Université de Toulouse

THÈSE

En vue de l'obtention du

DOCTORAT DE L'UNIVERSITÉ DE TOULOUSE

Délivré par :
Université Toulouse 2 Le Mirail (UT2 Le Mirail)

Présentée et soutenue par :
Clémentine Adam
Le vendredi 28 septembre 2012

Titre :

Voisinage lexical pour l'analyse du discours

ED CLESCO : Sciences du langage

Unité de recherche :
CLLE-ERSS

Directeur(s) de Thèse :
Cécile Fabre (PR), Université de Toulouse II / CLLE-ERSS
Nicholas Asher (DR) et Philippe Muller (MC), Université de Toulouse III / IRIT

Rapporteurs :

Thierry Poibeau (DR), CNRS / LaTTiCe
Pascale Sébillot (PR), INSA de Rennes / IRISA

Autre(s) membre(s) du jury :

Olivier Ferret (CR), CEA LIST, LVIC

Voisinage lexical pour l'analyse du discours

Clémentine Adam

Remerciements

...

Table des matières

Table des figures

Liste des tableaux

Introduction

La mise au jour des structures discursives dans les textes est un objectif crucial, notamment pour le développement des techniques de traitement automatique des langues visant à faciliter l'exploration de documents volumineux et l'accès automatisé à une information ciblée (fouille de données textuelles, systèmes de questions-réponses, résumé automatique, etc.). Ces structures discursives relèvent de plans variés, depuis des constituants élémentaires reliés par des relations de discours (relations de contraste, d'élaboration, d'explication, etc.) jusqu'à des segments de niveau supérieur dotés d'une fonction rhétorique et garantissant la cohérence et la lisibilité du texte (passages à unité thématique et/ou argumentative). Leur caractérisation et leur identification automatique s'appuient sur le repérage d'indices : indices typographiques et structurels, marqueurs de discours (*cue phrases*), etc. Dans cette thèse, nous nous intéressons à un type d'indices en particulier : les indices de nature lexicale.

Les mots d'un texte présentent tout un ensemble de relations sémantiques les uns avec les autres ; les liens qui les unissent tissent des motifs qui peuvent être exploités pour l'étude de la structure discursive à différents niveaux. Au niveau de l'organisation globale des textes, la densité du tissage lexical (ou cohésion lexicale) peut permettre de repérer des zones de continuité ou de rupture ; à un niveau local, certaines relations lexicales ciblées peuvent aider l'identification d'une relation de discours. Nous illustrons ces deux perspectives à partir de deux exemples.

L'exemple (1) présente deux paragraphes contigus extraits d'un article Wikipédia consacré au tatou. La continuité thématique à l'intérieur de chacun des paragraphes est signalée par de nombreuses relations lexicales, notamment celles connectant différents mots désignant des parties du corps dans le premier paragraphe (mots en italique), et différents mots relevant du vocabulaire associé aux relations sociales dans le second paragraphe (mots soulignés). Par contre, il est plus difficile d'identifier des relations lexicales connectant les deux paragraphes, ce qui souligne leur discontinuité.

(1) **Description**

La carapace sur le *dos* des tatous est formée de plaques osseuses articulées recouvertes de corne. Elles recouvrent la totalité du *dos* de l'animal, du *front* jusqu'à la *queue*, surface externe des *membres* comprise. Selon les espèces, elles forment soit des ceintures successives séparées par des replis cutanés souples, comme chez le tatou géant ; soit deux boucliers, l'un protégeant les *épaules*, l'autre les *hanches*, et séparés par des bandes d'*écailles* en nombre variable. Certaines espèces comme le tatou à trois bandes peuvent s'enrouler en boule en cas de danger.

Comportement

En dehors des périodes de reproduction, le tatou mène une vie solitaire. Il cherche plutôt à éviter la présence de congénères venant parasiter leurs sources de nourriture. Bien qu'indépendants, les tatous ne font preuve d'au-

> cune agressivité envers les autres individus. En général, les mâles défendent leur terrier, leurs sources de nourriture et leur femelle contre les autres mâles.

L'exemple (2) est extrait d'un article de l'Est Républicain. Du point de vue d'une analyse en relations de discours, les propositions « un véhicule a effectué une spectaculaire sortie de route » et « elle a quitté la chaussée sur sa droite » sont en relation d'élaboration, ce qui signifie que la seconde de ces propositions donne des détails sur la première. L'inférence de cette relation s'appuie sur des relations lexicales telles que « véhicule » / « voiture », « sortie » / « quitter » et « route » / « chaussée ».

(2) Un *véhicule* a effectué une spectaculaire *sortie* de *route*, hier vers 18h15, sur l'A36. La *voiture* circulait dans le sens Mulhouse-Montbéliard lorsqu'après être passée à hauteur du 35e RI, elle a *quitté* la *chaussée* sur sa droite.

Au terme de ces mini-analyses qui donnent une idée de la variété des phénomènes discursifs pouvant être abordés à partir d'une approche basée sur le lexique, on peut également noter que les relations sémantiques impliquées dans la cohésion lexicale sont de natures très variées, ce qui pose la question de leur repérage.

Notre objectif est d'exploiter une base de voisins distributionnels pour repérer la cohésion lexicale dans les textes. Les voisins distributionnels sont des paires de mots rapprochés par l'analyse distributionnelle automatique d'un corpus sur la base des contextes syntaxiques qu'ils partagent. L'analyse distributionnelle permet de capter une palette de relations sémantiques bien plus large que celle qui est habituellement recensée dans les dictionnaires ou visée par les outils d'acquisition sémantique.

Les deux principales hypothèses questionnées dans cette thèse sont ainsi les suivantes :

(a) les indices lexicaux constituent des éléments de signalisation de l'organisation du discours à différents niveaux ;

(b) une ressource acquise par analyse distributionnelle est adaptée pour appréhender ces indices lexicaux.

La vérification de ces hypothèses s'accompagne de la proposition de méthodes concernant la projection d'une ressource distributionnelle en texte (b) et les modalités d'exploitation des indices lexicaux ainsi identifiés (a).

Il s'agit donc, dans le domaine de l'analyse du discours, de prôner une prise en compte plus importante des aspects lexicaux ; de proposer une ressource pour ce faire ; de montrer par quelles méthodes on pourrait utiliser cette ressource dans différentes approches de l'organisation des textes.

Présentation du document

Dans la première partie de ce document, constituée du seul chapitre 1, nous dressons un panorama des recherches sur la cohésion lexicale et présentons notre problématique.

La seconde partie s'intéresse à la relation entre analyse distributionnelle et cohésion lexicale. Dans le chapitre 2, nous introduisons l'analyse distributionnelle et présentons la base distributionnelle que nous employons, les *Voisins de Wikipédia*. Nous donnons un aperçu de la richesse des relations lexicales qu'elle contient, mais montrons également qu'elle est fortement bruitée. Le chapitre 3 est consacré à la méthodologie de projection des voisins en texte et à la caractérisation des sorties obtenues. Enfin, le chapitre 4 aborde la question du filtrage des liens de voisinage projetés, afin d'aboutir à une détection de la cohésion lexicale qui soit la meilleure possible.

La troisième partie s'intéresse à la relation entre cohésion lexicale et organisation textuelle. En nous appuyant sur la méthodologie mise en place dans la deuxième partie, nous abordons différentes approches de l'organisation discursive, allant du très global vers le très local : le chapitre 5 est consacré à la segmentation thématique, le chapitre 6 à l'étude des structures énumératives, le chapitre 7 à la construction de la structure rhétorique du discours, et le chapitre 8 à l'étude d'une relation de discours en particulier, la relation d'élaboration.

Première partie

Contexte et problématique

Chapitre 1

Le projet VOILADIS, à l'interface du discours et du lexique

Sommaire

> L'air ou chant sous le texte, conduisant à la divination d'ici là, y applique son motif en fleuron et cul-de-lampe invisibles.
>
> Mallarmé, Le Mystère dans les lettres

Ce premier chapitre, essentiellement introductif, nous permet de poser la problématique de notre thèse en relation avec les recherches sur l'analyse du discours et plus précisément sur la cohésion lexicale. Après avoir présenté le champ de l'analyse du discours (section 1.1), nous nous focalisons sur le rôle de la cohésion lexicale : sur quelles relations lexicales repose-t-elle (section 1.2) ? Comment est-elle repérée, que ce soit du point de vue du lecteur ou dans une perspective de détection automatique (section 1.3) ? Et enfin, que peut-on retirer de son exploitation (section 1.4) ? Une fois ce panorama dressé, nous présentons les objectifs de notre travail de thèse, en les situant également dans le contexte de travaux antérieurs (ou partiellement contemporains) menés à Toulouse dans les laboratoires CLLE-ERSS et IRIT (section 1.5).

1.1 Analyse du discours

Cette thèse se situe dans le champ de l'analyse du discours. Dans cette section, nous introduisons un certain nombre de notions liées à ce champ de recherche : nous rappelons la relation entre texte et discours 1.1.1, évoquons la variété d'approches du discours à travers la distinction approches ascendantes / approches descendantes (section 1.1.2), et définissons les notions de cohérence et de cohésion (section 1.1.3.

1.1.1 Texte et discours

> Écrire quoi que ce soit, aussitôt que l'acte d'écriture exige de la réflexion, et n'est pas l'inscription machinale et sans arrêts d'une parole intérieure toute spontanée, est un travail de traduction [...]. L'esprit, au lieu d'épouser et de laisser s'émettre ce qui lui vient en réponse immédiate à ce qui l'excite, pense et repense [...] la chose qu'il veut exprimer, et qui n'est pas du langage ; et ceci, *en présence soutenue*, des conditions qu'il s'est données.
>
> Paul Valéry, Variations sur les Bucoliques

Nous précisons dans cette section les notions de *texte* et de *discours*. Le texte se définit comme la réalisation d'un discours dans un contexte particulier :

> [...] I view *text* (as a non-count noun) as denoting a typical instance of language *cum* other semiotic devices in use – i.e. occurring in some

> context and with the intention by the user of achieving some purpose or goal thereby. The term designates the connected sequence of verbal signs and non-verbal signals, vocal as well as non-vocal (i.e. visual, auditory, etc.) signals produced within the context of some utterance act. [...]
>
> *Discourse*, on the other hand, designates the hierarchically structured, mentally represented sequences of utterance and indexical acts which the participants are engaging in as the communication unfolds. Such sequences have as their *raison d'être* the accomplishment of some particular overall communicative goal [...] (Cornish (1999, pp. 33-34), cité par Péry-Woodley (2000, p. 13))

Alors que les représentations mentales qui forment le discours sont hiérarchiques et multi-dimensionnelles (« [The] mental representation is always multidimensional, i.e., has many relationships linking its elements. », Fayol, 1997, p. 157), le texte est strictement linéaire. Il en résulte que la production d'un texte implique des processus de linéarisation de l'information, et son interprétation des processus de délinéarisation :

> [...] the sequence you actually hear or see is like the tip of an iceberg – a tiny amount of matter an energy into which an enormous amount of information has been 'condensed' by a speaker or writer and is ready to be 'amplified' by a hearer or reader. If this transaction weren't so commonplace, it would be amazing : and we are still laboring to explain just how it can be done. (de Beaugrande, 1997, p. 11)

Une idée fondamentale pour ce qui suit est que les opérations de mise en texte laissent des traces à la surface des textes (Péry-Woodley, 2000). Ces traces peuvent être des signes verbaux ou non-verbaux – à l'écrit, les signes non-verbaux sont matérialisés par les propriétés visuelles de textes (Luc, 2000) –, et constituent une signalisation du discours, qui guide le lecteur dans son processus d'interprétation. L'étude de la signalisation du discours est essentielle pour la caractérisation linguistique de l'organisation textuelle, mais également dans une perspective de TAL ; les traces de la mise en texte sont utilisées comme des indices aiguillant la recherche de la structure sous-jacente des textes.

1.1.2 Différentes approches du discours

En partant du texte et des indices qui le jalonnent (traces de la mise en texte), il existe plusieurs manières d'aborder sa structure sous-jacente. Deux grands types d'approches peuvent être distinguées :

- les approches ascendantes qui visent une représentation complète du discours, vu comme un ensemble d'unités élémentaires reliés par des relations de cohérence (ou rhétoriques) ;

– les approches descendantes qui visent à identifier des structures discursives de haut niveau, avec l'hypothèse que ces dernières ont un impact sur l'interprétation locale des textes.

Ces types d'approches correspondent à des logiques différentes dans la façon d'envisager l'organisation discursive, comme l'exprime Widlöcher (2008, p. 38-39)

> Visant l'accès au plan du texte et à des intentions signifiantes, certaines approches en viendront à définir un ensemble de règles pouvant idéalement correspondre à des règles d'écriture ayant présidé à la production du texte. D'autres [...] entrent dans une logique de la manifestation et dans l'étude de mécanismes indiciaires permettant de mettre à jour la structure du texte sans faire d'hypothèses sur ses origines, et concentrent leur attention sur l'étude de traces, laissées peut-être de manière involontaire, mais pouvant enrichir, de manière incrémentale, une représentation de la structure possible du texte. D'un côté, on cherchera à justifier, de manière souvent ascendante et exhaustive, chaque élément de la structure du texte, en pensant le global depuis le local ; de l'autre, de manière plus descendante et en ciblant différents indices complémentaires plus diffus, on cherchera à rendre compte de phénomènes globaux, en ignorant ce qui, du point de vue de la compréhension de ces phénomènes, est insignifiant. (Widlöcher, 2008, p. 38-39)

1.1.3 Cohérence, cohésion

> Nous marchions, et il lui échappait des phrases presque incohérentes. Malgré mes efforts, je ne suivais ses paroles qu'à grand-peine, me bornant enfin à les retenir. L'incohérence d'un discours dépend de celui qui l'écoute. L'esprit me paraît ainsi fait qu'il ne peut être incohérent pour soi-même. Aussi me suis-je gardé de classer Teste parmi les fous.
>
> Paul Valéry, Monsieur Teste

Les notions de cohérence et de cohésion ont été largement discutées dans la littérature. La cohérence a rapport à l'interprétation des textes par le lecteur, comme l'explique Charolles (1997, p. 3) :

> L'interprétation des discours est soumise à un principe général de cohérence (Charolles, 1983, 1995) ou de pertinence (Sperber et Wilson, 1986) qui est de nature fondamentalement sémantique et pragmatique. Confronté à une séquence d'énoncés produits à la suite, le destinataire ne peut en effet que chercher à établir des relations entre ces énoncés, vu que, précisément, ils sont énoncés à la suite. L'établissement de ces liens fait

> appel à des opérations intellectuelles de haut niveau dans laquelle interviennent toutes sortes de compétences linguistiques et non linguistiques. Pour guider l'interlocuteur dans le processus de résolution de problèmes, le locuteur a à sa disposition un vaste ensemble de marques de cohésion qui codent des instructions relationnelles plus ou moins spécifiques.

La cohérence n'est ainsi pas tant une propriété des discours qu'un principe général gouvernant leur interprétation Ducrot (1972). La cohésion est une propriété linguistique des textes. Le concept de cohésion englobe un ensemble de marques qui permettent de relier entre elles les phrases d'un texte, créant sa *texture* (Halliday et Hasan, 1976). Les marques de cohésion participent ainsi à la signalisation du discours. Les procédés cohésifs relèvent de deux grandes catégories : la cohésion grammaticale et la cohésion lexicale. Nous détaillons ci-dessous les procédés relevant de la cohésion grammaticale.

- La référence est réalisée par les pronoms, articles démonstratifs ou adverbes dont l'interprétation dépend de leur rattachement à un autre élément texte (référence endophorique) ou du contexte extra-linguistique (référence exophorique) ; dans l'exemple (1), *they* fait référence à *three blind mice*, ce qui produit un lien cohésif.

  (1) Three blind mice, three blind mice.
  See how they run ! See how they run !
  (Halliday et Hasan, 1976, p. 31)

- La substitution est le remplacement d'un item par l'un des items suivants : one, ones, same (substitution nominale) ; do (substitution verbale) ; no, not (substitution clausale). L'ellipse est un forme particulière de substitution dans laquelle l'item est simplement omis. L'exemple (2) montre une substitution nominale (*one* remplace *axe*) ; l'exemple (3) montre une substitution verbale (*does* remplace *knows*) ; enfin, l'exemple (4) montre une ellipse verbale (*brought* est omis dans la seconde proposition).

  (2) My axe is too blunt. I must get a sharper one. (Halliday et Hasan, 1976, p. 89)

  (3) You think Joan already knows ? – I think everybody does. (Halliday et Hasan, 1976, p. 89)

  (4) Joan brought some carnations, and Catherine some sweet peas. (Halliday et Hasan, 1976, p. 143)

- La conjonction relie deux phrases ou groupes de phrases par l'emploi d'adverbes ou de syntagmes prépositionnels (exemple (5)) ; la relation établie par une conjonction peut être additive (a.), adversative (b.), causale (c.) ou temporelle (d.). La conjonction regroupe ainsi toute une gamme de marqueurs discursifs.

(5) For the whole day he climbed up the steep mountainside, almost without stopping.

a. And in all this time he met no one.

b. Yet he was hardly aware of being tired.

c. So by night time the valley was far below him.

d. Then, as dusk fell, he sat down to rest.

(Halliday et Hasan, 1976, p. 238-239)

La cohésion lexicale, à laquelle sont consacrées les sections qui suivent, est réalisée par des relations sémantiques entre items lexicaux du texte.

1.2 Nature des relations lexicales impliquées dans la cohésion lexicale

Dans leur chapitre consacré à la cohésion lexicale, Halliday et Hasan (1976) distinguent deux grands types de phénomènes cohésifs :
- La réitération ;
- la collocation.

1.2.1 La réitération

Dans sa définition stricte, la réitération est la reprise d'un item lexical par un autre, les deux items ayant le même référent. Le phénomène de réitération ne se limite pas à la répétitions d'items lexicaux et peut être réalisé par un continuum de relations lexicales : synonymie, quasi-synonymie, hypéronymie ou reprise par un « nom général » (*general noun*), comme illustré dans l'exemple (6).

(6) I turned to the ascent of the peak. $\left\{ \begin{array}{c} \text{The ascent} \\ \text{The climb} \\ \text{The task} \\ \text{The thing} \\ \text{It} \end{array} \right\}$ is perfectly easy.

(Halliday et Hasan, 1976, p. 279)

La classe des noms généraux contient des noms qui se trouvent au sommet des hiérarchies lexicales (par exemple : *person, man, thing, stuff*). Ces mots se trouvent à la frontière entre lexique et grammaire (dans la mesure où il s'agit d'une classe fermée) (Halliday et Hasan, 1976, p. 274). Ainsi, il n'y a pas de limite stricte entre ce continuum de relations lexicales et la reprise par un pronom tel que « it », dans

l'exemple donné. La frontière entre la réitération (cohésion lexicale) et la référence (cohésion grammaticale) est donc mince.

Si l'on s'en tient à cette première définition, la réitération est un phénomène bien délimité, impliquant :

- un ensemble fermé de relations lexicales ;
- un phénomène précis au sein du texte : la coréférence de deux items.

Toutefois, la correspondance entre ces deux éléments n'est pas systématique. En effet, il est tout à fait possible qu'un item soit la répétition d'un autre sans pour autant établir de lien de coréférence, comme le montre l'exemple (7).

(7) There's a boy climbing that tree.
a. The boy's going to fall if he doesn't take care.
b.Those boys are always getting into mischief.
c. And there's another boy standing underneath.
d. Most boys love climbing trees.
(Halliday et Hasan, 1976, p. 283)

Dans cet exemple, la relation référentielle entre les deux occurrences de « boy » évolue : (a.) identique ; (b.) inclusive ; (c.) exclusive ; (d.) indépendante. Mais selon Halliday et Hasan (1976, pp. 283-284), la variété des relations référentielles établies n'a aucun impact sur le phénomène de cohésion lexicale ; en effet, dans l'exemple (7), la répétition de « boy » crée un lien cohésif même dans le cas où sa référence est indépendante de celle de la première occurrence. Ainsi, si la possibilité de coréférer est à la base de la distinction entre relations de réitération et de collocation, Halliday et Hasan (1976, p. 284) concluent toutefois à la non-pertinence du lien entre référence et cohésion lexicale :

> Properly speaking, reference is irrelevant to lexical cohesion. It is not by virtue of any referential relation that there is a cohesive force set up between two occurrences of a lexical item ; rather, the cohesion exists as a direct relation between the forms themselves (and thus is more like substitution than reference).

La réitération se trouve ainsi réalisée par n'importe quel couple d'items lexicaux dont la relation lexicale qu'ils entretiennent (répétition, synonymie ou quasi-synonymie, hypéronymie ou reprise par un nom général) les rend potentiellement coréférentiels (ou substituables), sans égard à la réalisation effective de ce lien de coréférence.

