l'ensemble, sont inaugurés en 1952.

<s=201> Le Théâtre Scène nationale, la médiathèque et le palais de justice sont des ouvrages

à l'architecture contemporaine.

<s=202> René Iché, sculpteur français du XXe siècle a laissé quelques-unes de ses $\frac{1}{2}$uvres dans Narbonne :

<s=203> - Le buste d'Albarel, dans les jardins de la médiathèque ;

<s=204> - L'Art, le travail et le sport, hauts-reliefs au dessus des portes du palais de Travail ;

<s=205> - L'Athlète, au parc des Sports et de l'Amitié ;

<s=206> - Étude de lutteurs et La Petite Danseuse, au musée des Beaux-Arts.

<s=207>8.2. Patrimoine environnemental

<s=208> La ville de Narbonne possède 150 ha d'espaces verts, soit plus de 20 000 arbres et plus de 60 aires de jeux à entretenir. **<s=209>** Elle figure au palmarès des villes et villages fleuris avec 2 fleurs attribuées par le Conseil national des villes et villages fleuris de France.

<s=210> Plusieurs espaces verts sont à la disposition des Narbonnais, comme le jardin de la Révolution, le jardin du Plan Saint Paul, le jardin des Archevêques, récemment rénové, le parc de la Campane, les berges du canal de la Robine et le massif de l'abbaye de Fontfroide.

<s=211>9. Éducation

<s=212> Narbonne accueille plus de 4000 écoliers dans 31 écoles. **<s=213>** Elle possède également plusieurs collèges (Victor Hugo, Georges Brassens, Jules Ferry, Charles-Louis Montesquieu, Cité, Institution Sévigné)et lycées (Docteur Lacroix, Denis Diderot, Gustave Eiffel, Beauséjour).

<s=214> Elle est également dotée d'une antenne Université de Perpignan (faculté de Droit et de Sciences économiques)et d'un IUT carrières juridiques. **<s=215>** Plus de 1000 étudiants fréquentent le campus. **<s=216>** Une résidence étudiante est disponible depuis septembre 2007, ainsi qu'un CROUS et une cafétéria universitaire. **<s=217>** Un projet de développement des installations universitaires est en cours.

<s=218>10. Culture

<s=219>10.1. Narbonne dans la littérature classique

<s=220> Victor Hugo, dans un poème de la La Légende des Siècles intitulé Aymerillot, consacre près de 300 vers dans lesquels il relate la difficulté que rencontre Charlemagne pour trouver parmi ses hommes celui qui osera entreprendre le siège de Narbonne. **<s=221>** L'empereur, en découvrant cette « ville unique sous les cieux », assure : « J'aurai cette ville avant d'aller plus loin. » **<s=222>** Il demande tour à tour à ses plus vaillants chevaliers, comtes, ducs et seigneurs de prendre Narbonne, mais la ville est si bien défendue, et ses hommes si usés par

leurs précédents combats, que tous refusent. **<s=223>** C'est alors qu'un jeune homme inconnu de vingt ans, qui est pris « pour une fille habillée en garçon », et nommé Aymerillot, s'avance vers l'empereur et affirme : « J'entrerai dans Narbonne et je serai vainqueur. »

<s=224>10.2. Narbonne dans la littérature contemporaine

<s=225> Michel Sidobre, dans Et pourquoi pas Narbonne, décrit l'ombre des platanes et la douceur de vivre. **<s=226>** Il croque les gens simples affairés dans le Cimetière de Cité, restituant une ambiance méditerranéenne bon enfant<bon enfant_A|0>.

<s=227>10.3. Narbonne et le cinéma

<s=228> C'est à Narbonne que fut tourné Le Père Noël a les yeux bleus de Jean Eustache. **<s=229>** Le film est dédié à Charles Trenet, natif de Narbonne.

<s=230>10.4. Médias

<s=231> Les studios de la première radio occitane du Languedoc-Roussillon, Ràdio Lenga d'òc, sont situés à Narbonne.

<s=232> La ville compte également une radio locale, Radio Narbonne Méditerranée, une antenne de Virgin Radio ainsi qu'une WebTV tv-narbonne.com

<s=233>10.5. Festival Charles Trenet

<s=234> Le festival a été imaginé depuis de nombreuses années dans la ville natale du poète et chanteur, ce n'est qu'en 2008 qu'il a pu se réaliser et voir le jour grâce à l'insistance de René Coll.