> ... lexical reiteration takes place not only through repetition of an identical lexical item, but also through occurrence of a different lexical item that is systematically related to the first one, as a synonym or superordinate of it. This principle applies quite generally, irrespective of whether or not there is identity of reference. (Halliday et Hasan, 1976, p. 284)

On peut enfin remarquer que la relation de réitération apparaît comme une relation « en langue », puisqu'elle est décrite comme systématique et indépendante du texte (« systematically related », « a direct relation between the forms themselves », etc.).

1.2.2 La collocation

La relation de collocation est concernée par la « cohesion that is achieved through the association of lexical items that regularly co-occur » (Halliday et Hasan, 1976, p. 284). Alors que la réitération est une relation bien définie, et les sous-types de réitérations énumérés de manière exhaustive, la collocation, qualifiée de « the most problematical part of lexical cohesion », reste une notion assez vague, définie essentiellement par contraste avec la réitération :

> Here we shall simply group together (...) all the lexical cohesion that is not covered by what we have called 'reiteration' and treat it under the general heading of collocation, or collocational cohesion, without attempting to classify the various meaning relations that are involved. (Halliday et Hasan, 1976, p. 287)

Ainsi, toutes les relations lexicales qui participent à la cohésion lexicale d'un texte mais qui ne s'accompagnent pas d'une possible identité de référence tombent dans cette catégorie. La très vaste catégorie des collocations recouvre ainsi un grand nombre de relations, qui sont rapidement énumérées et exemplifiées :

- des relations dites « systématiques », qui peuvent notamment concerner :
  - des antonymes : *boys / girls*, *stand up / sit down*) ;
  - des paires de mots appartenant à une même série ordonnée : *Tuesday / Thursday*, *colonel / brigadier* ;
  - des paires de mots tirés d'ensembles non-ordonnés : *red / green* ;
  - un méronyme et son holonyme : *car / brake*, *box / lid* ;
  - des co-méronymes : *mouth / chin*, *verse / refrain* ;
  - des co-hyponymes : *chair / table*, *walk / drive* ;
- des relations qualifiées de « non systématiques », entre mots ayant tendance à partager le même environnement : *laugh / joke*, *ill / doctor*, *try / succeed*, *bee / honey*, etc.

L'étiquette « collocation » crée une certaine confusion, car ce terme évoque la collocation en tant qu'association statistiquement significative de deux mots, au sens de Firth (1957); Sinclair (1966). La catégorie des collocations chez Halliday et Hasan contient aussi bien des relations paradigmatiques classiques (antonymie, méronymie...) que des relations plus syntagmatiques, mais dont rien ne garantit qu'elles relèvent d'une association statistique.

Le caractère fourre-tout de la classe des collocations a été largement souligné, notamment par Hoey (1991, p. 7) :

Under this heading, Halliday and Hasan include a ragbag of lexical relations, many of which have no readily name.

1.2.3 Après Halliday et Hasan

D'autres typologies des relations lexicales porteuses de cohésion ont été proposées, notamment par Hasan (1984); Hoey (1991); Martin (1992) ; etc. Pour un panorama, on peut se référer à Tanskanen (2006, pp. 31-49), dont nous reproduisons la table récapitulative dans la figure 1.1. Nous nous contentons ici de formuler deux

Haliday & Hasan 1976	**Hasan 1984**	**Haliday 1985**	**McCarthy 1998**	**Morris & Hirst 1991**	**Hoey 1991**	**Martin 1992**	**Tanskanen 2006**
Reiteration : same word (repetition)	General : repetition	Repetition		Reiteration	Repetition : simple and complex	Taxonomic : repetition	Reiteration : simple and complex repetition
					Repetition : substitution		Reiteration : substitution
Reiteration : synonymy	General : synonymy	Synonymy	Equivalence		Repetition : simple paraphrase	Taxonomic : synonymy	Reiteration : equivalence
Reiteration : superordinate		Synonymy : superordinate	Inclusion : specific – general	Reiteration : superordinate	Repetition : superordinate	Taxonomic : hyponymy	Reiteration : generalisation
Collocation	General : hyponymy	Synonymy : hyponymy		Collocation : systematic semantic relation	Repetition : hyponymy		
	General : meronymy	Synonymy : meronymy	Inclusion : specific – general			Taxonomic : meronymy	Reiteration : specification
		Synonymy : co-hyponymy and co-meronymy				Taxonomic : co-hyponymy and co-meronymy	Reiteration : co-specification
	General : antonymy	Synonymy : antonymy	Opposition		Repetition : complex repetition or paraphrase	Taxonomic : contrast	Reiteration : contrast
		Collocation			Repetition : closed set*		Collocation : ordered set
				Collocation : non-systematic semantic relation	(Repetition : complex paraphrase)	Nuclear : extending and enhancing)	Collocation : activity-related
							Collocation : elaborative

* This category is introduced in Hoey (1994).

FIGURE 1.1 – Catégories de relations lexicales chez Halliday et Hasan (1976); Hasan (1984); Halliday (1985); McCarthy (1988); Morris et Hirst (1991); Hoey (1991); Martin (1992); Tanskanen (2006), synthétisées par Tanskanen (2006, p. 48)

remarques.

(a) La classe des collocations au sens d'Halliday et Hasan continue à être utilisée dans l'opposition de base *reiteration / collocation* (Tanskanen, 2006) ou itération / collocation (Legallois, 2004). Cette catégorisation semble donc fonctionner. Toutefois, on observe un glissement de sens du terme « collocation » ; alors qu'il désignait chez Halliday et Hasan l'ensemble des relations n'impliquant pas de possible identité de référence, dans les typologies plus récentes il ne désigne plus

que les relations non systématiques. Cette évolution a été amorcée dès Hasan (1984), qui remplace la classe des réitérations par une catégorie étiquetée *General*, qui inclut, aux côtés de la répétition, de la synonymie et de l'hypéronymie, les relations d'antonymie et de méronymie (relevant précédemment de la collocation). Plus récemment, chez Tanskanen (2006) et Legallois (2004), la classe des réitérations (ou itérations) inclut également ces relations, la spécification (méronymie) et la contradiction (antonymie et relations converses) constituant des types de réitérations.

(b) L'affinement de la typologie d'Halliday et Hasan (ou la proposition d'autres typologies) s'est très généralement accompagnée de plus de réticences vis-à-vis de la classe des collocations, qui se voit réduite à certaines relations bien délimitées (par exemple les collocations liées à des activités chez Martin (1992)). Les collocations sont jugées difficiles à appréhender et se retrouvent exclues de l'analyse. Ainsi, Hasan (1984, p. 195) considère que les collocations devraient être ignorées à cause de leur caractère trop subjectif :

> Altogether the notion of collocation proved problematic. While I firmly believe that behind the notion of collocation is an intuitive reality, I have come to accept the fact that unless we can unpack the details of the relations involved in collocation in the Firthian sense, it is best to avoid the category in research. The problems of inter-subjective reliability cannot be ignored. If someone felt that there is a collocational tie between *dive* 4 and *sea* 6 in A13 [texte analysé par Hasan], on what grounds could such a statement be either rejected or accepted ?

Hoey (1991), dont la typologie admet des relations proches de la collocation par le biais de la relation de « paraphrase complexe »[1], restreint cette relation à certains cas bien identifiés (relevant notamment de l'antonymie). Le problème mis en avant n'est pas cette fois la subjectivité, mais la trop forte productivité d'une telle relation :

> *Complex paraphrase* is, however, a can of lexical worms and for the purpose of our analysis will be recognized very conservatively. [...][It] will relate vast numbers of lexical items (for example, **sickness** and **doctor**, **carol** and **Christmas**) in ways that might be revealing about lexis but are not readily controllable [...]. (Hoey, 1991, p. 64)

Ainsi, la classe des collocations, foisonnante chez Halliday et Hasan, se voit finalement réduite à la portion congrue, certaines des relations qu'elle couvrait tombant sous le champ de la réitération, et d'autres se retrouvant rejetées car trop problématiques.

1. qui fait intervenir des relations triangulaires : par exemple *hot* est lié à *cold* par une relation de paraphrase simple, *hot* est lié à *heat* par une relation de répétition complexe, donc *cold* est lié à *heat* par une relation de paraphrase complexe.

Cependant, il apparaît que les relations lexicales non systématiques, qui vont au delà des relations habituellement recensées, ont un rôle prépondérant dans la perception de l'organisation discursive. Morris et Hirst (2004); Morris (2007) affirment la nécessité de prendre en compte les relations qu'ils qualifient de « non-classiques », et dont ils identifient trois principaux types :

- les relations entre mots au sein des catégories non-classiques de Lakoff (1987). Ces relations ne peuvent se nommer que par référence à la catégorie à laquelle ils appartiennent ; ainsi, *ball*, *field* and *umpire* sont reliés car ils font partie de l'activité *baseball* ;
- les relations thématiques (*case relations*) qui associent un prédicat à un actant typique (agent, objet ou instrument) : « chien » / « aboyer », « conduire » / « véhicule », etc.
- les relations relevant des *related terms* (Neelameghan et Rao, 1976; Neelameghan, 2001), auxquelles font appel des disciplines comme les sciences de l'information, et qui sont concernées par une très large gamme de relations associatives.

En proposant à des lecteurs d'identifier des groupes de mots reliés au sein de textes, Morris et Hirst (2004) montrent que ces derniers recourent massivement aux relations non-classiques, qui éclipsent les relations telles que la synonymie et l'hypéronymie (62% des relations annotées sont non-classiques, *cf.* Morris, 2007, p.198).

1.3 Capter la cohésion lexicale

1.3.1 Subtilité de la cohésion lexicale

De la variété des relations lexicales impliquées, il découle que les phénomènes relevant de la cohésion lexicale sont très difficiles à appréhender, que ce soit du point de vue du lecteur ou dans une perspective d'automatisation. Outre la nature des relations lexicales impliquées (dont on a vu que certaines sont ténues et subjectives), d'autres aspects entrent en jeu dans l'établissement d'une relation de cohésion lexicale. Ainsi, Halliday et Hasan (1976, pp. 289-290) citent deux facteurs qui ont une influence sur la force d'un lien collocationnel :

> The relative strength of the collocational tension is really function of two kinds of relatedness, one kind being relatedness in the linguistic system and the other being relatedness in the text. [...] There are degrees of proximity in the lexical system, a function of the relative probability with which one word tends to co-occur with another. Secondly, in the text there is relatedness of another kind, relative proximity in the simple sense of the distance separing one item from another, the number of words or clauses or sentences in between. The cohesive force that is exerted between any pair of lexical items in a passage of discourse is a function of their relative proximity in these two respects.

Ainsi, l'établissement d'une relation de cohésion lexicale ne dépend pas seulement de la relation « en langue » (« in the linguistic system ») entre les deux items, qui est déjà très relative[2], mais également d'une proximité dans le texte.

Ce qui transparaît ici et est appréhendé à travers le critère de la distance en texte, c'est la nature bidirectionnelle de la relation entre lexique et texte, exprimée ainsi par Hoey (1991, p. 8) :

> The text provides the context for the creation and interpretation of lexical relations, just as the lexical relations help create the texture of the text.

Un exemple d'instanciation textuelle d'une relation lexicale est donné par (Hoey, 1991, p. 220) :

> But it is not only the case that text is lexically signalled ; it is also the case that lexis is textually established. A pair of clauses, 'A was x, but B was only y, create an instantial contrastive relation between the lexical items x and y

Ce principe selon lequel les relations lexicales peuvent être instaurées par le discours relève d'une *discourse-based approach to lexis* (Nyyssönen, 1992). Certaines relations de cohésion lexicale sont très difficiles à anticiper car elles sont créées au sein d'une énonciation particulière et ne peuvent donc pas être répertoriées *a priori*, comme l'exprime (Mortureux, 1993) :

> Mais il convient de distinguer nettement les relations inscrites en langue (entre lexèmes correspondant aux vocables) de celles qu'instaure, à proprement parler, le discours lui-même ; les premières, en principe, sont répertoriées dans les dictionnaires de langue, et disponibles en permanence, quelle que soit la situation d'énonciation, tandis que les secondes sont liées étroitement à une énonciation donnée ; cette différence, on le verra, est pertinente pour l'interprétation des discours.

1.3.2 La notion de chaîne lexicale

Le repérage de relations cohésives fait souvent intervenir la notion de « chaînes lexicales » : « chains of cohesion » (Halliday et Hasan, 1976, p. 287), puis « lexical chains »(Morris et Hirst, 1991; Barzilay et Elhadad, 1997, etc.). Une chaîne lexicale est une séquence de mots liés entre eux ; elle se caractérise par sa densité (le nombre de maillons qui la composent) et sa longueur (son étendue dans le texte) ; à un point donné du texte, une chaîne lexicale est dite active si elle couvre ce point. De nombreuses méthodes de construction de chaînes lexicales ont été proposées.

2. « The relatedness is a matter of more or less ; there is no clearly defined cutoff point such as we can say that *sunset*, for example, is related to just this set of words and no other. » (Halliday et Hasan, 1976, p. 289)

1.3.3 Repérage de la cohésion lexicale par le lecteur / analyste

Un nombre important d'analyses faisant appel à la cohésion lexicale se sont basées sur un repérage manuel des liens, effectué directement par les auteurs en utilisant leurs propres procédures et ensembles de relations autorisées (Halliday et Hasan, 1976; Hasan, 1984; Hoey, 1991; Martin, 1992), ce qui ne permet pas d'évaluer précisément la complexité de cette tâche. Des difficultés sont cependant soulignées ; on a vu ainsi vu qu'elles ont pu mener à l'exclusion des relations relevant de la collocation, jugées particulièrement problématiques :

> The effect of lexical, especially collocational, cohesion on a text is subtle and difficult to estimate. (Halliday et Hasan, 1976, p. 288)

Plus récemment, quelques études ont permis d'apprécier la complexité et la subjectivité de la détection de la cohésion lexicale du point de vue des lecteurs.

Dans l'étude de Morris et Hirst (2004, 2005) , des lecteurs (5 sujets) sont confrontés au repérage de liens de cohésion lexicale à travers différentes tâches, dont notamment :

1. l'identification de groupes de mots associés dans un texte (donc de chaînes lexicales) en utilisant des crayons de couleur (chaque groupe est souligné avec une couleur différente) ;
2. l'identification des liens lexicaux à l'intérieur des groupes : pour chaque groupe, les lecteurs doivent préciser quelles paires de mots sont reliées ;

L'identification des principaux groupes a donné lieu à un bon accord selon les auteurs (« the subjects were in broad agreement about many of the groups of related words ») ; l'accord moyen sur les mots appartenant à un groupe donné est de 63% ; l'accord moyen sur les paires identifiées dans chaque groupe est de 13%. Ainsi, le repérage de liens de cohésion lexicale à proprement parler, c'est-à-dire de relations entre paires de mots dans un texte, donne ici lieu à un accord inter-annotateur très faible. Il faut bien sûr noter qu'il s'agit ici d'une tâche très ouverte, dont la vocation est de mettre au jour les relations sur lesquelles s'appuient des lecteurs pour interpréter un texte, et non d'annoter des liens de cohésion lexicale selon une typologie prédéfinie, ce qui explique partiellement cet accord très faible.

Hollingsworth et Teufel (2005) ont également proposé des évaluations de l'accord inter-annotateur pour l'annotation de chaînes lexicales. Dans leur étude, 3 sujets (dont les deux auteurs de l'article) annotent des chaînes lexicales en suivant une procédure assez similaire à celle de Morris et Hirst (2004, 2005). Hollingsworth et Teufel (2005)s'attachent davantage à la question méthodologique de l'appariement entre chaînes, en comparant différentes mesures d'accord ; les accords dont ils font état sont meilleurs que ceux reportés par Morris et Hirst (2004), mais restent assez bas.

Enfin Beigman Klebanov et Shamir (2006) proposent une exploration des relations de cohésion lexicale perçues par des lecteurs dans une approche différente, qui

ne repose pas sur le repérage global de chaînes lexicales, mais sur des phénomènes locaux d'ancrage (*anchoring*) : pour chaque mot d'un texte, le lecteur est invité à sélectionner parmi les mots qui précèdent ceux qui lui sont rattachés. La figure 1.2 montre un exemple d'annotation suivant une telle procédure. Pour cette expérience d'annotation impliquant 22 sujets, les auteurs reportent des coefficients d'accord inter-annotateurs compris entre 0.41 et 0.55.

the	be
stranger	sure # → maybe
albert	telegram → died
camus → { albert stranger}	from #
mother	home → mother
died	says
today	your #
or	passed
maybe	away
yesterday → today	funeral → { passed_away died }
i	tomorrow → { yesterday today }
can't	...

Possible annotation of the beginning of *The Stranger*. The notation $x \rightarrow \{ c\, d \}$ means each of c and d is an anchor for x, and $x \rightarrow \{c\_d\}$ means c and d together anchor x.

FIGURE 1.2 – Exemple d'annotation selon la procédure proposée par Beigman Klebanov et Shamir (2007, p. 32)

1.3.4 Détection automatique de la cohésion lexicale

Dans une perspective d'automatisation, en plus de la méthode de détection, le repérage des liens de cohésion lexicale pose la question de la ressource à mobiliser.

Morris et Hirst (1991) furent les premiers à proposer une méthode d'automatisation de la détection de chaînes lexicales, basée sur l'exploitation du *Roget's International Thesaurus* (*4th Edition*, 1977). Dans la mesure où ce dernier n'était pas disponible dans une version électronique, leur algorithme (reproduit dans la figure 1.3) a toutefois été appliqué manuellement. L'usage du thésaurus leur permet de prendre en considération des relations classiques (« systematic semantic relations ») et non-classiques (« nonsystematic semantic relations »). Ils font état de davantage de difficultés concernant ces dernières relations, qui donnent lieu à quelques faux positifs (des relations détectées mais qui ne semblent pas pertinentes), et à un nombre plus important de faux négatifs, dont une partie est due à un manque de connaissances sur le texte (« situational knowledge »), ce qui rejoint le problème que nous avons souligné concernant les relations instaurées en discours.

Par la suite, WordNet (Fellbaum, 1998) a été massivement utilisé dans le cadre d'approches automatiques de la détection de la cohésion lexicale, avec des varia-

```
REPEAT
      READ next word
      IF word is suitable for lexical analysis (see section 3.2.1) THEN
            CHECK for chains within a suitable span
            (up to 3 intermediary sentences, and no limitation on
            returns):
                  CHECK thesaurus for relationships (section 3.2.2).
                  CHECK other knowledge sources
                  (situational, general words, proper names).
            IF chain reslationship is found THEN
                  INCLUDE word in chain.
                  CALCULATE chain so far
                  (allow one transitive link).
            END IF
            IF there are words that have not formed a chain for a suitable
            number of sentences (up to 3) THEN
                  ELIMINATE words from the span.
            END IF
            CHECK new word for relevance to existing chains that
            are suitable for checking.
            ELIMINATE chains that are not suitable for checking.
      END IF
END REPEAT
```

FIGURE 1.3 – Algorithme de construction des chaînes lexicales (Morris et Hirst, 1991, p. 34)

tions dans les relations prises en compte (Barzilay et Elhadad, 1997; Hirst, 1997; Hirst et St-Onge, 1998; Stokes, 2004; Yang et Powers, 2006). Plutôt que d'utiliser directement les relations recensées par WordNet, certaines approches considèrent le réseau formé par cette ressource et évaluent la proximité sémantique entre deux mots en se fondant sur les chemins qui les relient dans ce réseau (*cf.* Budanitsky et Hirst, 2006, pour un panorama des approches utilisant WordNet). La modélisation de la proximité sémantique est un domaine de recherche très actif, et de nombreuses mesures ont été proposées, utilisant très souvent WordNet, mais également d'autres sources d'information. Par exemple, Sporleder *et al.* (2010) construisent des chaînes lexicales en exploitant une mesure nommée *Normalized Google Distance* (Cilibrasi et Vitanyi, 2007), qui est fondée sur le nombre de pages retournées par le moteur de recherche Google lorsqu'on lui soumet une paire de mots en requête. La similarité distributionnelle est également parfois utilisée (Marathe et Hirst, 2010). Ces mesures présentent l'avantage de permettre la prise en compte de relations non-classiques.

1.4 Exploiter la cohésion lexicale

Les chaînes lexicales ont été exploitées dans de nombreuses applications, comme la désambiguïsation lexicale (Okumura et Honda, 1994; Mihalcea et Moldovan, 2001; Ion *et al.*, 2011), la correction automatique « intelligente » (sensible au contexte) (Hirst et St-Onge, 1998), le résumé automatique (Barzilay *et al.*, 1999; Silber et McCoy, 2002), etc. La plupart de ces applications traitent les chaînes lexicales comme des sacs de mots associés à un texte, sans s'intéresser à leur structure et à leur position dans le texte. Par exemple, l'exploitation de chaînes lexicales pour la correction automatique repose sur le principe suivant : si un mot n'appartient à aucune chaîne calculée à partir du texte, et est orthographiquement proche d'un mot qui pourrait être intégré à une chaîne, ce mot est peut-être une faute.

Nous nous intéressons ici plus particulièrement aux travaux qui exploitent la cohésion lexicale dans le cadre d'approches de l'organisation discursive.

1.4.1 Prépondérance de la cohésion lexicale dans la signalisation du discours

Le rôle important de la cohésion lexicale dans l'élaboration de l'organisation discursive a été souligné (Hoey, 1991, « lexis organizes text »,).

La première raison de cette importance est la forte couverture des marques de cohésion lexicale. Déjà dans les analyses de textes proposées par Halliday et Hasan (1976, pp. 340-355) à la fin de leur livre, les liens de cohésion lexicale représentent près de la moitié de l'ensemble des liens cohésifs. À l'inverse des autres types de liens cohésifs, la présence de liens de cohésion lexicale est présentée comme indispensable à la formation d'un texte :

> However luxuriant the grammatical cohesion displayed by any piece of discourse, it will not form a text unless this is matched by cohesive patterning of a lexical kind. (Halliday et Hasan, 1976, p. 292)

Une autre propriété distingue la cohésion lexicale des autres marques de cohésion : sa capacité à former des relations multiples : un item lexical est généralement connecté à plus d'un autre item, tandis que les liens de cohésion grammaticale connectent seulement deux items (à part parfois la relation de référence). Cette propriété de la cohésion lexicale est essentielle car elle rend possible la formation de motifs (*patterns*), qui font de la cohésion lexicale le principal vecteur de texture :

> Lexical cohesion is the only type of cohesion that regularly forms multiple relationships [...]. If this taken into account, lexical cohesion becomes the dominant mode of creating texture. In other words, the study of the greater part of cohesion is the study of lexis [...]. (Hoey, 1991, p. 10)

Enfin, une dernière propriété intéressante de la cohésion lexicale est sa capacité à former des liens distants (*long-distance lexical relations*), pouvant établir des

connections entre des zones de texte très éloignées : cette caractéristique de la cohésion lexicale a été analysée par Phillips (1985), qui montre que les connections lexicales entre différents chapitres d'un livre peuvent être utilisées pour mettre au jour des aspects organisationnels.

1.4.2 Analyses du rôle de la cohésion lexicale dans des approches de l'organisation discursive

Martin (1992); Parsons (1996); Fang et Cox (1998) ont montré que la cohésion lexicale était corrélée avec la cohérence, en utilisant des textes rédigés par des sujets appartenant à différents groupes d'âges. Morris et Hirst (1991) ont montré des correspondances entre les chaînes lexicales qu'ils détectent et une analyse de la structure discursive de leurs textes selon le modèle de Grosz et Sidner (1986).

Hoey (1991, 1994, 1997) propose une approche de l'organisation textuelle uniquement basée sur le lexique, principalement développée dans son livre *Patterns of Lexis in Texts*. Cette approche ne repose pas sur l'identification de chaînes lexicales, mais sur des réseaux phrastiques (*nets*), qui exploitent le nombre de liens lexicaux entretenus par chaque phrase avec les autres phrases du texte. Chaque item lexical d'une phrase ne peut être connecté qu'avec un seul item par autre phrase.

Nous reprenons le premier exemple proposé par Hoey (1991, pp. 35-41) dans la figure 1.4. La matrice recensant les nombres de liens entre chacune des 5 phrases de ce texte est répliquée dans le tableau 1.1. On peut par exemple voir que la phrase 5 est connectée à la phrase 1 par 4 liens lexicaux : il s'agit des répétitions de *drug*, *grizzly* et *bears*, et de la relation de paraphrase entre *produce* et *responsible*. À partir de ces analyses, Hoey montre que certaines phrases fortement connectées peuvent être considérées comme centrales, alors que d'autres sont jugées marginales. Les motifs formés par les réseaux phrastiques permettent d'interpréter certaines phrases comme constituant des ouvertures ou des clôtures thématiques (*topic opening* / *topic closing*).

	1				
2	4	2			
3	2	2	3		
4	5	5	2	4	
5	4	5	1	2	5

TABLEAU 1.1 – Matrice obtenue par l'analyse du texte 1.4

La méthode de Hoey (1991) a inspiré de nombreuses analyses, notamment Károly (2002), et, pour le français, Legallois (2003, 2004). Elle a également été appliquée au résumé automatique (Benbrahim et Ahmad, 1994; Benbrahim, 1996) et à la segmentation thématique (Sardinha, 2001).

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.

FIGURE 1.4 – Exemple tiré de *Patterns of lexis in texts* (Hoey, 1991, p. 35)

1.4.3 Exploitation d'indices lexicaux dans des approches automatiques du discours

Les approches descendantes du discours (visant l'identification de segments de haut niveau) qui utilisent des indices lexicaux les associent particulièrement à la notion de cohérence thématique. Une forte cohésion lexicale est perçue comme un signe d'homogénéité thématique. La tâche de segmentation thématique (Hearst, 1997), se limitant à ce seul plan de cohérence thématique, propose un découpage des textes en segments uniquement basé sur le critère de cohésion lexicale.

Concernant les approches ascendantes du discours, de nombreux travaux qui essaient de prédire des relations rhétoriques entre unités discursives utilisent une notion de proximité lexicale (Harabagiu et Maiorano, 1999; Marcu, 2000; Wellner et Pustejovsky, 2007; Subba et Di Eugenio, 2009). Toutefois, la cohésion lexicale n'est jamais un indice central (elle est un indice parmi d'autres) et son impact n'est pas précisément évalué.

Nous reviendrons plus en détail sur ces différentes approches du discours et sur leur prise en compte de l'information lexicale dans la III^{e} partie de cette thèse.

1.5 Le projet VOILADIS : problématique et objectifs

Dans cette section, nous présentons notre projet de thèse, le projet VOILADIS[3] Nous évoquons tout d'abord le contexte scientifique dans lequel il a vu le jour, en présentant les recherches qui l'ont précédé et motivé au sein des laboratoires CLLE-ERSS et IRIT (section 1.5.1). Nous précisons ensuite sa problématique (section 1.5.2).

1.5.1 Héritages et motivations

1.5.1.1 Le projet ANNODIS

L'analyse du discours a fait l'objet d'un projet ANR, le projet ANNODIS (2007-2010) (Péry-Woodley, 2000), coordonné par M.-P. Péry-Woodley et réunissant trois laboratoires de recherche : CLLE-ERSS, Toulouse ; IRIT, Toulouse et GREYC, Caen. Ce projet s'est concentré sur la construction d'un corpus de textes annotés au niveau discursif pour le français et sur le développement d'outils pour l'annotation et l'exploitation de corpus. Le corpus constitué dans le cadre d'ANNODIS contient notamment des textes issus de Wikipédia.

Une des originalités du projet ANNODIS a été d'aborder le discours à partir de deux perspectives complémentaires :

- l'approche descendante (aussi appelée « macro ») considère les textes du point de vue de leur organisation globale ; elle se base sur des faisceaux d'indices pré-marqués pour identifier des structures de haut niveau ;
- l'approche ascendante analyse les discours de façon compositionnelle : elle part des unités minimales du discours pour en construire la structure par le biais de relations rhétoriques.

Notre projet de recherche prend en partie sa source dans la volonté d'explorer les aspects lexicaux liés aux structures discursives annotées dans le cadre d'ANNODIS. Il peut être vu comme une direction de recherche complémentaire, mais qui s'affranchit largement du projet ANNODIS de par les problématiques différentes qu'elle engendre.

1.5.1.2 Les voisins distributionnels

Un programme d'analyse distributionnelle automatique, UPÉRY, a été développé au sein du laboratoire CLLE-ERSS. L'analyse distributionnelle automatique calcule des rapprochements sémantiques entre mots d'un corpus sur la base des contextes syntaxiques qu'ils partagent, afin de capter des classes lexicales opérant dans ce corpus ; les paires de mots ainsi formées sont nommées « voisins distributionnels ».

3. Le projet VOILADIS (VOIsinage Lexical pour l'Analyse du DIScours), financé par le PRES de Toulouse et coordonné par Cécile Fabre, implique les laboratoires CLLE-ERSS (axe TAL) et IRIT (équipe LILaC).

Cet outil a notamment été appliqué à un corpus contenant 10 ans d'articles du journal Le Monde (menant à la constitution des *Voisins du Monde*) et à un corpus contenant l'intégralité de l'encyclopédie en ligne Wikipédia (constitution des *Voisins de Wikipédia*)[4].

À titre d'exemple, dans la base des *Voisins de Wikipédia*, le mot « protection » a pour voisins les mots « protéger », « menacer », « défense », « préserver », « destruction », etc. ; ces rapprochements ont été effectués sur la base de contextes communs tels que « couche d'ozone », « zone humide », « écosystème », « habitat », « environnement ». La méthode UPÉRY permet de capter une palette de relations lexicales beaucoup plus large que celle qui est habituellement recensée dans les dictionnaires ou visée par les outils d'acquisition sémantique.

Mais les bases distributionnelles constituent également des ressources difficiles à caractériser et à exploiter, de par leur caractère pléthorique et la grande quantité de paires peu ou pas interprétables qu'elles font émerger ; les apports de l'analyse distributionnelle automatique à la linguistique et au TAL restent ainsi largement à explorer. Dans cette perspective de caractérisation des voisins distributionnels, une piste a été amorcée par de premiers travaux (Fabre et Bourigault, 2006; Morlane-Hondère et Fabre, 2010) : celle de la recontextualisation des voisins *via* leur reprojection dans le corpus d'où ils ont été extraits, de manière à fournir des indices supplémentaires permettant de corroborer ou de préciser les relations qu'ils entretiennent.

Le projet VOILADIS peut être vu comme une extension et une systématisation de cette piste de recherche, puisque nous proposons des méthodes pour reprojeter les voisins sur leur corpus d'origine à plus grande échelle, mais aussi de manière plus contrôlée et spécifiée. La recontextualisation n'est plus uniquement considérée du point de vue du retour qu'elle offre sur la ressource, mais également du point de vue des phénomènes discursifs qu'elle permet de mettre au jour dans les textes.

1.5.2 Problématique du projet VOILADIS

1.5.2.1 Objectifs

Nous avons présenté les problématiques de recherche à la croisée desquelles se situe le projet VOILADIS. D'une part, nous souhaitons explorer l'apport de la prise en compte de la cohésion lexicale dans le cadre de différentes approches du discours pour lesquelles ce type d'indices n'est pas toujours envisagé ; nous bénéficions pour cela des annotations réalisées dans le cadre du projet ANNODIS, qui offrent un double regard sur l'organisation discursive ; mais nous avons vu que la cohésion lexicale est subtile et difficile à détecter, notamment parce qu'elle repose en grande partie sur des relations non-classiques (*cf.* section 1.2), parfois instaurée par le discours (*cf.* section 1.3). D'autre part, l'analyse distributionnelle permet de détecter une très

4. Ces deux ressources sont consultables en ligne sur le site de REDAC (REssources Développées à CLLE-ERSS) : http ://redac.univ-tlse2.fr

large gamme de relations de proximité sémantique, classiques ou non-classiques, qui se construisent au sein d'un corpus donné ; des expériences préliminaire de recontextualisation de ces relations dans des textes issus de ce même corpus ont montré l'intérêt d'une telle démarche.

L'hypothèse centrale de notre travail de thèse est que l'analyse distributionnelle peut être utilisée pour détecter la cohésion lexicale, car la richesse des relations qu'elle fait émerger peut être rapprochée de la richesse de relations impliquées dans la cohésion lexicale, et qu'elle permet une certaine prise en compte de la nature bidirectionnelle du lien entre lexique et texte (les relations lexicales peuvent être créées par le texte, mais participent également à la création de sa texture, *cf.* section 1.3.1). Partant, notre objectif est :

- d'explorer cette hypothèse en proposant une méthodologie de projection des voisins distributionnels sur des textes et en caractérisant les sorties obtenues ; nous nous appuyons pour cela sur les *Voisins de Wikipédia* ;
- d'intégrer des indices basés sur la cohésion lexicale détectée par voisinage distributionnel dans différentes approches de l'organisation textuelle, et de montrer leur apport à ces approches ; nous bénéficions pour cela des annotations effectuées sur des textes issus de Wikipédia lors de la campagne d'annotation ANNODIS, mais prévoyons également d'explorer des directions différentes.

La poursuite de ce double-objectif nous situe sur un terrain de recherche très vaste et qui ne peut être entièrement couvert ; dans la section qui suit nous précisons quelques pistes que nous ne suivront pas.

1.5.2.2 Délimitation de la problématique

Certains aspects ne seront pas abordés dans cette thèse, dans un souci de circonscription de la problématique. Nous en citons ici deux principaux.

Combiner plusieurs ressources pour mieux capter la cohésion lexicale La combinaison de différentes sources d'informations (dictionnaires, collocations, etc.) pourrait sans doute aboutir à une meilleure détection de la cohésion lexicale. Toutefois, nous nous intéressons spécifiquement dans cette thèse à l'exploitation d'une ressource distributionnelle, les *Voisins de Wikipédia* et n'envisagerons pas de la compléter avec des relations lexicales issues d'autres ressources. Notre but n'est pas de capter les phénomènes de cohésion lexicale le plus exhaustivement possible, mais d'explorer les possibilités offertes par l'analyse distributionnelle. Si nous faisons appel à une autre ressource lexicale dans cette thèse (plus particulièrement à un dictionnaire de synonymes), c'est dans une perspective de comparaison.

Projeter les voisins sur des textes n'appartenant pas au corpus ayant servi à les construire L'encyclopédie Wikipédia, sur laquelle est basée la construction des *Voisins de Wikipédia* et dont certains articles ont été annotés dans le cadre du projet ANNODIS, suffit largement à nous pourvoir en données textuelles pour les

expériences que nous menons dans cette thèse. Nous n'envisagerons pas d'appliquer nos méthodes à des textes provenant d'autres sources, qui demanderaient de se poser la question de la portabilité de la ressource, alors même que l'une de nos motivations pour faire appel à cette ressource est qu'elle permet d'extraire des relations opérant dans un corpus donné. De plus, dans la mesure où l'analyse distributionnelle et la projection des voisins en texte sont des procédures entièrement automatiques, il est tout à fait possible d'imaginer que la même chaîne de traitement (SYNTEX-UPÉRY-VOILADIS) pourrait être appliquée à un autre corpus.

1.5.2.3 Plan de réalisation

Étant donné les objectifs que nous nous sommes fixés, le plan de réalisation consiste en deux principales étapes.

1. Mise en place d'un environnement permettant de projeter les voisins distributionnels en texte : développement d'une procédure automatisée de projection, élaboration des stratégies de filtrage des liens projetés et caractérisation des sorties obtenues afin de faciliter leur exploitation.
2. Définition de modalités d'exploitation des liens de voisinage projetés au sein d'approches variées du discours, en s'appuyant notamment sur les données d'ANNODIS ; caractérisation de l'apport de la méthode à l'étude de l'organisation textuelle.

La première étape s'intéresse au lien entre voisinage distributionnel et cohésion lexicale ; elle fait l'objet de la IIe partie de cette thèse. La seconde étape se concentre sur le lien entre cohésion lexicale (telle que détectée à l'issue de l'étape 1) et organisation textuelle ; elle est abordée dans la III partie de cette thèse.

Deuxième partie

De l'analyse distributionnelle à la détection de la cohésion lexicale : parcours

Chapitre 2

Les Voisins de Wikipédia : Construction, description, exploitation

Sommaire

Ce chapitre est consacré à la ressource utilisée lors du projet VOILADIS, les *Voisins de Wikipédia*. Nous présentons d'abord l'hypothèse qui sous-tend l'analyse distributionnelle, ainsi que ses grands principes (2.1) ; nous détaillons ensuite la mise en oeuvre informatique de l'analyse distributionnelle par la chaîne de traitement SYNTEX-UPÉRY(2.2), qui aboutit à la construction des *Voisins de Wikipédia* dont nous donnons des éléments de description (2.3) ; nous évoquons enfin la question de l'exploitation d'une telle ressource (2.3.3).

2.1 L'analyse distributionnelle : des origines à la mise en oeuvre informatique

2.1.1 Les origines de l'analyse distributionnelle

> Eadem sunt quorum unum potest substitui alteri salva veritate. [a]
>
> *a.* Deux termes dont l'un peut toujours être substitué à l'autre sans changer la valeur de vérité d'une proposition sont identiques.
>
> Leibniz (1704)

La linguistique structurale s'est développée à partir du « Cours de linguistique générale » de Ferdinand de Saussure (Saussure, 1916), qui est considéré comme le père de la linguistique moderne. Dans le point de vue structural, la langue est constituée d'ensembles d'objets arbitraires, organisés en systèmes définis par les relations qu'entretiennent ces objets entre eux. Le distributionalisme est issu du courant structuraliste américain ; il a été esquissé par Leonard Bloomfield dans son ouvrage fondateur « Language » (Bloomfield, 1933), et formalisé par son élève Zellig Sabbettai Harris (Harris, 1951).

Le distributionalisme est né sous l'influence des théories psychologiques behavioristes, selon lesquelles le comportement humain serait totalement explicable, et on pourrait en étudier la mécanique. Il se caractérise donc par son empirisme, avec notamment le recours aux corpus pour observer les formes linguistiques, et le rejet des théories *mentalistes* sur la notion de sens, qui selon Bloomfield ne peut pas être définie clairement [1] :

> Adherents of mentalistic psychology believe that they can avoid the difficulty of defining meanings, because they believe that, prior to the utterance of a linguistic form, there occurs within the speaker a non-physical process, a *thought*, *concept*, *image*, *feeling*, *act* of *will*, or the

1. « We cannot with certainty define meanings », « the linguist cannot define meanings » (Bloomfield, 1933, p. 145)

> like, and that the hearer, likewise, upon receiving the sound-waves, goes through an equivalent or correlated mental process. The mentalist, therefore, can define the meaning of a linguistic form as the characteristic mental event which occurs in every speaker and hearer in connection with the utterance or hearing of the linguistic form. The speaker who utters the word apple has had a mental image of an *apple*, and this word evokes a similar image in a hearer's mind. (Bloomfield, 1933, p. 142)

Au mentalisme, Bloomfield (1933, pp. xv-xvi) oppose le *mécanisme*. Selon l'approche mécaniste, la parole ne peut pas s'expliquer comme un effet de la pensée, et on doit pouvoir rendre compte de sa structure hiérarchisée sans émettre aucun postulat concernant les intentions des locuteurs et leurs états mentaux.

> The mechanist does not accept this solution. He believes that *mental images*, *feelings*, and the like are merely popular terms for various bodily movements. (Bloomfield, 1933, p. 142)

Plutôt que de définir le sens comme un « état mental », Bloomfield fait appel à la notion de *classes*, c'est-à-dire d'ensembles d'unités pouvant se substituer les unes aux autres dans la chaîne parlée ; chaque classe est dotée d'un sens (*classmeaning*) : deux unités qui apparaissent dans un même environnement présenteront des similitudes du point de vue de leur sens.

> If the difference male : female is defined for the linguist, he can assure us that this is the difference between he : she, lion : lioness, gander : goose, ram : ewe. The linguist has this assurance in very many cases, where a language, by some recognizable phonetic or grammatical feature, groups a number of its forms into form-classes : in any one form-class, every form contains an element, the classmeaning, which is the same for all forms of this form-class. Thus, all English substantives belong to a form-class, and each English substantive, accordingly, has a meaning, which, once it is defined for us (say, as 'object'), we can attribute to every substantive form in the language. English substantives, further, are subdivided into the two classes of singular and plural ; granted a definition of the meanings of these two classes, we attribute one of these meanings to every substantive. (Bloomfield, 1933, p. 146)

La notion mentaliste de sens est ainsi remplacée par la notion de distribution (la somme des environnements dans lequel apparaît une unité linguistique).

Issue de ces idées, la méthode distributionnelle développée par Harris (1951, 1954, 1962, 1968), consiste à « définir avec rigueur une méthode formelle de segmentation de la chaîne parlée en unités distinctives, définies par les seules relations qu'elles entretiennent dans cette chaîne, c'est-à-dire par leur environnement » (Dubois et Dubois-Charlier, 1970, p. 3). Selon l'hypothèse distributionnelle, le sens des unités lexicales dépend (au moins en partie) de leurs propriétés distributionnelles, c'est-à-dire des contextes linguistiques dans lesquels elles apparaissent ; le degré de similarité sémantique entre deux unités lexicales dépend alors de la similarité entre

leurs distributions. L'hypothèse distributionnelle est formulée par Harris (1954) en ces mots :

> If we consider words or morphemes A and B to be more different in meaning than A and C, then we will often find that the distributions of A and B are more different than the distributions of A and C. In other words, difference in meaning correlates with difference of distribution.