<s=235> Le prix Sacem/ Trenet soutenu par la Sacem et le Centre Régional de la Chanson, récompense les nouveaux talents de la scène Française après une longue sélection. **<s=236>** Un jury de personnalités et de professionnels juge alors la prestation des jeunes artistes.

<s=237> Auteurs, compositeurs, interprètes sont les bienvenus pour participer à ce concours national. **<s=238>** Les différents concerts gratuits étaient présentés par Laurent Boyer avec la participation notamment du Grand Orchestre de René Coll, de Chico et les Gypsies, de Sheryfa Luna, de Sofia Mestari ou encore de Frédéric Lerner.

<s=239>11. Sports

<s=240> Narbonne est également connue pour son club de rugby le RC Narbonne qui évolue cette saison en PRO D2 après être resté pendant 100 ans en Top 14. **<s=241>** Grand centre de formation du rugby français, le RCNM a remporté 2 fois le titre de champion de France

et possède le record de victoires en Challenge Yves du Manoir (coupe de France de rugby). **<s=242>** Le club est arrivé en finale du Bouclier européen en 2001. **<s=243>** Ce club mythique du rugby francais a fourni un nombre considérable de joueurs à l'équipe de France, ainsi le grand Walter Spanghero, son frère Claude ou encore François Sangalli et Franck Tournaire et le petit prince Didier Codorniou. **<s=244>** Le RCNM joue au stade de l'Egassiairal (12000 places). **<s=245>** Le club de rugby fait partie intégrante de la culture et de la vie narbonnaises.

<s=246> Le volley est également très présent à Narbonne, l'UVBN joue en ProA au Palais des Sports.

<s=247> Le Football Union Narbonne, plus connu sous le nom de « FUN », est le club de football de la Ville de Narbonne. **<s=248>** Toutes ses équipes évoluent au plus haut niveau régional.

<s=249>12. Garnison

<s=250> Le 100e régiment d'infanterie était en garnison à Narbonne durant la révolte des vignerons du Languedoc en 1907, et fut consigné cinq dimanches de suite. **<s=251>** Cependant, des groupes d'appelés acclament les manifestants et entonnent l'Internationale. **<s=252>** Le régiment est envoyé en man$\frac{1}{2}$uvres dans le Larzac, puis en garnison à Tulle.

<s=253>13. Distinctions

<s=254> - Le label « Ville d'Art et d'Histoire »

<s=255> - Ville fleurie : 2 fleurs depuis 2002

<s=256> - Ville sportive

<s=257> - Ville touristique : 2 étoiles au guide Michelin

<s=258> - Label France station nautique

<s=259> - Station touristique et balnéaire (dossier en cours de classement

<s=260> - Classée à deux reprises Première ville de sa catégorie pour l'accueil des entreprises par le magazine l'Entreprise

<s=261> - Pavillon Bleu attribué depuis 1988

<s=262> - Classée " Ports propres " (pour la Nautique et Narbonne-Plage)

<s=263> - Prix Éco-Maire en 2005

<s=264> - Prix des " Rubans du développement durable " en 2006

<s=265> - 4@ aux Labels Villes Internet en 2006.

A.2 Résumé introductif de l'article *Narbonne* étiqueté

<s=2> Narbonne**<narbonne_N>** (en occitan**<occitan_N>** Narbona**<narbona_N>**) est une commune**<commune_N>** française**<français_A>**, située**<situer_V>** dans le département-**<département_N>** de l'Aude**<aude_N>** et la région**<région_N>** Languedoc-Roussillon**<languedoc-roussillon_N>**.

<s=3> La commune**<commune_N>** est traversée**<traverser_V>** par le canal**<canal_N>** de la Robine**<robine_N>**, classé**<classer_V>** au Patrimoine**<patrimoine_N>** mondial**<mondial_A>** de l'humanité**<humanité_N>** par l'UNESCO**<unesco_N>** depuis 1996**<1996_N>**.

<s=4> Elle est classée**<classer_V>** 120e ville**<ville_N>** de France**<france_N>** en termes de population**<population_N>**.