Cette hypothèse a d'abord été mise en œuvre de façon manuelle par Harris, dans le cadre de discours spécialisés. L'analyse distributionnelle a donné lieu à des tentatives de mise en œuvre informatique à partir des années 1970 (Hirschman et Grishman, 1975) ; toutefois, les premières implémentations restent limitées par la taille réduite des corpus utilisés et la difficulté à disposer de données annotées syntaxiquement. Hindle (1990) se base sur un corpus d'articles de presse de 6 millions de mots annoté syntaxiquement par l'analyseur syntaxique robuste Fidditch ; l'observation des données obtenues lui fait conclure à la plausibilité de l'hypothèse distributionnelle (Hindle, 1990, p. 274). On peut également citer les travaux fondateurs de Lund et Burgess (1996) et de Landauer et Dumais (1997). Dans la section qui suit, nous présentons dans leurs grandes lignes les principes de construction automatique d'une ressource distributionnelle.

2.1.2 Mise en oeuvre informatique de l'hypothèse distributionnelle

Nous décrivons dans cette section les grands principes de l'analyse distributionnelle automatique d'un corpus, qui aboutit à la construction d'une base distributionnelle telle que les *Voisins de Wikipédia* (dont la construction est spécifiquement décrite dans la section 2.2). Dans cette thèse, on parlera de « base distributionnelle » ou de « ressource distributionnelle ». De nombreux équivalents existent dans la littérature, notamment « *Word-Space (Model)* » (Schütze, 1992, 1993; Sahlgren, 2006; Baroni et Lenci, 2008), « *Semantic Space* » (Lund et Burgess, 1996; Baroni *et al.*, 2009) ou « *Distributional Semantic Model* » (Lenci, 2008; Evert et Lenci, 2009).

2.1.2.1 Construction d'une ressource distributionnelle : vision d'ensemble

Le principe de l'analyse distributionnelle automatique est le suivant : dans un corpus donné, on compte combien de fois certains *mots-cibles*[2] apparaissent en relation avec certains *contextes*, ce qui permet de construire pour chaque mot-cible un vecteur représentant sa distribution. Les vecteurs similaires représentent ainsi des mots qui ont une distribution similaire. C'est l'hypothèse distributionnelle qui fait interpréter cette similarité distributionnelle comme le reflet d'une similarité sémantique. La similarité distributionnelle correspond à ce que Grefenstette (1994a)

2. *target word* (Peirsman *et al.*, 2007; Bullinaria, 2008; Baroni *et al.*, 2009, ...)

nomme une affinité de deuxième ordre entre mots. Deux mots sont en affinité de premier ordre s'ils tendent à apparaître dans un même contexte (les collocations sont des exemples d'affinités de premier ordre) ; deux mots qui tendent à avoir les mêmes affinités de premier ordre sont dits en affinité de deuxième ordre.

Si les grands principes de l'analyse distributionnelle automatique restent les mêmes, on observe des différences d'une mise en oeuvre à l'autre, dues à la variété de réponses que l'on peut apporter à différentes questions telles que :

- Comment sélectionne-t-on les mots cibles ?
- De quelle nature sont les « contextes » qui permettent de rapprocher ces mots-cibles ?
- Quelle mesure utilise-t-on pour calculer la similarité entre vecteurs de distribution ?

Ces questions essentielles constituent différents paramètres à fixer pour effectuer l'analyse distributionnelle d'un corpus. Nous listons ces paramètres dans les sections qui suivent, en les regroupant en deux grands types : paramètres linguistiques (2.1.2.2) et paramètres mathématiques (2.1.2.3).

2.1.2.2 Paramètres linguistiques

Les paramètres linguistiques interviennent dans les premières étapes de la construction d'une ressource distributionnelle. Ils concernent notamment :

- le choix du corpus : constitue-t-on un corpus en langue de spécialité, en langue générale ? de quelle taille ? etc.
- le pré-traitement du corpus et l'ajout d'annotations linguistiques : comment s'effectue la segmentation des textes ? (mots graphiques ou prise en compte d'unité polylexicales ?) Le corpus est-il soumis à un prétraitement ? (lemmatisation, analyse morpho-syntaxique / syntaxique, etc.)
- la sélection des mots-cibles : sont-ils sélectionnés sur leurs catégories morpho-syntaxiques, sur leur fréquence, etc.
- la sélection des contextes : document, phrase, fenêtre d'un certain nombre de mots, mots en relation syntaxique avec le mot-cible ?

2.1.2.3 Paramètres mathématiques

Au terme des traitements linguistiques, il est possible d'extraire la distribution de tous les mots-cibles du corpus. Il s'agit alors de comparer ces distributions pour rapprocher les mots-cibles similaires. Différents paramètres mathématiques interviennent dans ce calcul :

- la pondération des contextes : on peut accorder plus de poids à certains contextes sur la base par exemple d'une mesure d'association comme l'information mutuelle ;
- la mesure de similarité employée ;

– l'utilisation de méthodes de réduction dimensionnelle sur les vecteurs de distribution.

Si l'on sait que tous ces paramètres (linguistiques et mathématiques) ont beaucoup d'influence sur la ressource obtenue, la compréhension de cette influence est encore lacunaire. De nombreux travaux récents s'attachent à explorer systématiquement l'influence de certains paramètres (Ferret, 2010; Otero, 2009; Bullinaria et Levy, 2006; Bullinaria, 2008; Weeds, 2003).

2.1.3 Utilisation et évaluation des ressources distributionnelles

L'analyse distributionnelle a été pensée pour être appliquée sur des textes en langue spécialisée, afin d'extraire des classes lexicales opérant dans le cadre de *sous-langages* (par exemple langages de disciplines scientifiques ou techniques (Harris, 1988, 1990, 1991). Les sous-langages se caractérisent par des restrictions de sélection bien plus nettes que celles que l'on observe dans la langue générale :

> Le caractère distinctif d'un sous-langage, c'est que pour certains sous-ensembles des phrases du langage, les restrictions de sélection, pour lesquelles on ne peut pas fournir de règles pour le langage dans son ensemble, intègrent la grammaire. Dans un sous-langage, les classes lexicales ont des frontières relativement tranchées qui reflètent la division des objets du monde en catégories qui sont clairement différenciées dans le domaine. (Sager, 1986)

Parmi les travaux fondateurs de l'analyse distributionnelle, ont trouve ceux menés par Harris *et al.* (1989); Ryckman (1990) sur le discours pharmaceutique et biologique, par Sager *et al.* (1987) sur le langage médical, ou encore par Grishman et Kittredge (1986) en explorant différents domaines. En français, on trouve notamment les travaux de Nazarenko *et al.* (1997) et Bouaud *et al.* (2000) portant sur des textes issus du domaine médical.

L'analyse distributionnelle connait actuellement un regain d'intérêt, avec l'émergence d'approches se détournant des domaines spécialisés pour se consacrer à l'analyse de grands corpus relevant davantage de la langue générale, avec notamment l'exploitation de textes journalistiques (Bourigault et Galy, 2005; Dias *et al.*, 2010), ou de textes tout-venant issus du Web et rassemblés en vue de construire des corpus de très grande taille (Terra et Clarke, 2003; Turney, 2008). Ces approches produisent des résultats beaucoup plus difficiles à appréhender et à évaluer ; mais elles permettent d'envisager d'exploiter les bases distributionnelles construites dans une large gamme de tâches sémantiques (Baroni *et al.*, 2009). Parmi les tâches pour lesquelles des bases distributionnelles ont été exploitées, on peut notamment citer la désambiguïsation lexicale (Schütze, 1998), l'expansion de requêtes en recherche d'information (Grefenstette, 1994b), la désambiguïsation de l'attachement de syntagmes

prépositionnels en analyse syntaxique (Pantel et Lin, 2000).

L'évaluation d'une ressource distributionnelle repose sur la question de savoir quelles sont les paires de mots qui présentent une relation de sens pertinente, et quelles sont celles qui sont catégorisées comme étant du *bruit*. Trois méthodes principales peuvent être utilisées pour évaluer une ressource lexicale.

- La comparaison à une ressource de référence ou *gold standard*, en faisant appel à une ressource construite manuellement par des experts comme WordNet ou EuroWordNet (Vossen, 1998) : si la comparaison à un *gold standard* peut constituer un point de départ intéressant pour évaluer et caractériser une ressource distributionnelle, elle n'offre toutefois qu'un point de vue très limité (Poibeau et Messiant, 2008). Le principe de l'analyse distributionnelle est de faire émerger des relations sémantiques sans *a priori* sur leur nature, susceptibles de faire évoluer les ressources existantes.
- La confrontation à des jugements humains : soit en utilisant des données générées lors de tâches d'association (Lindsey *et al.*, 2007; Wandmacher *et al.*, 2008), soit en demandant à des locuteurs d'exercer leur intuition sur un ensemble de paires extraites de la ressource (Evert et Krenn, 2001, 2005).
- L'évaluation « par la tâche » : il s'agit d'évaluer la ressource à l'aune des performances d'un système dans lequel elle a été implémentée. Des évaluations de ressources lexicales ont notamment été conduites par le biais de tâches telles que la désambiguïsation lexicale (Weeds et Weir, 2005) ou la détection de malapropismes (Budanitsky et Hirst, 2006). Cette méthode d'évaluation est dite extrinsèque ou indirecte (Curran, 2003), par opposition aux deux premières approches qui évaluent la ressource en elle-même et pour elle-même.

2.2 Construction des *Voisins de Wikipédia*

> « Qu'est-ce que signifie "apprivoiser" ?
> – C'est une chose trop oubliée, dit le renard.
> Ça signifie "créer des liens..." »
>
> Antoine de Saint-Exupéry, Le Petit Prince

Les *Voisins de Wikipédia* sont calculés à partir du corpus WikipédiaFR2007 (2.2.1) par la chaîne de traitement SYNTEX-UPÉRY : SYNTEX (2.2.2) effectue l'analyse syntaxique du corpus et UPÉRY (2.2.3) se base sur cette analyse syntaxique pour mener l'analyse distributionnelle. La construction du corpus WikipédiaFR2007 et l'application de la chaîne SYNTEX-UPÉRY sur ce corpus sont dûs à F. Sajous. Les *Voisins de Wikipédia* sont consultables *via* une interface en ligne [3].

3. http ://redac.univ-tlse2.fr/voisinsdewikipedia

2.2.1 Le corpus : WikipédiaFR2007

Le corpus utilisé pour construire les *Voisins de Wikipédia* est une version de l'encyclopédie en ligne Wikipédia. Plus précisément, il s'agit de l'intégralité de l'encyclopédie Wikipédia francophone dans sa version d'avril 2007. Ce corpus présente l'avantage d'être relativement bien caractérisé et homogène en genre (mais couvrant des domaines très variés), tout en étant de taille conséquente. Il s'agit donc d'une alternative intéressante à la stratégie consistant à constituer de très gros corpus avec des textes tout-venant.

Le tableau 2.1 résume les caractéristiques essentielles de ce corpus (désormais WikipédiaFR2007). Nous y précisons les nombres de noms (propres ou communs), verbes et adjectifs (que nous regroupons sous « mots pleins ») car il s'agit des mots concernés par le calcul des voisins.

Nombre de textes (articles)	471075
Nombre de phrases	20745434
Moyenne phrases / article	44.04
Nombre de mots	206 852 826
Moyenne mots / article	439.11
Nombre de noms communs	51 999 482
Nombre de noms propres	27 319 303
Nombre de verbes	12 037 757
Nombre d'adjectifs	12 905 959
Tot. « mots pleins »	104 262 501
« mots pleins » / mots	50.40%

TABLEAU 2.1 – Caractéristiques du corpus WikipédiaFR2007

Avec plus de 200 millions de mots, le corpus utilisé est de taille imposante mais encore loin de la taille de certains corpus utilisés dans le cadre de l'analyse distributionnelle, notamment pour l'anglais (Terra et Clarke, 2003; Turney, 2008). On peut également remarquer que le nombre de mots par article est assez faible. On observe en effet une forte variété dans la longueur des articles, avec la présence d'articles très longs mais d'autre part une grande quantité d'articles quasi-vides.

2.2.2 Analyse syntaxique par SYNTEX

SYNTEX a été développé au début des années 2000 (Bourigault et Fabre, 2000). Il est décrit par Bourigault (2007) comme un analyseur syntaxique « opérationnel », c'est-à-dire robuste et utilisable par une large gamme d'applications sur des corpus de genre variés. SYNTEX prend en entrée des textes étiquetés par le TreeTagger et a un fonctionnement en cascade, c'est-à-dire qu'il traite chaque phrase en plusieurs passes successives, un module étant consacré à chaque relation syntaxique. Les relations sont résolues dans l'ordre suivant : coordination, objet, sujet, adjectif épithète,

prépositions. Il effectue une analyse ascendante en dépendance. Il identifie progressivement les relations syntaxiques jusqu'à aboutir à une représentation syntaxique globale de la phrase ; le mode de représentation en dépendance part du principe que la présence de chaque mot est légitimée par la présence d'un autre mot (son gouverneur). Pour un exemple d'analyse syntaxique détaillée pas à pas, on peut se reporter à Bourigault (2007, pp. 72-74).

Les sorties de SYNTEX se présentent sous la forme d'un fichier (dit « fichier anasynt ») qui pour chaque phrase fournit trois lignes d'informations :

- un identifiant unique de la phrase (identifiant unique du texte suivi du numéro de la phrase dans le texte) ;
- le texte de la phrase segmenté en mots ;
- l'analyse syntaxique effectuée : une étiquette par mot de la forme `catégorie|lemme|forme|numéro|régi par|recteur de`

La figure 2.1 présente un exemple de sortie pour la phrase « Le Kilimandjaro éveille l'intérêt des explorateurs. »[4]. Dans cette analyse, on peut par exemple voir

```
<SEQ id=1137ceb1351e5c66e897e6392bd4956b_13; analyse=1;>
<TXT>Le Kilimandjaro éveille l' intérêt des explorateurs.
<ETIQ>
DetMS|le|Le|1|DET;2|
NomPr|Kilimandjaro|Kilimandjaro|2|SUJ;3|DET;1
VCONJP|éveiller|éveille|3||SUJ;2,OBJ;5
Det??|le|l'|4|DET;5|
Nom?S|intérêt|intérêt|5|OBJ;3|DET;4,PREP;6
Prep|de|des|6|PREP;5|NOMPREP;7
Nom?P|explorateur|explorateurs|7|NOMPREP;6|
Typo|.|.|8||
```

FIGURE 2.1 – Extrait d'une sortie de SYNTEX

que le nom « Kilimandjaro » est gouverneur du déterminant « le » *via* la relation `DET`, et qu'il est dépendant du verbe « éveiller » *via* la relation `SUJ`.

2.2.3 Analyse distributionnelle par UPÉRY

UPÉRY (Bourigault, 2002) est un module d'analyse distributionnelle développé spécifiquement pour traiter des corpus étiquetés par SYNTEX. Il prend donc en entrée des corpus entièrement analysés syntaxiquement. Les mots-cibles (dont on veut extraire la distribution) sont l'ensemble des noms (ou syntagmes nominaux), verbes et adjectifs du corpus dépassant un certain seuil (paramétrable) de fréquence ;

4. Adaptation d'une phrase extraite de l'article « Kilimandjaro » dans WikipédiaFR2007.

sont également exclus certains adjectifs et verbes trop fréquents[5]. Les contextes considérés sont des contextes syntaxiques. Un mot peut constituer un contexte pour un autre mot s'il est rattaché à lui par une relation de dépendance syntaxique (objet, sujet, modifieur) ; ces relations de dépendances peuvent être labellisée par une préposition (« à », « de », etc.) ou un groupe prépositionnel (« au bord de », « dans le cadre de », etc.).

Nous détaillons dans ce qui suit les étapes de l'analyse distributionnelle opérée par UPÉRY à partir de l'exemple déjà utilisé : « Le Kilimandjaro éveille l'intérêt des explorateurs ».

1. Dans un premier temps, l'ensemble des triplets <`gouverneur`, `relation`, `dépendant`> sont extraits. Dans notre exemple, il s'agira des triplets :
   <`éveiller`, `obj`, `intérêt`>
   <`éveiller`, `obj`, `intérêt de explorateur`>
   <`éveiller`, `suj`, `Kilimandjaro`>
   <`intérêt`, `de`, `explorateur`>

2. Ces triplets sont ensuite ramenés à des couples <`prédicat`, `argument`> en accolant la relation au gouverneur. Ainsi la différence entre gouverneurs et dépendants est conservée, puisque les prédicats (définis comme un gouverneur + une relation) ne seront rapprochés qu'avec d'autres prédicats et les arguments (dépendants syntaxiques) seront rapprochés avec d'autres arguments. Dans notre exemple, on obtient les couples <`prédicat`, `argument`> suivants :
   <`éveiller_obj`, `intérêt`>
   <`éveiller_obj`, `intérêt de explorateur`> <`éveiller_obj`, `Kilimandjaro`>
   <`intérêt_de`, `explorateur`>
   On peut constater qu'un même mot (ici « intérêt » peut être prédicat dans certains couples et argument dans d'autres).

3. Pour chaque couple, un score d'information mutuelle est calculé. Ce score ne sert pas à pondérer les contextes, mais à exclure certains couples dont l'information mutuelle serait trop faible (seuil paramétrable).

4. Enfin, la similarité des distributions de l'ensemble des prédicats et de l'ensemble des arguments extraits est calculée en utilisant le score de Lin (1998). Soient deux mots w_1 et w_2, pour lesquels on a mesuré un ensemble de caractéristiques $F(w)$. Lin envisage cet ensemble de caractéristiques comme une description du mot, cette description portant une information $I(F(w))$. Le score de similarité de Lin $sim(w_1, w_2)$ est calculé comme le rapport de l'information commune aux deux mots sur la moyenne de l'information portée par la description de chacun

5. les adjectifs « même », « autre », « seul », « tel » et les verbes « être », « avoir » et « permettre ».

des mots :

$$sim(w_1, w_2) = \frac{2 \times I\left(F(w_1) \cap F(w_2)\right)}{I\left(F(w_1)\right) + I\left(F(w_2)\right)} \tag{2.1}$$

pour chaque prédicat, respectivement argument d'un corpus est le rapport du nombre de prédicats/arguments avec lesquels ils se combinent dans le corpus, sur la totalité des prédicats/arguments avec lesquels ils pourraient se combiner. Dans le cas du calcul de similarité distributionnelle, l'information commune à deux unités est le nombre de contextes qu'elles partagent. Ainsi, le prédicat `éveiller_obj` a ainsi une forte similarité distributionnelle avec le prédicat `exciter_obj`, grâce à des contextes partagés tels que `curiosité`, `convoitise`, `appétit` ou encore `imagination`. L'argument `intérêt` est quant à lui rapproché de l'argument `importance` grâce à des contextes partagés tels que `juger_sans`, `attacher_obj` ou `se mesurer_suj`.

Nous avons mentionné en 2.1.2 la variété de paramètres qui doivent être spécifiés lors de la construction d'une ressource distributionnelle. Ces paramètres sont autant de leviers pouvant être manipulés pour modifier la ressource résultante ; cette ressource est donc fondamentalement modulable, dynamique. Dans le cas des *Voisins de Wikipédia*, les paramètres que nous avons appelés « linguistiques » correspondent à des choix motivés : les partis-pris consistant à utiliser un corpus annoté syntaxiquement et à se baser sur cette annotation syntaxique pour calculer les distributions d'unités linguistiques bien définies (mots ou syntagmes) sont des des choix théoriques forts visant à construire une base distributionnelle sur des critères réellement linguistiques. Par contre, les paramètres mathématiques (les différents seuils paramétrables, la mesure de similarité utilisée, etc.) ont été fixés empiriquement de manière relativement arbitraire et sont donc davantage questionnables. De plus, il faut ajouter à ces paramètres mathématiques inhérents à la construction de la ressource un ensemble de filtres pouvant être appliqués *a posteriori* (que nous abordons dans le chapitre 4) : par exemple, placer un seuil sur le nombre de voisins admis pour chaque item, à la place ou en complément d'un seuil sur le score de Lin.

2.3 Description des *Voisins de Wikipédia*

La description linguistique d'une base distributionnelle reste un défi de par son volume et le manque d'étalons auxquels la comparer. Des méthodes visant à mieux spécifier la nature des relations mises au jour sont notamment développées dans Morlane-Hondère et Fabre (2012). Dans cette section, nous donnons quelques éléments de description de la base distributionnelle obtenue au terme des traitements détaillés en 2.2. Après l'avoir caractérisée brièvement 2.3.1, nous donnons un aperçu de la richesse des relations lexicales qu'elle exhibe (2.3.2) ; enfin, nous posons la question de son exploitation et le problème du bruit qu'elle contient (2.3.3).

2.3.1 Caractérisation des *Voisins de Wikipédia*

Les *Voisins de Wikipédia* recensent 1383774 couples de voisins distributionnels. La répartition des scores de Lin est illustrée par les figures 2.2 et 2.3 ; 97.2% des paires sont dotées d'un score compris entre 0.1 et 0.29. Ainsi, les voisins sont très tassés dans une petite fourchette de scores très faibles, ce qui incite à la prudence lors de la mise en place éventuelle d'un seuil sur le Lin.

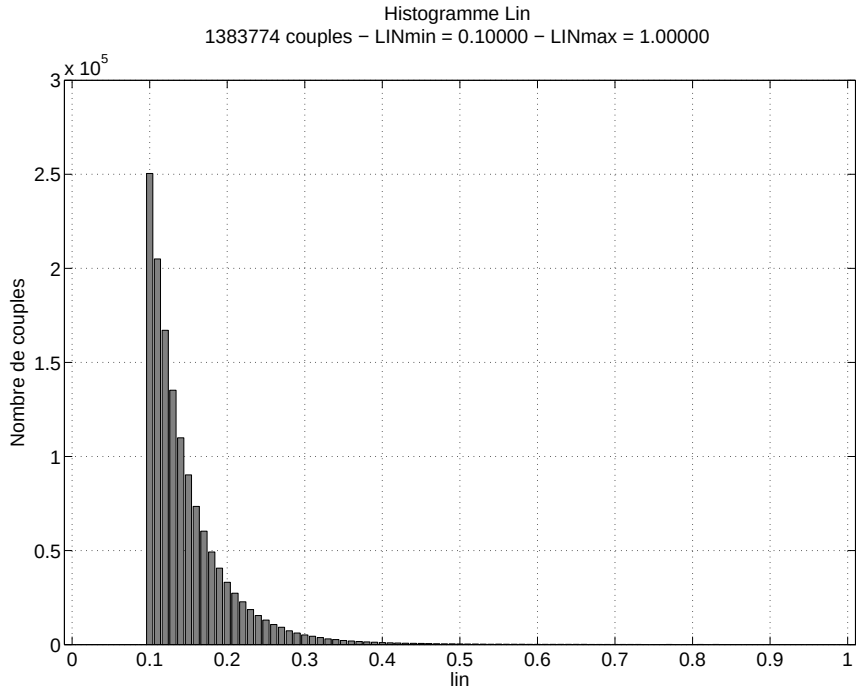


FIGURE 2.2 – Histogramme du score de Lin des *Voisins de Wikipédia*

La figure 2.4 présente la répartition des voisins selon leurs catégories. Les couples nom/nom sont de loin les plus nombreux (51.7%), suivis par les couples les couples nom/verbe (17.4%) et verbe/verbe (15.0%). En dessous des 10.0%, on trouve les couples adjectif/adjectif (7.4%). Les autres de types de rapprochement (impliquant notamment des syntagmes et des noms propres) ne dépassent pas les 2.5%.

Les couples verbe/verbe et nom/verbe émergent du rapprochement de prédicats ; les couples adjectif/adjectif sont dûs au rapprochement d'arguments. Les couples noms/noms peuvent être prédicatifs (lorsqu'ils partagent des arguments comme par exemple des adjectifs) ou argumentaux (lorsqu'ils sont rapprochés sur la base de prédicats – par exemple de verbes – communs). Ainsi, les prédicats sont légèrement plus représentés que les arguments (58.8% / 41.2%, *cf.* figure 2.5 dans la base des *Voisins de Wikipédia*.

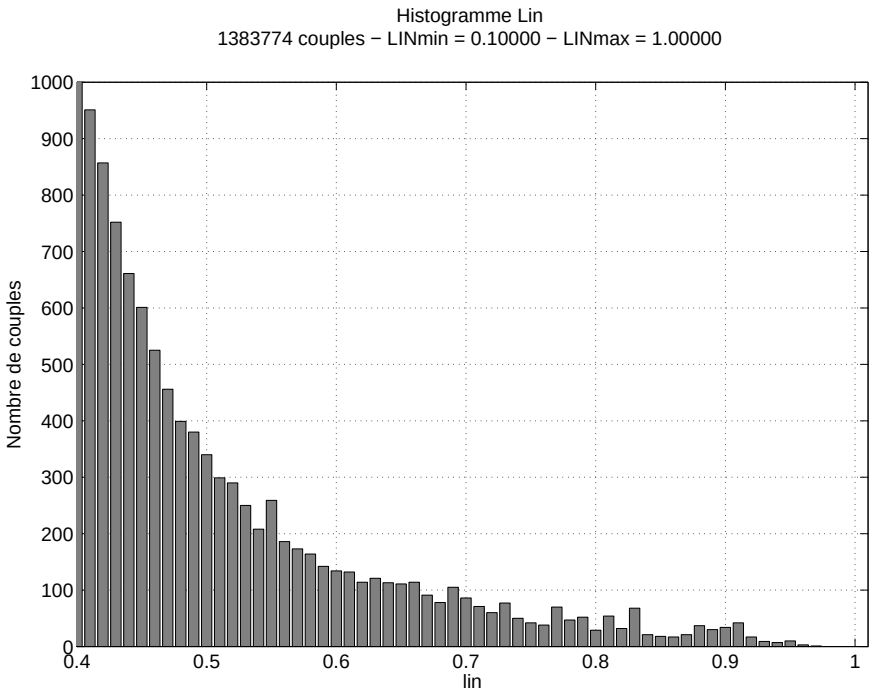


FIGURE 2.3 – Histogramme du score de Lin des *Voisins de Wikipédia*, détail pour les scores entre 0.4 et 1.0

2.3.2 La richesse des liens extraits

Dans cette section nous donnons à voir la large palette de relations lexicales présentes dans les *Voisins de Wikipédia*.

Les relations lexicales classiques (Lyons, 1977; Cruse, 1986) sont la synonymie, l'antonymie, l'hypo/hypéronymie et la méro/holonymie. Ces relations sont définies comme paradigmatiques, et peuvent toutes s'observer parmi les *Voisins de Wikipédia*. Nous en donnons quelques exemples dans le tableau 2.2.

Néanmoins les relations exhibées par les voisins vont bien au delà de ces relations classiques avec la présence de toute une gamme de relations non-classiques (Morris et Hirst, 2004, *cf.* section 1.2.3). Le tableau 2.3 montre des exemples de telles relations, dont on peut voir qu'elles recouvrent les différentes catégories considérées par Morris et Hirst (2004) :

- relations entre mots au sein de catégories non-classiques : par exemple, les relations `joueur`/`match` et `ballon`/`maillot` peuvent s'interpréter par référence à l'activité *football* ;
- relations thématiques (*case relations*) : `professeur` est un agent typique de l'action `enseigner` ;
- autres *related terms*.

La mise au jour de relations intercatégorielles par le biais des rapprochements de

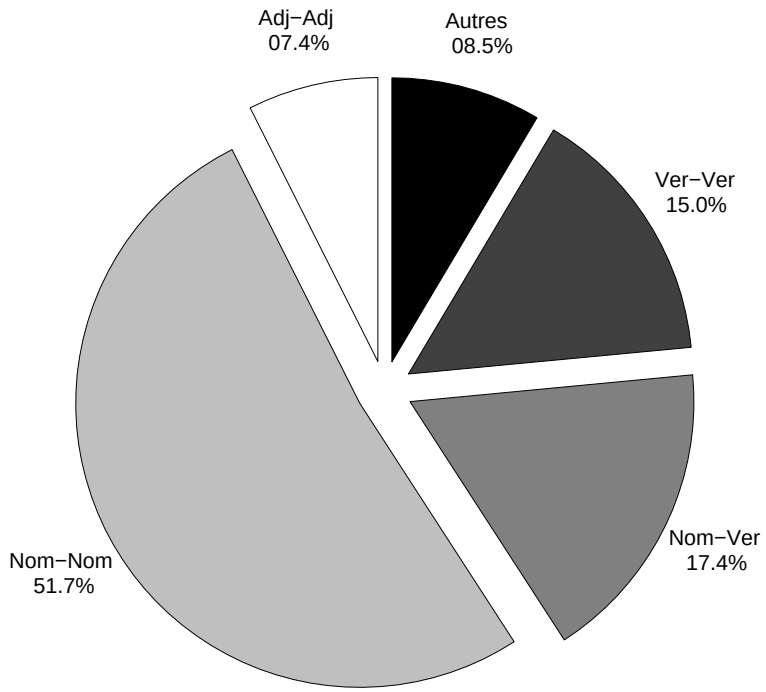


FIGURE 2.4 – Répartition des catégories des *Voisins de Wikipédia*

FIGURE 2.5 – Répartition des *Voisins de Wikipédia* entre arguments et prédicats.

Relation	Couples	Contextes communs
Synonymie	thèse_de/doctorat_de immortel/éternel	université de Paris, médecine gloire_mod, âme_mod, amour_mod
Antonymie	perdre_obj/gagner_obj vrai/faux	temps précieux, pari, match jumeau_mod, départ_mod
Hyponymie	pomme/fruit français/langue	salade_de, croquer_obj, cueillir_obj traduire_vers, livre_en, locuteur_de
Co-hyponymie	linguiste/sociologue lundi_mod/mardi_mod	étudier_suj, développer_suj, travail_de soir, gras, noir
Méronymie	voiture_mod/moteur_mod doigt/main	performant, diesel, hybride appuyer_avec, se brûler_obj
Co-méronymie	pouce/index fleur_mod/feuille_mod	tenir_entre, phalange_de, bout_de sessile, sécher, couper

TABLEAU 2.2 – Voisinage et relations classiques

Couples	Contextes communs
professeur_de/enseigner_obj	belles-lettres, botanique, littérature française
joueur_de/match_de	série éliminatoire, foot, ligue 1
ballon/maillot	porteur_de, jouer_avec, s'emparer_de
protéger_obj/protection_de	patrimoine, population civile, écosystème
pratiquer_obj/adepte_de	marche, art martial, bouddhisme
protester_contre/attidude_de	gouvernement, population, parti

TABLEAU 2.3 – Voisinage et relations non-classiques

prédicats est un intérêt de l'analyse distributionnelle. Ces relations restent peu étudiées et recensées au delà du critère morphologique. Or, comme le montrent Fabre et Bourigault (2006), les relations Nom/Verbe émergeant de l'analyse distributionnelle ne se limitent pas à des liens morphologiques (*cf.* les exemples protéger/protection, mais également pratiquer/adepte).

Le principe même de l'analyse distributionnelle est d'extraire des couples de mots entretenant une relation d'ordre paradigmatique, c'est-à-dire pouvant se substituer dans certains contextes. On observe ainsi dans les voisins, toutes sortes de relations paradigmatiques classiques et non classiques. Néanmoins, un grand nombre des relations mises au jour par l'analyse distributionnelle ressemblent à ce qu'on pourrait obtenir en extrayant des collocations (professeur/enseigner, protester/attitude). Ceci dit, il faut souligner que même les relations classiques, définies comme paradigmatiques, présentent une tendance à apparaître en cooccurrence. Nous avons calculé la proportion de couples parmi les *Voisins de Wikipédia* qui cooccurrent au moins une fois au sein d'un même paragraphe dans le corpus ayant servi à générer la base. Cette proportion est de 69.9%. La tendance à présenter un lien syntagmatique est donc une propriété importante des voisins distributionnels.

Enfin, nous notons que l'analyse distributionnelle extrait des relations de voisi-

nage opérant dans un corpus donné, et donc dépendantes de ce corpus. Dans le cas d'un corpus en langue de spécialité, c'est ce qui est recherché ; dans le cas d'un corpus de langue générale comme c'est le cas avec les *Voisins de Wikipédia*, on espère que la dépendance des liens extraits au corpus d'origine est moindre. Toutefois, on peut s'attendre à trouver parmi les *Voisins de Wikipédia* des relations de sens n'opérant que dans un contexte discursif particulier. Cela pose la question de la portabilité de la ressource à d'autres corpus, mais également la question de son applicabilité à l'ensemble du corpus (multi-domaines) à partir duquel elle a été construite. Nous abordons le problème du bruit contenu par la ressource dans la section suivante.

2.3.3 L'exploitation de la ressource et la question du bruit

L'approche distributionnelle est ascendante : elle fait émerger du corpus analysé des relations sémantiques sans faire d'*a priori* sur les informations à extraire (par opposition à une approche descendante qui partirait de connaissances sur le fonctionnement de la langue et les projetterait en corpus). C'est donc une caractéristique essentielle des ressources distributionnelles que de contenir des relations non étiquetées par des typologies classiques, induites par cette approche ascendante. Il est dès lors délicat d'aborder la question du bruit dans ce type de ressource, tout en ne souhaitant pas se priver de la richesse des relations sémantiques mises au jour.

Nous faisons ici l'hypothèse qu'il existe des paires de voisins qui entretiennent effectivement une relation sémantique, qu'elle soit classique ou non classique, et des paires de voisins qui ne sont pas reliés par une telle relation (et que l'on catégoriserait donc comme du bruit). Nous présentons dans le tableau 2.4 des exemples d'associations lexicales selon nous indésirables.

Couples	Contextes communs
`recherche/nord`	s'orienter_vers, pôle_de, département_de
`bouteille/dé`	jeter_obj, lancer_obj, utiliser_obj
`bâtir_obj/orgue_de`	basilique, cathédrale, église
`rédiger_obj/exister_dans`	constitution, dictionnaire, bible
`traditionnel/sombre`	costume_mod, esthétique_mod, chanson_mod

TABLEAU 2.4 – Voisinage et associations indésirables

L'évaluation d'une ressource distributionnelle repose sur la question de savoir quelles sont les paires de mots qui présentent une relation de sens pertinente, et quelles sont celles qui sont catégorisées comme étant du *bruit*. À cette question, chacun des modes d'évaluation que nous avons listés (section 2.1.3) apporte une réponse différente.

– Si l'on mène une évaluation extrinsèque, un lien pertinent est un lien exploitable, dont la prise en compte améliore les performances d'une tâche donnée. De nombreuses applications, comme la segmentation thématique que nous

abordons dans le chapitre 5, reposent sur la prise en compte de liens de proximité sémantique au sens large (parfois appelées *loose associative relations* van der Plas (2008)).
- Dans le cas de l'évaluation par comparaison à un *gold standard*, un lien pertinent est un lien répertorié dans des ressources faisant référence. Cela revient souvent à considérer comme « pertinents » les seuls couples de mots qui présentent une relation classique, recensée dans les typologies traditionnelles : synonymie, antonymie, hyperonymie, méronymie, etc.
- Enfin, l'évaluation par confrontation à des jugements humains n'apporte pas *a priori* de réponse définitive à cette question : un lien pertinent est un lien intuitivement perçu comme tel par des locuteurs ; il est toutefois nécessaire de proposer une consigne d'annotation à ces locuteurs, et donc de dégager des éléments de réponse.

Ainsi, la question de la pertinence des paires lexicales évalués est étroitement liée à la question de la contextualisation de ces paires : on ne peut réellement juger de la pertinence d'un lien que dans un contexte donné – contexte applicatif dans lequel le lien est ou non exploitable, ou cotexte où se réalise la relation sémantique entre les deux mots qui le composent.

S'il est ainsi très difficile d'évaluer la proportion de bruit contenue par notre ressource distributionnelle, nous notons toutefois dès à présent qu'il ne s'agit absolument pas d'un phénomène marginal : l'observation des *Voisins de Wikipédia* donne très vite à voir une très grande proportion de couples dont la relation est très difficilement interprétable. Cela amène à s'interroger sur le filtrage des voisins ; comme nous l'avons vu, il est possible pour cela de jouer sur de nombreux critères, dont le plus évident est le score de Lin.

Dans cette thèse, nous aborderons la question de la pertinence des paires de voisins conjointement à celle de la détection de la cohésion lexicale lors d'une phase d'annotation de liens (chapitre 4), puis par le biais de l'apport des voisins distributionnels à la tâche de segmentation thématique (chapitre 5).

2.4 Bilan

Nous avons dans ce chapitre présenté les grands principes de l'analyse distributionnelle et détaillé la chaîne de traitement ayant abouti à la construction de la ressource qui est utilisée dans le cadre du projet VOILADIS, les *Voisins de Wikipédia*. Nous avons donné des éléments de description de cette ressource en mettant en parallèle des observations faites sur la ressource et des réflexions menées sur son mode de constitution. Cela nous a permis de souligner des points essentiels dans la problématique de cette thèse.
- La variété des liens lexicaux mis au jour par l'analyse distributionnelle peut être rapprochée de la variété des liens impliqués dans la cohésion lexicale.
- Le fait que les relations extraites, bien que par nature paradigmatiques, pré-

sentent des propriétés syntagmatiques incite à envisager la reprojection de ces liens en corpus.
- Enfin, nous avons soulevé le délicat problème du bruit contenu par cette ressource, très important dans le cadre de son exploitation.

Le prochain chapitre est consacré au problème de la re-projection des liens de voisinage en texte.

Chapitre 3

(Re)projection des Voisins de Wikipédia en texte

Sommaire

Afin d'exploiter la ressource décrite dans le chapitre 2 dans le cadre d'approches du discours, il est nécessaire de projeter les liens contenus dans cette ressource sur les textes que l'on souhaite traiter. Ce chapitre est consacré à cette question de la projection des voisins en texte. C'est ici que commence la description de notre travail de thèse à proprement parler : alors que la construction des *Voisins de Wikipédia* est antérieure – et indépendante – au projet VOILADIS, la question de leur reprojection sur leur corpus d'origine n'avait pas été jusque-là traitée.

3.1 Préambule

Dans ce chapitre, nous envisageons la projection des *Voisins de Wikipédia* sur des textes issus de WikipédiaFR2007. On pourrait en fait parler de *re*projection, puisque les liens projetés ont été calculés sur la base de ce même corpus. Comme nous l'avons déjà expliqué en 1.5.2.2, nous n'envisagerons pas dans cette thèse la projection sur des textes n'appartenant pas au corpus d'origine. Il s'agit donc d'un aller-retour, d'une projection des liens calculés globalement sur la totalité d'un corpus (sans faire aucun cas de la notion de texte) sur des textes particuliers de ce corpus. Aucune intervention manuelle ne prend place dans la totalité de la chaîne de traitement (extraction des voisins par SYNTEX-UPÉRY → reprojection de ces voisins sur leur corpus d'origine), afin d'assurer la portabilité de cette chaîne de traitement à un corpus différent.

La reprojection des voisins distributionnels dans leur corpus d'origine a déjà été envisagée dans le cadre d'autres problématiques. Fabre et Bourigault (2006), qui étudient les couples nom/verbe mis au jour par l'analyse distributionnelle, proposent d'utiliser un critère de cooccurrence pour sélectionner un sous-ensemble de paires à caractériser. La remise en contexte leur permet de sélectionner des couples de voisins plus pertinents (la proximité textuelle étant un indice supplémentaire de proximité sémantique), mais également de guider l'évaluation des relations sémantiques par leur visualisation au sein du discours. Morlane-Hondère et Fabre (2010) croisent voisinage distributionnel et projection de patrons morpho-syntaxiques dans leur corpus d'origine afin d'étudier les manifestations de la relation d'antonymie.

Comme nous l'avons souligné, ce chapitre présente les premiers apports du projet VOILADIS. Notre objectif, rappelons-le, est de prendre les *Voisins de Wikipédia* en entrée et de mettre en place l'ensemble des traitements permettant de les exploiter dans le cadre de l'analyse du discours, de leur projection en texte à la modélisation et à l'exploitation des liens projetés. Dans la suite de ce chapitre, nous décrivons les traitements mis en place pour la projection des voisins (section 3.2). Nous décrivons ensuite les sorties obtenues au terme de ces traitements (section 3.3). Puis nous discutons de ce que l'on peut faire émerger de ces sorties très imposantes (section 3.4) : des objets plus complexes que de simples liens de voisinage, ou encore des scores évaluant la cohésion lexicale d'un pan de texte.

3.2 Réalisation matérielle

> [...] l'histoire est entièrement vraie puisque je l'ai imaginée d'un bout à l'autre. Sa réalisation matérielle proprement dite consiste essentiellement en une projection de la réalité, en atmosphère biaise et chauffée, sur un plan de référence irrégulièrement ondulé et présentant de la distorsion. On le voit, c'est un procédé avouable s'il en fut.
>
> Boris Vian, L'écume des jours (Avant-propos)

Dans cette section, nous expliquons comment nous avons procédé pour projeter les *Voisins de Wikipédia* dans des textes issus de WikipédiaFR2007 (3.2.1). Nous posons ensuite la question de la visualisation des liens projetés (3.3.3). Enfin, nous dressons un premier bilan de cette projection (3.3.2).

3.2.1 Approche globale, approche locale

Différents dispositifs de projection des voisins ont été mis en place durant cette thèse. Le plus couramment utilisé a consisté à annoter tous les liens de voisinage présents dans un texte ou dans une portion de texte. Nous avons également, plus ponctuellement, mis en place des dispositifs spécifiques pour projeter les voisins dans des conditions particulières. Par exemple, nous avons été amenée à rechercher les liens de voisinage reliant deux portions de texte – et non pas les liens internes à chaque portion de texte – ou encore à vérifier si deux mots, ou deux segments de textes sélectionnés pour des conditions très spécifiques, étaient reliés par un lien de voisinage – il ne s'agit plus alors à proprement parler de « projection », puisqu'on sélectionne d'abord des mots candidats, puis qu'on vérifie s'ils sont reliés ou non par un lien de voisinage.