<s=5> Située**<situer_V>** au cœur**<cœur_N>** du 51e parc**<parc_N>** naturel**<naturel_A>** de France**<france_N>** « parc**<parc_N>** naturel**<naturel_A>** régional**<régional_A>** de la Narbonnaise**<narbonnaise_N>** en Méditerranée**<méditerranée_N>** », Narbonne**<narbonne_N>** possède**<posséder_V>** également d'autres sites**<site_N>** naturels**<naturel_A>** classés**<classer_V>**, comme le massif**<massif_N>** de la Clape**<clape_N>** et celui de l'abbaye**<abbaye_N>** Sainte-Marie**<sainte-marie_N>** de Fontfroide**<fontfroide_N>**, ainsi que les étangs**<étang_N>** de Bages-Sigean**<bages-sigean_N>**.

<s=6> Fondée**<fondé_A>** par les Romains**<romain_N>** en 118 av. J. C., elle était leur plus ancienne**<ancien_A>** colonie**<colonie_N>** en Gaule**<gaule_N>** et son centre**<centre_N>** urbain**<urbain_A>** garde**<garder_V>** trace**<trace_N>** de nombreux**<nombreux_A>** siècles-

<siècle_N> d'histoire**<histoire_N>** (cathédrale**<cathédrale_N>** Saint-Just-et-Saint-Pasteur**<saint-just-et-saint-pasteur_N>**, palais**<palais_N>** des Archevêques**<archevêque_N>**, restes**<reste_N>** de la voie**<voie_N>** Domitienne**<domitien_A>**).

<s=7> La ville**<ville_N>** est entourée**<entourer_V>** d'un environnement**<environnement_N>** exceptionnel**<exceptionnel_A>** fait de garrigues**<garrigue_N>** et de vignes**<vigne_N>** ; proche**<proche_A>** du littoral**<littoral_N>** d'une région**<région_N>** très touristique**<touristique_A>**, elle possède**<posséder_V>** une plage**<plage_N>** de cinq kilomètres**<kilomètre_N>** de sable**<sable_N>** fin**<fin_A>** à Narbonne-Plage**<narbonne-plage_N>**.

<s=8> Les habitants**<habitant_N>** de Narbonne**<narbonne_N>** sont appelés**<appeler_V>** les Narbonnais**<narbonnais_N>**.

Bibliographie

[Adam, 1999] Adam, J.-M. (1999). *Linguistique textuelle : des genres de discours aux textes*. Nathan (Paris).

[Apothéloz, 1985] Apothéloz, D. (1985). Un point de vue cognitif sur l'activité de discours. *Bulletin de psychologie*, 371.

[Barzilay et Elhadad, 1997] Barzilay, R. et Elhadad, M. (1997). Using lexical chains for text summarization. *In Proceedings of the ACL Workshop on Intelligent Scalable Text Summarization*.

[Benbrahim, 1996] Benbrahim, M. (1996). *Automatic text summarisation through lexical cohesion analysis*. Thèse de doctorat, University of Surrey.

[Benbrahim et Ahmad, 1994] Benbrahim, M. et Ahmad, K. (1994). *Computer-aided lexical cohesion analysis and text abridgement*. Rapport technique, University of Surrey.

[Berber Sardinha, 2001] Berber Sardinha, T. (2001). Lexical segments in text. *In* Scott, M. et Thompson, G., éditeurs : *Patterns of text : in honour of Michael Hoey*, pages 213–237. John Benjamins.

[Biber, 1988] Biber, D. (1988). *Variation across Speech and Writing*. Cambridge University Press (Cambridge).

[Bourigault, 2002] Bourigault, D. (2002). Upery : un outil d'analyse distributionnelle étendue pour la construction d'ontologies à partir de corpus. *In Actes de la 9^{e} conférence sur le Traitement Automatique de la Langue Naturelle* (24-27 juin 2002), Nancy.

[Bourigault et Galy, 2005] Bourigault, D. et Galy, E. (2005). Analyse distributionnelle de corpus de langue générale et synonymie. *In 4es Journées de la linguistique de corpus* (15–17 septembre 2005), Lorient.

[Brown et Yule, 1983] Brown, G. et Yule, G. (1983). *Discourse analysis*. Cambridge University Press (Cambridge).

[Brunn *et al.*, 2001] Brunn, M., Chali, Y. et Pinchak, C. J. (2001). Text summarization using lexical chains. *In Proceedings of the Document Understanding Conference (DUC 2001)*.