La figure 3.1 illustre la procédure suivie pour projeter globalement les voisins sur des textes (ou des portions de texte). À partir du corpus ayant servi à calculer la base de voisins distributionnels, un sous-corpus est constitué sur des critères spécifiques à l'expérience menée. Ce corpus subit un prétraitement, puis tous les liens de voisinage qu'il contient sont identifiés. La sortie produite, très riche et difficilement exploitable (pour des analyses linguistiques manuelles) en l'état, subit généralement d'autres traitements spécifiques avant la phase d'analyse. Cette figure, générale, est reprise et adaptée à une expérimentation particulière dans la section 5.3.1 (figure 5.4 page 164).

La figure 3.2 illustre la procédure suivie pour interroger la base de voisins localement, sur des données ciblées. À partir du corpus ayant servi à calculer la base de voisins distributionnels, certaines données sont extraites : il peut s'agir d'objets discursifs annotés lors d'une campagne d'annotation, ou de segments de texte repérés par des patrons morpho-syntaxiques ; des items lexicaux, dont on veut savoir s'ils

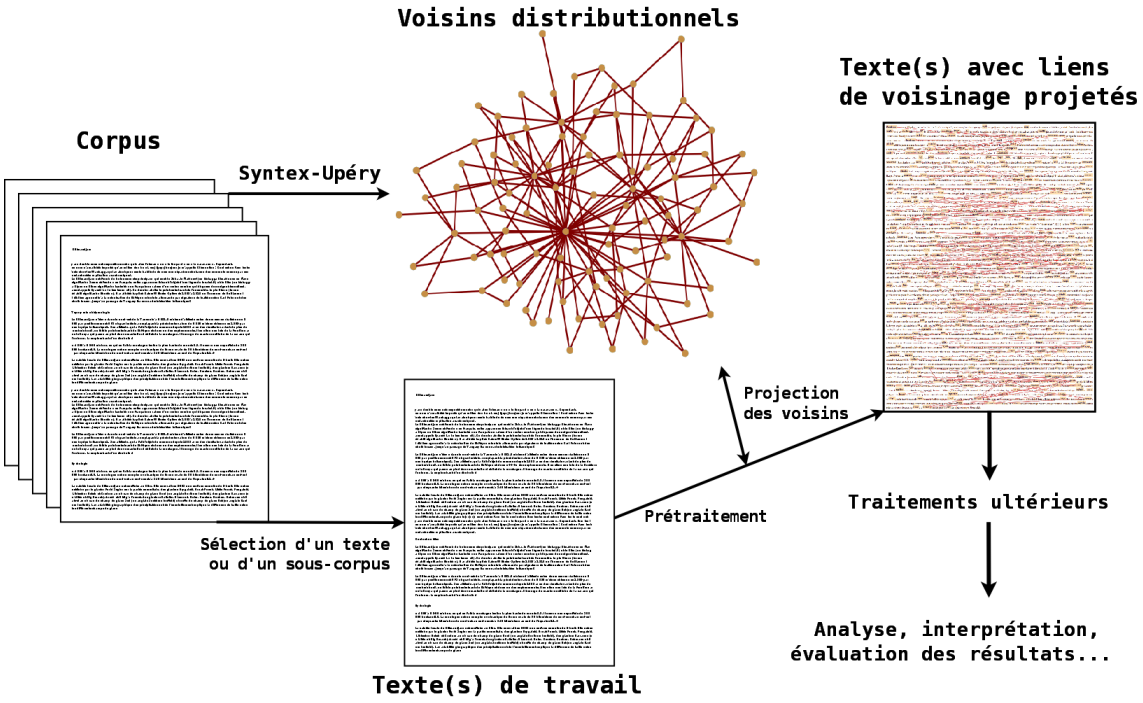


FIGURE 3.1 – Projection des voisins en texte

sont reliés par un lien de voisinage, sont identifiés. On interroge alors la base de voisins distributionnels sur ces mots candidats. Les données annotées produites peuvent directement être analysées. Cette figure est reprise et adaptée à une expérimentation particulière dans la section 8.3.2 (figure 8.2 page 247).

Les deux procédures illustrées sont très proches, leur principale différence résidant dans le point de vue global ou local (ciblé) que l'on adopte sur les données textuelles traitées.Dans les sections qui suivent, nous décrivons, conjointement pour ces deux types de procédures, les différentes étapes citées ci-dessus : sélection des données de travail (section 3.2.2), prétraitement de ces données (section 3.2.3), projection des voisins distributionnels (section 3.2.4).

3.2.2 Sélection des données de travail

Les données textuelles utilisées varient selon les expériences menées ; elles sont par contre toujours extraites du corpus WikipédiaFR2007 précédemment décrit en 2.2.1). Nous donnons ici un aperçu des critères auxquels il a été fait appel pour sélectionner ces données de travail.

Afin de constituer des sous-corpus de travail lors des différentes expérimentations menées au cours du projet VOILADIS, nous avons utilisé différents critères, parfois cumulés entre eux :

– textes sélectionnés de manière arbitraire, afin de présenter un échantillon varié du corpus (*cf.* chapitre 4) ;

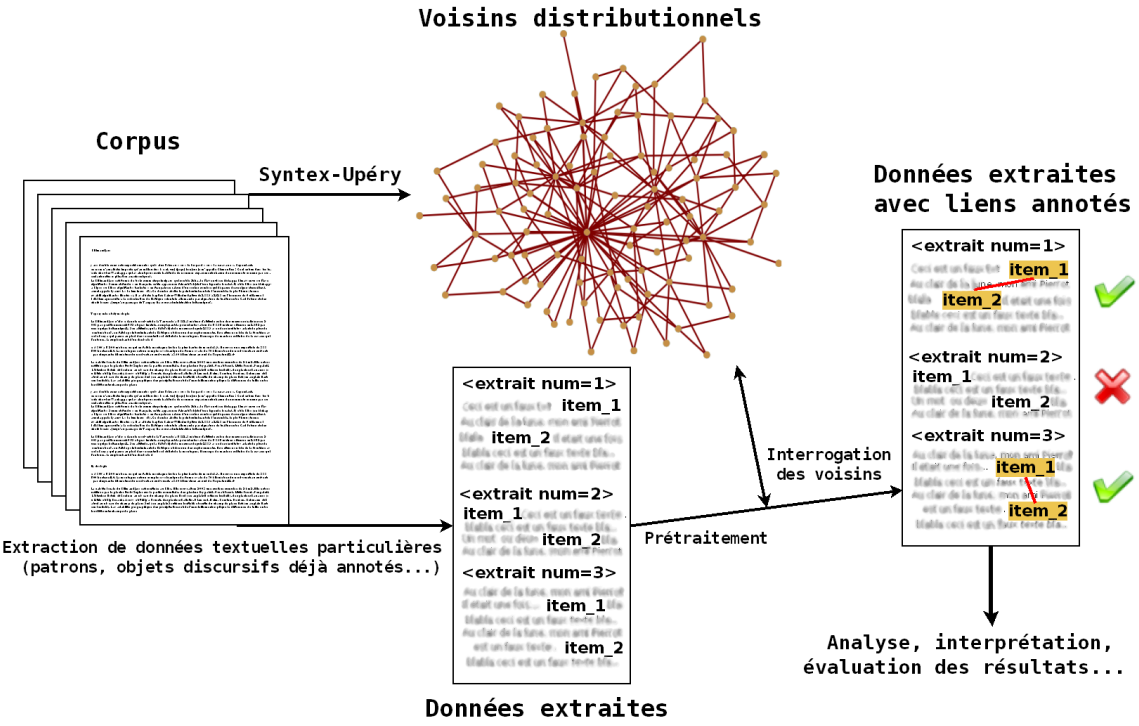


FIGURE 3.2 – Interrogation de la base de voisins sur des liens lexicaux ciblés

- textes extraits sur la base de leur caractéristiques structurelles : longueur, nombre de niveaux de titres, etc. (*cf.* chapitre 5) ;
- textes sélectionnés pour leur appartenance à un domaine particulier, par exemple la géographie (*cf.* chapitre 5) ;
- textes annotés lors de la campagne ANNODIS dans son versant « descendant » (*cf.* chapitre 6).

Lorsque nous nous situions dans une approche locale, nous avons sélectionné des données :

- sur la base de patrons morpho-syntaxiques, afin d'extraire par exemple des phrases contenant un syntagme gérondif dans la section 8.3.1 (*cf.* chapitre 8) ;
- sur la base d'annotations discursives réalisées lors de la campagne ANNODIS dans son versant « ascendant » (*cf.* chapitre 8).

Ces choix sont développés et justifiés dans les sections concernées.

3.2.3 Prétraitements

Les textes sont d'abord annotés avec le TreeTagger (Schmid, 1994). La segmentation en mots et l'annotation en catégories morpho-syntaxiques utilisées par la suite sont donc celles du TreeTagger. Rappelons que les *Voisins de Wikipédia* ont été calculés à partir d'une analyse syntaxique complète du corpus WikipédiaFR2007, effectuée par l'analyseur SYNTEX. Notons toutefois que cet analyseur syntaxique

prend lui-même en entrée l'étiquetage effectué par le TreeTagger. Malgré cela, en se basant sur des textes seulement annotés morpho-syntaxiquement on est bien sûr confronté à une perte d'information, pour deux raisons.

D'une part, pour des raisons de segmentation : lors de la projection, on se base sur la segmentation en mots effectuée par TreeTagger. Or, certains items parmi les *Voisins de Wikipédia* sont des unités plus complexes, notamment des syntagmes nominaux. Ces items sont donc ignorés dans nos traitements. Nous n'avons pas cherché de solution permettant de pallier ce problème, dans la mesure où il ne concerne qu'une faible proportion des paires recensées dans les *Voisins de Wikipédia* : seules 2.14% de ces paires comprennent un item de taille supérieure à un mot graphique.

D'autre part, parce que l'on perd l'information de nature syntaxique encodée avec les prédicats (2.2.2). Par exemple, les prédicats `manger_obj` et `manger_suj` sont deux items différents dans les *Voisins de Wikipédia* (et ils n'ont donc pas les mêmes voisins), alors que dans un texte annoté morpho-syntaxiquement, il n'y a qu'un seul verbe « manger ». Concernant ce point, nous avons fait le choix de fusionner tous les prédicats construits à partir d'un même lemme et d'une même catégorie ; ainsi, nous ne considérons qu'un seul item `manger`, ayant pour voisins l'union des ensembles de voisins de `manger_obj` et `manger_suj` [1]. Cette stratégie paraît particulièrement défendable dans un exemple comme celui de `manger`. En effet, une occurrence de ce verbe au sein d'un texte sera généralement concernée par les deux ensembles de voisins (puisqu'elle pourra être conjointement pourvue d'un sujet et d'un objet). Par contre, il peut paraître délicat de fusionner deux prédicats comme `donner_obj` (qui a par exemple pour voisins `obtenir_obj` et `posséder_obj`) et `donner_sur` (qui a par exemple pour voisins `ouvrir_sur` et `vue_sur`). En effet, dans ce cas, les deux prédicats correspondent à deux usages distincts du verbe « donner », qui ne seront jamais combinés pour une même occurrence de ce verbe ; la stratégie adoptée pourrait donc, ici, donner lieu à du bruit. Il faut cependant relativiser cette dernière observation en gardant à l'esprit que même en absence de fusion les problèmes de polysémie ne seraient pas totalement évités.

Nous avons choisi de nous en tenir à une annotation morpho-syntaxique car la perte d'informations résultante nous paraissait avoir des conséquences minimes par rapport à la complexité nouvelle engendrée par la prise en compte d'informations plus riches, c'est-à-dire d'une annotation syntaxique en entrée de nos traitements. En effet, il est bien plus facile de supposer qu'un item dans le texte est équivalent à un seul item dans la base de voisins, que d'y associer un nombre variable de prédicats, comme c'est le cas dans l'exemple (1) où le verbe « donner » est concerné par les trois prédicats `donner_suj`, `donner_obj` et `donner_à`.

(1) En revanche, écrire que « le préfet Poubelle a donné son nom à l'ustensile

1. Si un voisin est présent dans les deux ensembles, on recalcule le score de Lin en faisant la moyenne des deux scores existants

éponyme » est incorrect. (Wikipédia, article « Éponymie »)

De plus, bien qu'elle ne soit pas abordée dans cette thèse, nous gardons à l'esprit l'éventuelle portabilité de la ressource et de nos algorithmes à des textes n'ayant pas servi à sa construction. Nous tenions donc à n'utiliser dans nos traitements que des outils disponibles, ce qui est le cas de TreeTagger mais pas de SYNTEX.

3.2.4 Algorithme de projection des voisins

Pour projeter les voisins en texte, nous utilisons une base de données dans laquelle tous les prédicats ayant le même lemme et la même catégorie ont été fusionnés, afin de correspondre aux items annotés morpho-syntaxiquement lors du prétraitement décrit dans la section précédente. Pour reprendre l'exemple cité précédemment, on ne considère plus qu'un item de voisinage (`manger_V`) au lieu de 9 (`manger_avec`, `manger_dans`, `manger_de`, `manger_en`, `manger_obj`, `manger_par`, `manger_suj`, `manger_sur` et `manger_à`). Les voisins de `manger_V` sont constitués par l'union des ensembles de voisins des différents prédicats concernés. Si plusieurs prédicats ont un même voisin, les scores de Lin des deux (ou plus) paires sont moyennés pour obtenir le score de Lin de la nouvelle paire. Même si d'autres formes de combinaison des scores de Lin auraient été envisageables (par exemple, sélection du score maximal), la moyenne apparaît ici appropriée car elle laisse s'exprimer l'information de chacun des scores individuels. Les rangs (*cf.* section 2.3.3) sont recalculés sur la base du nouveau classement obtenu. Le tableau 3.1 présente un court extrait de la base de données obtenue.

item1	cat1	item2	cat2	rang1-2	rang2-1	Lin
atteindre	V	taille	N	537	180	0.19
mètre	N	taille	N	143	1050	0.12
taille	N	volume	N	28	9	0.25

TABLEAU 3.1 – Format de la base de donnée utilisée pour la projection

Nous avons développé des scripts Perl qui interrogent la base de données des voisins distributionnels afin de les projeter dans les textes. Dans le cas d'une interrogation locale, des items lexicaux candidats sont déjà identifiés, et il suffit de vérifier si les paires lexicales concernées correspondent à une entrée de la base des voisins. Dans le cas d'une projection globale, nous procédons en deux temps, que nous illustrons à partir de la phrase (2) dans laquelle les numéros des items apparaissent.

(2)

Redressés,	les	gorilles	atteignent	une	taille	de	1,75	mètre.
704	705	706	707	708	709	710	711	712

(Wikipédia, article « Gorille »)

– Le texte est d'abord parcouru linéairement ; pour chaque mot plein (verbes, noms et adjectifs), on vérifie s'il fait partie de la base des voisins ; si oui, il est relevé et sa position est mémorisée. Par exemple, pour la phrase (2), les mots suivants sont relevés :
`<voisin id="atteindre_V" items="`*`707`*`,741,778,974">`
(Comprendre : le verbe « atteindre » a 4 occurrences dans le texte « Gorille », aux positions 707, 741, 778 et 974.)
`<voisin id="mètre_N" items="619,`*`712`*`,743,976">`
`<voisin id="taille_N" items="`*`709`*`">`
– Le graphe de voisinage correspondant aux noeuds identifiés est ensuite construit. Dans notre exemple, deux couples peuvent être trouvés dans la base de données (3.1) :
`<couple id="c1285" voisins="atteindre_V,taille_N">`
`<couple id="c5947" voisins="mètre_N,taille_N">`
Par contre, `atteindre`/`mètre` n'est pas un couple de la base des *Voisins de Wikipédia*.

Lors de la phase de projection, il est possible de faire varier les différents paramètres émanant de la construction des *Voisins de Wikipédia* : score de Lin, rangs, catégories considérées, etc. (*cf.* section 2.2 page 57)

3.3 Caractérisation des sorties produites

3.3.1 Format de sortie

Au terme de la projection des voisins, les informations que l'on récupère en sortie sont les suivantes :

– le texte d'origine, dans lequel certains mots sont identifiés comme faisant partie des voisins ;
– l'ensemble des liens de voisinage reliant ces mots.

Nous montrons ici le format pour lequel nous avons opté pour consigner ces différentes informations. La figure 3.3 p. 77 montre la sortie correspondant à l'exemple (3), qui est un court paragraphe de deux phrases, extrait de l'article « Gorille ».

(3) Redressés, les gorilles atteignent une taille de 1,75 mètre, mais ils sont en fait un peu plus grands car ils ont les genoux fléchis. L'envergure des bras dépasse la longueur du corpus et peut atteindre 2,75 mètres. (Wikipédia, article « Gorille »)

Dans la section qui suit, nous caractérisons les sorties obtenues à la suite de ce traitement.

Dans ce paragraphe, 10 mots font partie des voisins (ils sont listés avec les balises `<voisin />`) :

– les verbes « atteindre » et « dépasser » ;

```
<texte id="gorille">
<titre niv="1" num="1">
Gorille
</titre>
[...]
<titre niv="2" num="4">
<item id="648" lcat="caractéristique_N">Caractéristiques</item>
</titre>
[...]
<paragraphe num="14">
Redressés , les gorilles <item id="707">atteignent</item> une <item
id="709">taille</item> de 1,75 <item id="712">mètre</item> , mais ils
sont en fait un peu plus <item id="722">grands</item> car ils ont les
<item id="727">genoux</item> fléchis . L' <item id="731">envergure</item>
des <item id="733">bras</item> <item id="734">dépasse</item> la <item
id="736">longueur</item> du <item id="738">corps</item> et peut <item
id="741">atteindre</item> 2,75 <item id="743">mètres</item> .
</paragraphe>
[...]
<voisin id="atteindre_V" items="707,741,778,974">
<voisin id="bras_N" items="733">
<voisin id="corps_N" items="738">
<voisin id="dépasser_V" items="734">
<voisin id="envergure_N" items="731">
<voisin id="genou_N" items="727">
<voisin id="grand_A" items="43,722,748,1440,1711">
<voisin id="longueur_N" items="736,885">
<voisin id="mètre_N" items="619,712,743,976">
<voisin id="taille_N" items="709">

<couple id="c1216" voisins="atteindre_V,corps_N">
<couple id="c1219" voisins="atteindre_V,dépasser_V">
<couple id="c1246" voisins="atteindre_V,longueur_N">
<couple id="c1285" voisins="atteindre_V,taille_N">
<couple id="c1472" voisins="bras_N,corps_N">
<couple id="c2722" voisins="corps_N,taille_N">
<couple id="c3087" voisins="dépasser_V,taille_N">
<couple id="c4107" voisins="envergure_N,longueur_N">
<couple id="c4111" voisins="envergure_N,taille_N">
<couple id="c5667" voisins="longueur_N,mètre_N">
<couple id="c5692" voisins="longueur_N,taille_N">
<couple id="c5947" voisins="mètre_N,taille_N">
</texte>
```

FIGURE 3.3 – Projection des *Voisins de Wikipédia* : Format de sortie

– les noms « bras », « corps », « envergure », « genou », « longueur », « mètre » et « taille » ;
– l'adjectif « grand ».

Si l'on considère ces 10 noeuds dans le graphe des *Voisins de Wikipédia*, on observera 12 arcs, c'est-à-dire que 12 relations de voisinage connectent ces 10 items. Les 12 couples de voisins concernés sont listés à l'aide des balises `<couple />`. Le sous-graphe de voisinage correspondant est représenté dans la figure 3.4. Dans ce sous-graphe, nous avons représenté tous les mots du texte faisant partie de la base des voisins. Les voisins appartenant à un même couple sont reliés d'un trait. On peut remarquer qu'un mot du texte, « genou », n'est impliqué dans aucun couple.