[Chaffin et Herrmann, 1984] Chaffin, R. et Herrmann, D. J. (1984). The similarity and diversity of semantic relations. *Memory and Cognition*, 12(2):134–141.

[Charaudeau et Maingueneau, 2002] Charaudeau, P. et Maingueneau, D., éditeurs (2002). *Dictionnaire d'analyse du discours*. Éditions du Seuil (Paris).

[Charolles, 1978] Charolles, M. (1978). Introduction aux problèmes de la cohérence des textes. *Langue française*, 31(1):7–41.

[Choi, 2000] Choi, F. Y. Y. (2000). Advances in domain independent linear text segmentation. *In Proceedings of the first conference on North American chapter of the Association for Computational Linguistics*, pages 26–33, San Francisco.

[Coseriu, 1970] Coseriu, E. (1970). Structure lexicale et enseignement du vocabulaire. *In* Rey, A., éditeur : *La lexicologie*. Klincksieck (Paris).

[Dubreil, 2008] Dubreil, E. (2008). Collocations : définitions et problématique. *Texto !*, 13(1).

[Détrie *et al.*, 2001] Détrie, C., Siblot, P. et Vérine, B. (2001). *Termes et concepts pour l'analyse du discours : une approche praxématique*. Honoré Champion (Paris).

[Ellman, 2000] Ellman, J. (2000). *Using Roget's Thesaurus to determine similarity of texts*. Thèse de doctorat, University of Sunderland.

[Fabre et Bourigault, 2006] Fabre, C. et Bourigault, D. (2006). Extraction de relations sémantiques entre noms et verbes au-delà des liens morphologiques. *In Actes de la 13^e conférence sur le Traitement Automatique de la Langue Naturelle* (10–03 avril 2006), Leuven.

[Fayol, 1986] Fayol, M. (1986). Cohérence et cohésion : une revue des travaux français de psychologie expérimentale. *In* Charolles, M., Petöfi, J.-M. et Sözer, E., éditeurs : *Research in Text Connexity and Text Coherence : A survey.*, page 128. Helmut Buske Verlag (Hambourg).

[Haberlandt, 1982] Haberlandt, K. (1982). Les expectations du lecteur dans la compréhension de texte. *Bulletin de psychologie*, 256(35):733–740.

[Halliday et Hasan, 1976] Halliday, M. A. K. et Hasan, R. (1976). *Cohesion in English*. Longman (Londres).

[Harris, 1991] Harris, Z. S. (1991). *A theory of language and information : a mathematical approach*. Clarendon (Oxford).

[Hirst et Budanitsky, 2001] Hirst, G. et Budanitsky, A. (2001). Correcting real-word spelling errors by restoring lexical cohesion. *Natural Language Engineering*, 11(1):87–111.

[Hirst et St. Onge, 1997] Hirst, G. et St. Onge, D. (1997). Lexical chains as representations of context for the detection of malaproprisms. *In* Fellbaum, C., éditeur : *WordNet : an electronic lexical database*. The MIT Press (Cambridge).

[Ho-Dac *et al.*, 2004] Ho-Dac, M., Jacques, M.-P. et Rebeyrolle, J. (2004). Sur la fonction discursive des titres. *In* Porhiel, S. et Klinger, D., éditeurs : *L'unité texte*, pages 125–152.

[Hoey, 1983] Hoey, M. (1983). *On the surface of discourse*. Allen and Unwin (Londres).

[Hoey, 1991] Hoey, M. (1991). *Patterns of lexis in text*. Oxford University Press (Oxford).

[Jarmasz et Szpakowicz, 2003] Jarmasz, M. et Szpakowicz, S. (2003). Not as easy as it seems : automating the construction of lexical chains using Roget's Thesaurus. *In Canadian Conference on AI*, pages 544–549.

[Lafourcade, 2007] Lafourcade, M. (2007). Making people play for lexical acquisition. *In 7th Symposium on Natural Language Processing* (13–15 décembre 2007), Pattaya.

[Legallois, 2004] Legallois, D. (2004). Cohésion lexicale et réseaux phrastiques dans la constitution du texte expositif. *In* Porhiel, S. et Klinger, D., éditeurs : *L'unité texte*. Association perspectives.

[Lin, 1998] Lin, D. (1998). An information-theoretic definition of similarity. *In Proceedings of the 15th International Conference on Machine Learning*, pages 296–304. Morgan Kaufmann.

[Marcu, 2000] Marcu, D. (2000). The rhetorical parsing of unrestricted texts : a surface-based approach. *Computational linguistics*, 26(3):395–448.

[Meurer, 2003] Meurer, J.-L. (2003). Relationships between cohesion and coherence in essays and narratives. *Fragmentos*, (25):147–154.

[Morley, 2006] Morley, J. (2006). Lexical cohesion and rhetorical structure. *International Journal of Corpus Linguistics*, 11.

[Morris, 2004] Morris, J. (2004). Readers' perceptions of lexical cohesion in text. *In* Julien, H. et Thompson, S., éditeurs : *Proceedings of the 32nd Annual Conference of the Canadian Association of Information Science*, University of Manitoba (Winnipeg).

[Morris et Hirst, 1991] Morris, J. et Hirst, G. (1991). Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Computational linguistics*, 17(1):21–48.

[Morris et Hirst, 2004] Morris, J. et Hirst, G. (2004). Non-classical lexical semantic relations. *In* Moldovan, D. et Girju, R., éditeurs : *Proceedings of the HLT Workshop on Computational Lexical Semantics*, pages 46–51, Boston.

[Morris et Hirst, 2005] Morris, J. et Hirst, G. (2005). The subjectivity of lexical cohesion in text. *In* Shanahan, J. G., Qu, Y. et Wiebe, J., éditeurs : *Computing Attitude and Affect in Text*, pages 41–48. Springer (New York).

[Mosenthal et Tierney, 1985] Mosenthal, J. H. et Tierney, R. J. (1985). Cohesion : problems with talking about text. *Reading Research Quarterly*, 19(2):240–244.

[Ngai et Holliman, 2002] Ngai, S. et Holliman, M. (2002). *Judging relevance through identification of lexical chains*. Rapport de projet, University of Stanford.

[Niobey *et al.*, 2007] Niobey, G., De Galiana, T., Jouannon, G. et Lagane, R. (2007). *Dictionnaire analogique*. Larousse (Paris).

[Péry-Woodley, 2000] Péry-Woodley, M.-P. (2000). *Une pragmatique à fleur de texte : approche en corpus de l'organisation textuelle*. Mémoire d'HDR, Université de Toulouse II - Le Mirail.

[Péry-Woodley, 2008] Péry-Woodley, M.-P. (2008). *Discours : modèles et traitements*. Cours dispensé à l'Université Toulouse II - Le Mirail.

[Rebeyrolle et Jacques, 2006] Rebeyrolle, J. et Jacques, M.-P. (2006). Étude en corpus de la fonction des intertitres dans la construction du discours. *In Actes de la troisième Rencontre fribourgeoise de la linguistique sur corpus appliquée aux langues romanes*, Fribourg en Brisgau.

[Rebeyrolle *et al.*, 2009] Rebeyrolle, J., Jacques, M.-P. et Péry-Woodley, M.-P. (2009). Titres et intertitres dans l'organisation du discours. *Journal of French Language Studies*, 19(2).

[Richer, 2006] Richer, J.-J. (2006). Essai de définition du blog comme genre de discours. *In Journées d'étude "Plurilinguisme et multimédia"* (16 mars 2006), Lyon.

[Sanford, 2006] Sanford, A. (2006). Coherence : Psycholinguistic approach. *In* Brown, K., éditeur : *Encyclopedia of Language and Linguistics*. Elsevier (Amsterdam), 2e édition.

[Silber et McCoy, 2002] Silber, H. G. et McCoy, K. F. (2002). Efficiently computed lexical chains as an intermediate representation for automatic text summarization. *Computational linguistics*, 28(4):487–496.

[Stokes, 2004] Stokes, N. (2004). *Applications of lexical cohesion analysis in the topic detection and tracking domain*. Thèse de doctorat, University College Dublin.

[Tanskanen, 2006] Tanskanen, S. K. (2006). *Collaborating towards coherence : lexical cohesion in english discourse*. John Benjamins (Amsterdam).

[Widdowson, 1978] Widdowson, H. G. (1978). *Teaching language as communication*. Oxford University Press (Oxford).