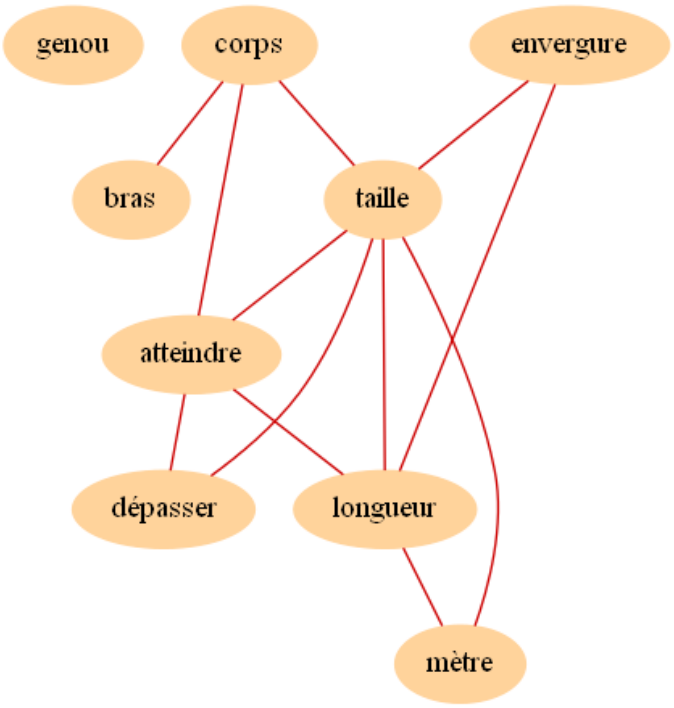


FIGURE 3.4 – Graphe des voisins impliqués dans l'exemple (3)

Replaçons-nous à présent dans le texte. Il est important de noter que les 12 couples de voisins listés représentent en fait plus de 12 liens dans le texte. En effet, deux des voisins identifiés (`atteindre` et `mètre`) correspondent chacun à deux occurrences dans le texte (items 707 et 741 et items 712 et 743) ; tous les couples dans lesquels sont présents ces voisins correspondent donc à deux liens dans le texte. La projection des *Voisins de Wikipédia* sur cet extrait permet ainsi de mettre au jour 18 liens de voisinage.Ces liens sont représentés dans la figure 3.5.

Nous parlerons désormais de « couple de voisins » pour désigner une paire de la base des *Voisins de Wikipédia* (par exemple le couple `atteindre`/`taille`), et de lien de voisinage pour désigner une occurrence particulière d'un couple au sein d'un texte. Ce sont ces liens projetés dans le texte qui nous intéressent dans cette thèse,

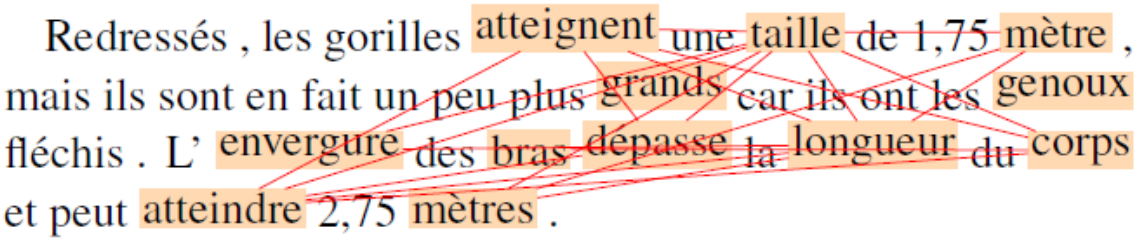


FIGURE 3.5 – Liens projetés dans le texte

ceux dont on veut déterminer s'ils rendent compte de la cohésion lexicale du texte considéré : si l'on veut utiliser les voisins afin de dégager des éléments de la structure des textes, on a besoin de l'information de la position particulière de chaque paire de mots en relation de voisinage. Néanmoins, le graphe de voisinage du texte, tel que présenté en 3.4, constitue également un point de vue intéressant pour aborder le texte qu'il concerne (comme nous le verrons dans la section 3.4).

3.3.2 Analyse quantitative

Nous n'avons jamais, au cours du projet VOILADIS, appliqué notre algorithme à l'ensemble du corpus WikipédiaFR2007, mais uniquement à des sous-corpus construits selon des besoins spécifiques (*cf.* section 3.2.2). Nous dressons ce bilan « chiffré » sur la base du sous-corpus de 42 textes utilisé dans le chapitre 4 et décrit en 4.2.3. Nous avons projeté les *Voisins de Wikipédia* sur ce sous-corpus sans faire intervenir aucun filtre ou seuil sur les scores de similarité.

Le tableau 3.2 permet d'apprécier la quantité de liens projetés dans les textes. Nous donnons uniquement les moyennes par texte qui sont plus interprétables[2]. En moyenne, un peu moins de 30% des mots d'un texte correspondent à une entrée dans la base des *Voisins de Wikipédia*, ce qui correspond à plus de 60% des mots « pleins » du texte (*cf.* tableau 2.1 page 58). La projection de voisins permet donc une très bonne couverture des textes. Mais ce qu'il faut avant tout remarquer, c'est l'énorme quantité de liens projetés. On observe en moyenne près de 83000 liens de voisinage par texte. Il s'agit certes d'un maximum, puisque l'on a projeté tous les voisins, sans aucun seuillage en aval (différents seuillages ont cours dans le processus de construction des voisins). Mais ces nombres imposants incitent à la prudence. On voit ici qu'on manipule des objets sur lesquels il est très difficile de « garder la main », du fait de leur profusion.

Les très grands écarts-types visibles dans la troisième colonne du tableau 3.2 montrent que les quantités consignées sont extrêmement variables d'un texte à l'autre. Cette variabilité n'est pas arbitraire mais fondamentalement dépendante

2. De plus, additionner les nombres de liens par texte pour donner un nombre de liens pour tout le corpus n'a pas vraiment de sens puisque les liens ne dépassent pas la portée du texte.

	Nombre moyen / texte	Écart-type
Mots	2245.4	1344.3
Voisins (types)	369.7	183.5
Items (occurrences)	643.8	403.1
Couples de voisins	19076.1	19153.6
Liens de voisinage	82780.6	114135.6
Liens de répétition	943.1	1532.4
Liens (total)	83723.7	115413.4

TABLEAU 3.2 – Bilan quantitatif de la projection

de la taille des textes. En effet, le nombre de liens projetés augmente avec l'augmentation de la taille des textes. Ce phénomène n'est en rien surprenant et on pouvait intuitivement le prédire, puisque pour un couple de voisins, le nombre de liens correspondants est le produit des fréquences des deux voisins considérés. Mais nous tenions à souligner son ampleur et à en donner un aperçu.

Le tableau 3.3 reprend les informations du tableau 3.2, mais pour deux textes particuliers, un texte dit « court » (925 mots) et un texte dit « long » (6007 mots). Ce tableau permet de mieux comprendre les écarts-types précédemment soulignés. La figure 3.6 p. 81 permet de visualiser la progression quadratique de la quantité de liens de voisinage projetés à mesure que l'on augmente la taille du texte : on voit qu'un extrait de texte de 500 mots contient déjà plus de 3000 liens de voisinage.

	Texte « court »	Texte « long »
Mots	985	6007
Voisins (types)	266	1796
Items (occurrences)	186	803
Couples de voisins	3189	76577
Liens de voisinage	6193	483016
Liens de répétition	117	4929
Liens (total)	6310	487945

TABLEAU 3.3 – Comparaison entre deux textes de longueur différente

L'explosion du nombre de liens de voisinage avec l'augmentation de la taille de texte prise en compte pose souvent problème, notamment lorsque l'on veut comparer des structures linguistiques de tailles non-équivalentes. Nous ne faisons ici que poser ce problème ; nous préciserons par la suite les solutions que nous y avons apportées.

Un autre phénomène dont nous souhaitons donner un aperçu concerne la longueur des liens de voisinage, c'est-à-dire la distance qui sépare deux items lexicaux connectés. Le nombre de liens connectant deux pans de texte de même taille décroit régulièrement avec la distance séparant ces deux pans de texte. C'est ce que montre la figure 3.7 qui présente la répartition des liens de voisinage selon leurs

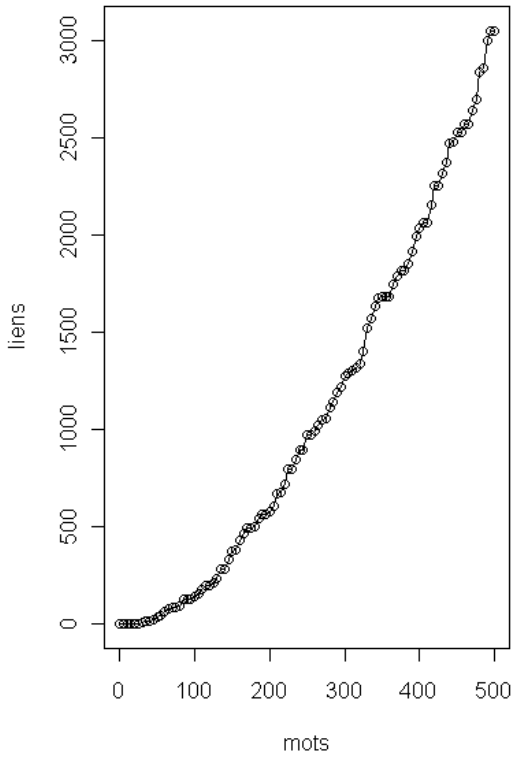


FIGURE 3.6 – Augmentation du nombre de liens projetés en fonction du nombre de mots d'un texte

longueurs. Ainsi, la proximité dans le texte (de deux phrases, segments...) a une influence directe sur le nombre de liens repérés entre ces deux unités.

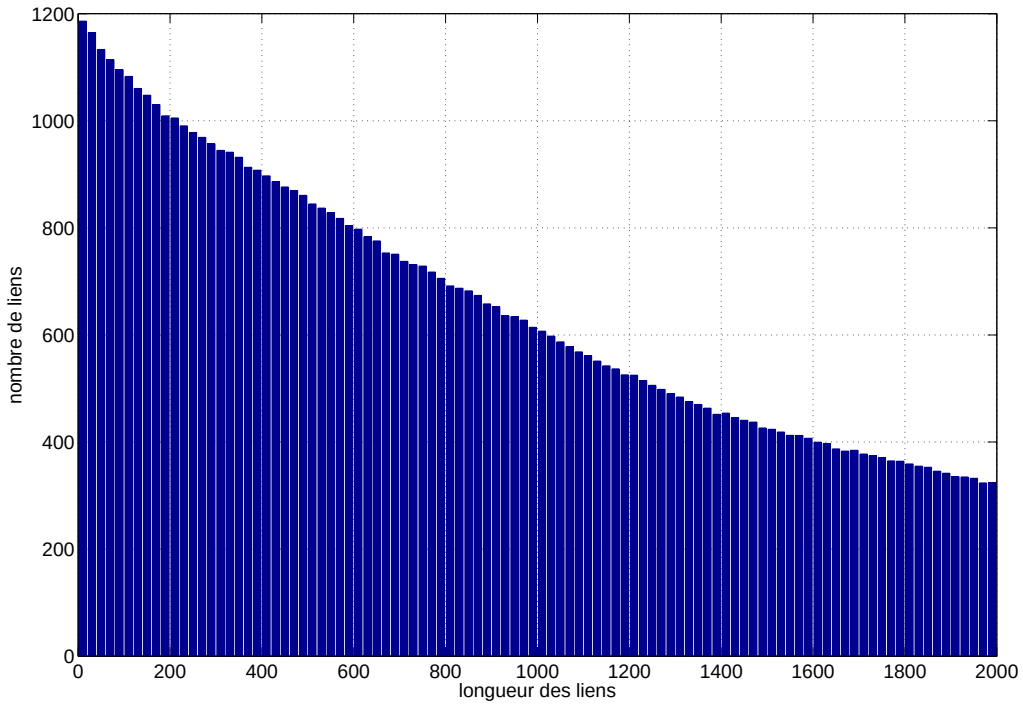


FIGURE 3.7 – Histogramme des longueurs des liens de voisinage

Au terme de cette section 3.2, nous avons présenté la mise en place de la projection des *Voisins de Wikipédia* sur des textes extraits de WikipédiaFR2007. Nous avons insisté sur la quantité de liens projetés, qui rend les sorties difficiles à exploiter pour le linguiste. On sait par ailleurs (2.3.3) que la base de voisins contient beaucoup de bruit, et on suppose naturellement que ce bruit s'est reporté sur les sorties de nos traitements. La suite de ce chapitre et le chapitre suivant sont consacrés à différentes stratégies mises en place pour regagner de la maîtrise sur ces sorties :

- par une modélisation différente des liens projetés, en prenant en considération des objets plus grands que le lien de voisinage, ou le calcul de scores de cohésion (section 3.4) ;
- par la mise en place d'une annotation des liens en contexte, permettant d'explorer différents critères permettant de limiter le nombre de liens projetés (chapitre 4).

3.3.3 Problème de la visualisation des liens en texte

Nous avons donné dans la section 3.3.1 (figure 3.3) un exemple de sortie sur un extrait de texte très court. On le voit, la lisibilité de ce format de sortie est réduite et rend impossible, sur des textes de longueur plus importante, l'exploitation

« humaine » des textes (analyse linguistique des liens projetés, etc.) : on est confronté au problème de la visualisation des liens en texte. Il serait possible de contourner cette question si l'on s'intéressait uniquement à l'exploitation informatisée des liens projetés et à l'évaluation de leur apport dans différentes approches du discours (par exemple, pour procéder à la segmentation thématique d'un texte en se basant sur les liens de voisinage qu'il contient, nul besoin de pouvoir visualiser ces liens). Mais dès lors que l'on veut un retour possible sur la ressource, que l'on veut analyser les liens utilisés, présenter des exemples, etc., se pose la question de la visualisation.

La visualisation des liens de voisinage dans les textes rencontre deux principales pierres d'achoppement.

L'existence d'outils adaptés Si de nombreux outils existent pour la visualisation de graphes (rendant par exemple possible une visualisation dynamique des *Voisins de Wikipédia*), il est par contre difficile de visualiser les liens au sein d'un texte, c'est-à-dire de visualiser un graphe en conservant l'information de l'agencement linéaire des mots du texte, dont certains constituent des nœuds du graphe et d'autres non. En bref, tracer automatiquement des traits entre certains mots particuliers d'un texte n'a rien d'anodin, et malgré nos recherches nous n'avons trouvé que peu d'outils permettant cela.

La quantité d'informations à visualiser Comme nous l'avons souligné dans la section 3.3.2, la quantité de liens projetés par texte est très importante (près de 500000 liens pour un texte de 6000 mots). Même lorsque l'on dispose d'un outil permettant de visualiser des liens en texte, cet outil ne permet pas toujours de manier une telle quantité d'informations. Et, s'il est possible d'afficher ces informations, elles sont alors difficiles à analyser du fait de leur profusion. Le problème de la visualisation est ici la trace de quelque chose qui pose également des problèmes d'ordre algorithmique : la difficulté à faire émerger une information que l'on puisse analyser, interpréter (du point de vue du linguiste) ou exploiter (d'un point de vue applicatif) parmi la masse d'informations présentes.

On peut noter que la problématique de la visualisation des données linguistique est un enjeu qui prend de l'importance en linguistique de corpus ; elle a notamment fait l'objet de l'un des tutoriels d'ACL 2008 (Collins *et al.*, 2008) Nous présentons dans la suite de cette section quelques outils et stratégies auxquels nous avons fait appel dans le cadre du projet VOILADIS.

3.3.3.1 Le *package* TikZ pour LaTeX

Nous avons recouru à l'utilisation d'un module LaTeX, le *package* TikZ. C'est ce *package* qui a permis de générer la figure 3.5, ainsi que toutes les figures d'apparence similaire dans cette thèse. La visualisation avec TikZ est particulièrement utile lorsque l'on veut visualiser en une image la totalité des liens projetés sur un texte ou

une portion de texte. Bien sûr, à l'échelle du texte, la trop grande quantité de liens rend les figures créées peu lisibles. La figure 3.8 montre un exemple de visualisation sur l'article Wikipédia « Domestication », ici tronqué à 3000 mots. Seules les paires de voisins dont le score de Lin dépasse 0.5 sont identifiés.

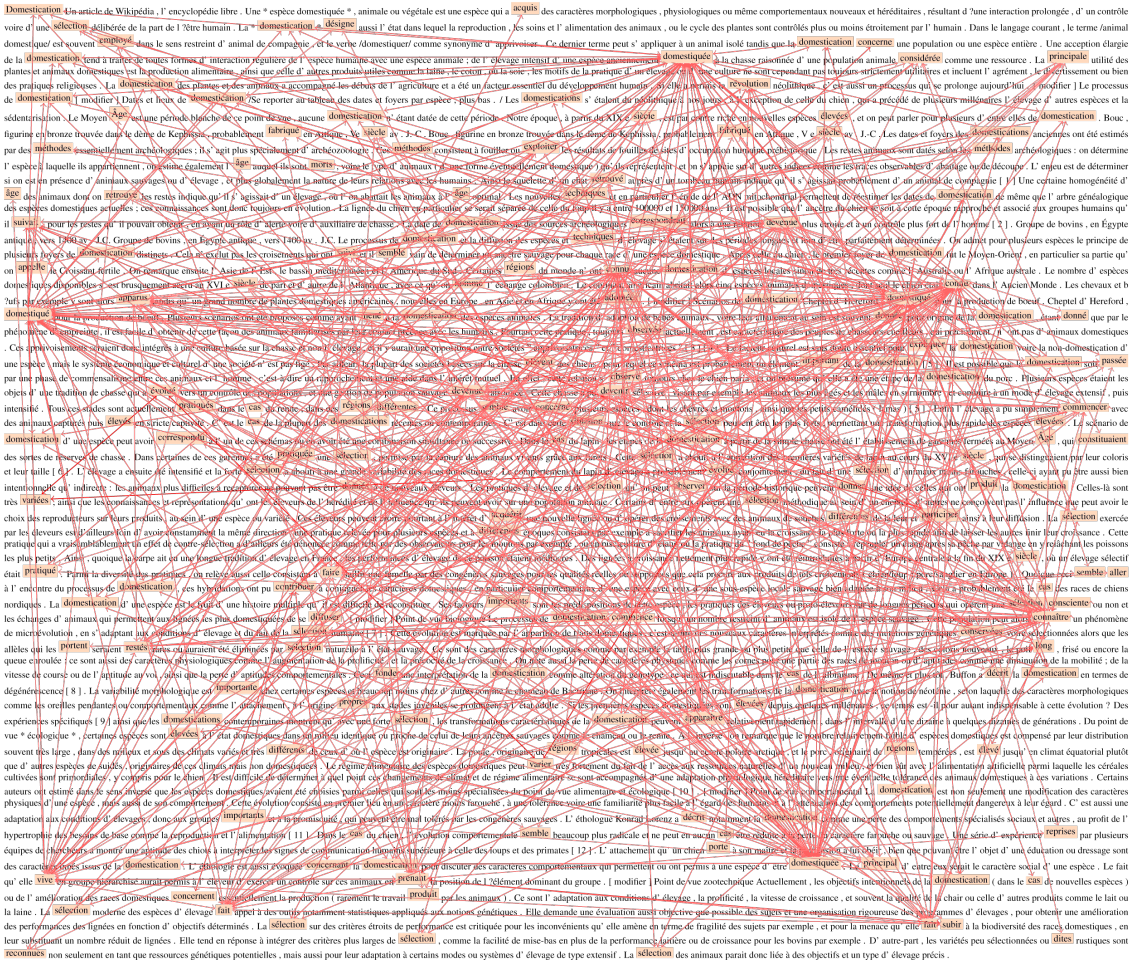


FIGURE 3.8 – Visualisation avec TikZ des liens projetés ($Lin >= 0.5$) sur le texte « Domestication » tronqué à 3000 mots

3.3.3.2 La plate-forme d'annotation Glozz

La plate-forme d'annotation Glozz (Mathet et Widlöcher, 2009) a été développée au GREYC (Caen) dans le cadre du projet ANNODIS. Elle offre un environnement générique pour l'exploration et l'annotation discursive de corpus. Elle permet de visualiser des liens entre mots en les reliant d'une flèche, de la même manière que TikZ. Dans la mesure où les annotations effectuées au cours du projet ANNODIS sont au format Glozz, cette plate-forme constitue une solution intéressante pour visualiser conjointement ces annotations discursives et les liens de voisinage projetés

dans les segments de texte concernés. Par contre, l'affichage des liens de voisinage à l'échelle du texte pose problème à cette plate-forme, qui peine à gérer une telle quantité d'informations.

La figure 3.9 présente un exemple de visualisation de liens lexicaux avec la plate-forme Glozz, à partir de l'exemple (2).

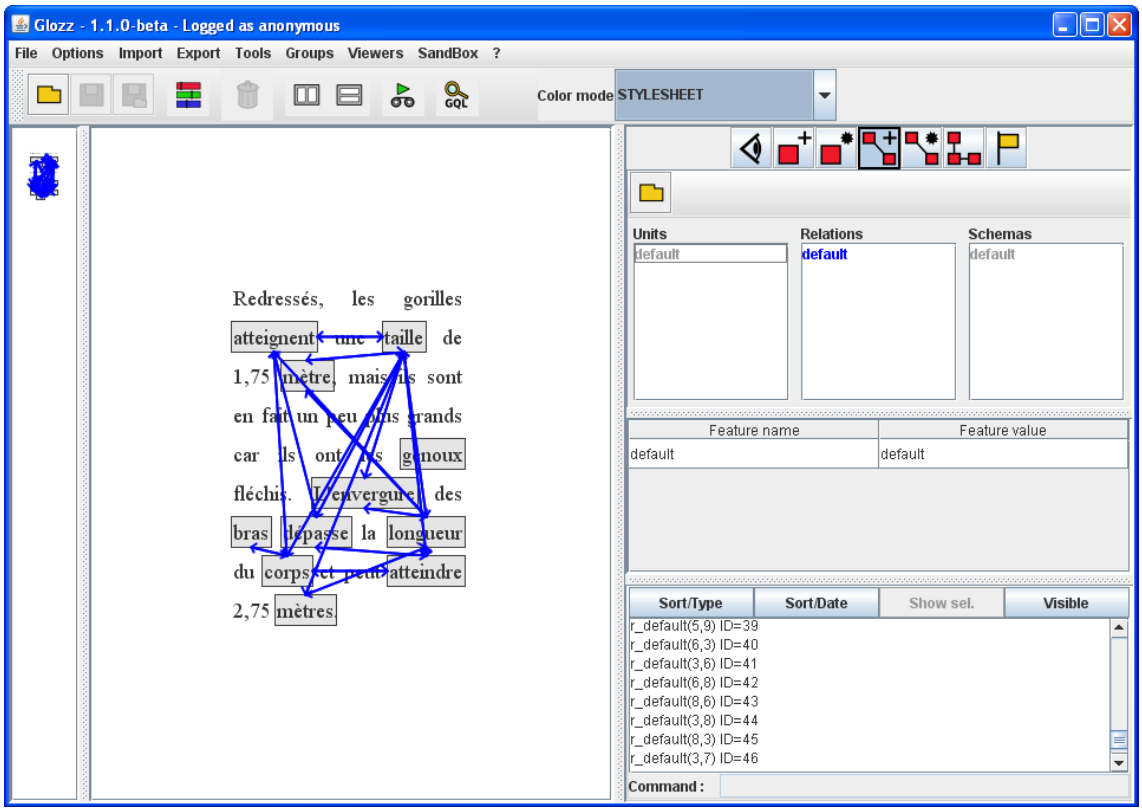


FIGURE 3.9 – Visualisation de liens lexicaux avec Glozz

3.3.3.3 Programmation d'interfaces de visualisation en JavaScript

Enfin, nous avons programmé des interfaces de visualisation dynamiques en JavaScript, permettant par exemple de cliquer n'importe quel item du texte pour mettre en surbrillance tous ses voisins. Cette solution a été particulièrement secourable pour observer les liens de voisinages projetés en se plaçant à l'échelle d'un texte entier : plutôt que de limiter la taille du texte visualisé, on limite cette fois l'information lexicale visible grâce à un affichage dynamique. Un exemple d'une telle interface est donné dans la figure 4.1 page 106.

3.4 Que peut-on faire émerger de la projection des voisins en texte ?

Nous avons expliqué dans ce chapitre comment nous procédions pour projeter les *Voisins de Wikipédia* en texte (section 3.2), et présenté les sorties « brutes » obtenues (section 3.3). Dans cette section, nous discutons des informations qui peuvent émerger des sorties présentées et être exploitées dans le cadre du projet VOILADIS, sans nous intéresser pour l'instant à l'analyse de ces informations.

Nous avons jusqu'ici parlé de liens de voisinage, reliant deux items lexicaux au sein d'un texte. Nous montrons dans cette section que des objets plus complexes, faisant intervenir plusieurs liens de voisinage, peuvent être envisagés (section 3.4.1). Faire émerger ces objets permet une modélisation différente des liens projetés, faisant émerger une information lexicale plus « lisible ».

Une autre façon d'exploiter les liens projetés est de les synthétiser *via* des scores de cohésion (section 3.4.2), qui donnent une évaluation de la cohésion globale d'un pan de texte, d'un objet discursif, ou de l'évolution de la quantité de liens de cohésion lexicale au fil d'un texte.

3.4.1 Des objets de taille supérieure au lien

Dans cette section, nous nous intéressons plus particulièrement au graphe des couples de voisins apparaissant dans un texte donné, qui est un sous-graphe des *Voisins de Wikipédia* (*cf.* figure 3.4). Nous présentons deux objets définis dans la théorie des graphes (Diestel, 2010) et les illustrons à partir des nos données : les cliques (section 3.4.1.1) et les composantes connexes (section 3.4.1.2).

3.4.1.1 Cliques

En théorie des graphes, une clique est un sous-ensemble de sommets tel que chacun de ces sommets est connecté à tous les autres. On parlera de k-clique pour désigner une clique de taille k, c'est-à-dire comprenant k sommets. Par exemple, dans la figure 3.10, les sommets `taille`, `atteindre` et `longueur` forment une 3-cliques, car chacun de ces sommets est connecté aux deux autres. Ce graphe contient de nombreuses autres 3-cliques (par exemple <`taille, envergure, longueur`> ou <`taille, longueur, mètre`>), mais aucune 4-clique.

Nous illustrons l'application du concept de clique à nos données à partir de l'exemple (4), toujours tiré de l'article « Gorille ».

(4) Les gorilles habitent les forêts et sont actifs le jour. Tandis que les gorilles des pays plats préfèrent les forêts tropicales humides, les gorilles des montagnes vivent plutôt dans les forêts secondaires. Les gorilles des montagnes se tiennent la plupart du temps au sol. Les gorilles des pays plats grimpent souvent dans les arbres à la recherche de nourriture, même les mâles lourds

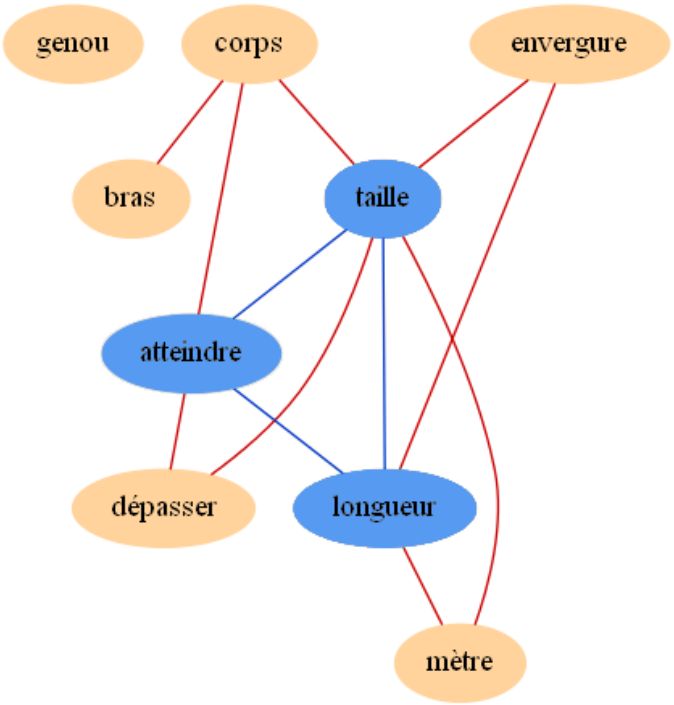


FIGURE 3.10 – Un exemple de 3-clique

> montent fréquemment dans des arbres dont la hauteur peut atteindre vingt mètres. À terre, les gorilles marchent à quatre pattes en s' appuyant sur les phalanges de leurs mains et non sur les paumes comme d' autres singes plus franchement quadrupèdes. Chaque nuit, pour se reposer, ils construisent un nid de feuilles en à peine cinq minutes. Les gorilles de montagne ont leurs nids à terre la plupart du temps, les gorilles des pays plats dans les arbres. (Wikipédia, article « Gorille »)

La figure 3.11 montre la totalité des liens de voisinage projetés dans ce paragraphe (208 liens correspondant à 129 couples). Nous avons calculé toutes les 3-cliques du sous-graphe de voisins correspondant (qui ne contient aucune clique de taille supérieure à 3). Les 18 cliques obtenues sont données dans le tableau 3.4

À ce stade, on peut déjà remarquer que le calcul de cliques constitue un moyen intéressant de sélectionner et de présenter l'information. Rien, bien sûr, ne permet de supposer que l'information sélectionnée est plus « pertinente » que celle qui ne l'a pas été. Mais il est certain que les 18 cliques listées offrent une vision plus synthétique du contenu lexical du paragraphe considéré que ne le ferait une liste des 129 couples de voisins impliqués. Sans pousser plus avant l'analyse, on peut constater que dans cet extrait, les cliques permettent de faire ressortir le thème principal de l'habitat du gorille. Des mots appartenant à des thèmes plus « accidentels » (« mâle », « patte », « phalange », « nourriture », etc.) n'apparaissent pas dans la liste des

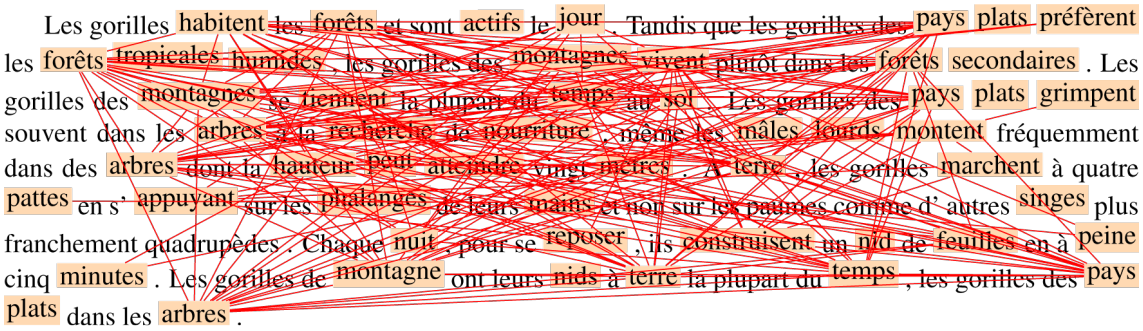


FIGURE 3.11 – Habitat du gorille : tous les liens

`hauteur_N mètre_N sol_N`	`appuyer_V pouvoir_V tenir_V`
`hauteur_N mètre_N temps_N`	`pays_N temps_N vivre_V`
`jour_N nuit_N temps_N`	`pays_N terre_N vivre_V`
`jour_N temps_N vivre_V`	`arbre_N sol_N terre_N`
`montagne_N pays_N terre_N`	`arbre_N feuille_N forêt_N`
`montagne_N pays_N vivre_V`	`recherche_N temps_N vivre_V`
`montagne_N sol_N terre_N`	`recherche_N terre_N vivre_V`
`montagne_N sol_N vivre_V`	`sol_N terre_N vivre_V`
`montagne_N terre_N vivre_V`	`atteindre_V construire_V habiter_V`

TABLEAU 3.4 - Listes des cliques calculées

cliques calculées car ils ne sont pas suffisamment connectés dans cet extrait de texte.

La figure 3.12 reprend l'exemple (4), en y projetant cette fois toutes les 3-cliques calculées et listées dans le tableau 3.4. Si cette figure est bien plus allégée que la figure 3.11, elle reste difficile à interpréter.

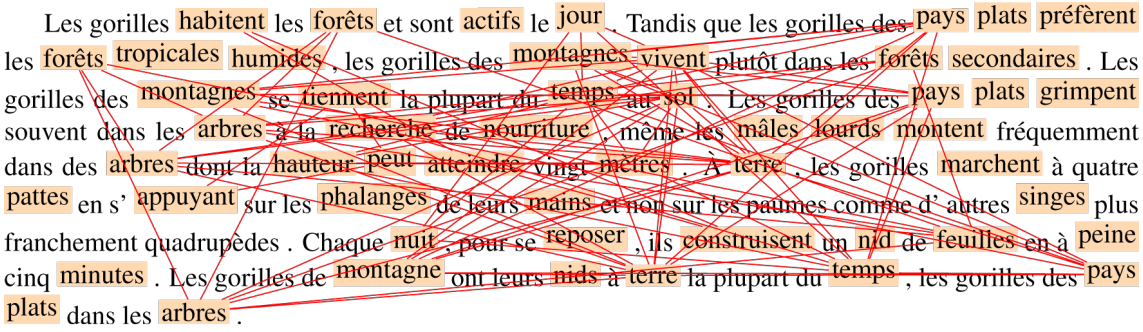


FIGURE 3.12 – Habitat du gorille : 3-cliques

Afin de mieux visualiser les cliques calculées en texte, nous avons opéré de la manière suivante. Nous avons calculé un score de Lin pour chaque clique, défini comme la moyenne des scores de Lin de tous les couples impliqués dans cette clique. Nous avons ensuite sélectionné les 4 cliques présentant les meilleurs scores :

– `jour_N nuit_N temps_N` ($Lin = 0.45$)

– `arbre_N feuille_N forêt_N` ($Lin = 0.41$)
– `hauteur_N mètre_N temps_N` ($Lin = 0.40$)
– `montagne_N pays_N terre_N` ($Lin = 0.40$)

Ce processus a donné lieu à la figure 3.13, ou chacune des 4 cliques sélectionnées apparaît dans une couleur différente.

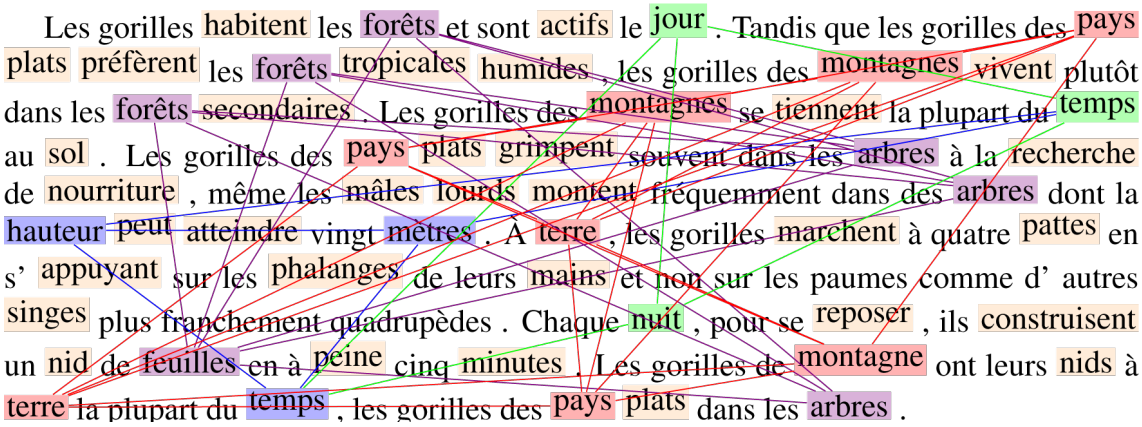


FIGURE 3.13 – Habitat du gorille : quatre 3-cliques ayant le meilleur score de Lin

L'information présentée dans cette figure apparaît beaucoup plus lisible et peut plus facilement donner lieu à des interprétations.

Le calcul de cliques sur des textes entiers fait émerger des cliques de taille plus importantes. L'annexe A.1 (page 285) montre les cliques maximales calculées sur un texte, l'article Wikipédia « Albanie ».

3.4.1.2 Composantes connexes

En théorie des graphes, un graphe est dit connexe s'il existe un chemin reliant n'importe lequel de ses sommets à tous les autres. Par exemple, le graphe de la figure 3.14 n'est pas connexe car aucun chemin ne permet de relier les sommets `genou` et `envergure`. Pour un graphe donné, une composante connexe est un sous-graphe connexe maximal de ce graphe. Ainsi, un graphe connexe contient une et une seule composante connexe. Le graphe de la figure 3.14 contient quant à lui deux composantes connexes : l'une est formée de l'unique sommet `genou`, la seconde de tous les autres sommets du graphe.

Pour illustrer le concept de composante connexe appliqué à nos données, nous nous basons sur le texte de la figure 3.15. Il s'agit des cinq paragraphes qui constituent la section intitulée « Caractéristiques », dans l'article Wikipédia « Gorille ». Dans ce texte ont été projetés 273 liens impliquant 213 couples différents.

Nous avons ensuite extrait toutes les composantes connexes du sous-graphe de voisins concerné. Nous avons obtenu 19 composantes :

– Une grosse composante de 25 sommets, comprenant les voisins : `sexe_N exister_V grand_A fait_N taille_N pouvoir_V atteindre_V`

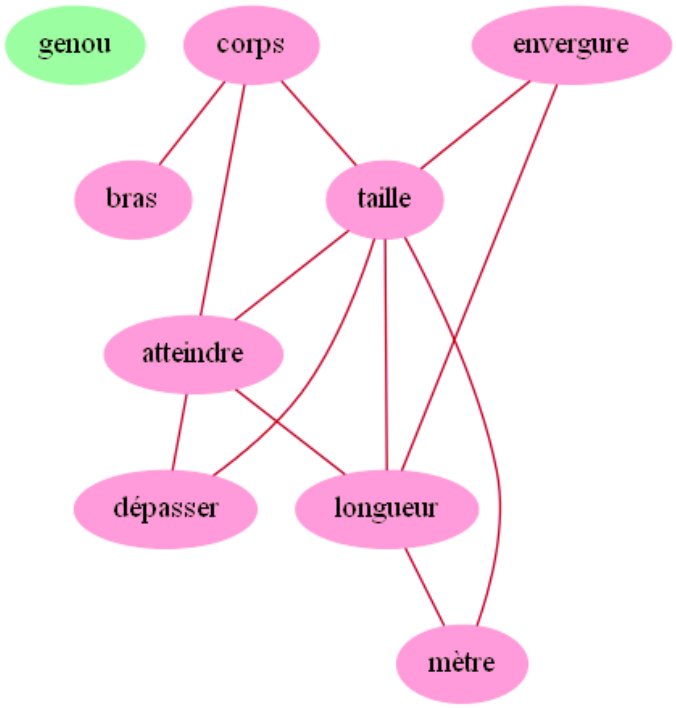


FIGURE 3.14 – Composantes connexes

Le gorille est après le bonobo et le chimpanzé, du point de vue génétique, l'animal le plus proche de l'humain. Cette parenté a été confirmée par les similitudes entre les chromosomes et les groupes sanguins. Notre génome ne diffère que de 2% de celui du gorille.
Redressés, les gorilles atteignent une taille de 1,75 mètre, mais ils sont en fait un peu plus grands car ils ont les genoux fléchis. L'envergure des bras dépasse la longueur du corps et peut atteindre 2,75 mètres
Il existe une grande différence de masse entre les sexes : les femelles pèsent de 90 à 150 kilogrammes et les mâles jusqu'à 275. En captivité, particulièrement bien nourris, ils atteignent 350 kilogrammes.
Le pelage dépend du sexe et de l'âge. Chez les mâles les plus âgés se développe sur le dos une fourrure gris argenté, d'où leur nom de « dos argentés ». Le pelage des gorilles de montagne est particulièrement long et soyeux.
Comme tous les anthropoïdes, les gorilles sont dépourvus de queue. Leur anatomie est puissante, le visage et les oreilles sont glabres et ils présentent des torus supra-orbitaires marqués.

FIGURE 3.15 – Texte d'exemple : caractéristiques du gorille

longueur_N mètre_N envergure_N corps_N bras_N différence_N masse_N dépasser_V groupe_N développer_V dépendre_V nom_N point_N vue_N animal_N confirmer_V différer_V peser_V

- 7 composantes comprenant 2 à 3 sommets :
  - génétique_A humain_A proche_A
  - dos_N pelage_N fourrure_N
  - oreille_N queue_N visage_N
  - chromosome_N génome_N
  - parenté_N similitude_N
  - âgé_A âge_N
  - femelle_N mâle_N
- 11 composantes connexes ne comprenant qu'un seul sommet (il s'agit donc de voisins non connectés au sein de l'extrait présenté, à la manière de « genou » dans la figure 3.14) : captivité_N gris_A nourrir_V kilogramme_N présenter_V sanguin_A montagne_N anatomie_N puissant_A genou_N long_A

Les composantes connexes les plus interprétables selon nous sont celles de taille intermédiaire, comprenant des voisins qui n'appartiennent pas au sous-graphe « principal » du texte mais sont néanmoins connectés. On peut remarquer qu'alors que les 3-cliques faisaient ressortir les zones les plus denses du graphe de voisinage du texte, comprenant des mots fortement connectés, ces composantes connexes de petite taille permettent de mettre au jour des couples de voisins isolés au sein du texte.

La figure 3.16 fait apparaître les composantes connexes de taille 2 et 3 dans le texte.

On peut remarquer que cette méthode fait ressortir différents thèmes au sein du texte (la génétique, la description physique, etc.). Il est d'ailleurs notable que chaque composante connexe reste interne à un seul paragraphe, à l'exception de la composante <femelle_N, mâle_N> qui s'étend également au paragraphe suivant.

Le calcul de composantes connexes sur des textes entiers implique de filtrer davantage les voisins, car sinon le graphe de voisinage associé au texte tend à ne comprendre qu'une seule composante (ou un nombre très faible). L'annexe A.2 page 287 présente un ensemble de composantes extraites d'un texte en faisant varier un seuil sur le score de Lin pour filtrer les voisins projetés.

3.4.2 Des mesures de cohésion lexicale

L'exploitation de la cohésion lexicale passe souvent pas le calcul de scores de cohésion, qui synthétisent l'information lexicale. Il peut s'agir de mesurer :

- la cohésion lexicale d'un pan de texte, d'un certain objet linguistique, comme les structures énumératives (*cf.* chapitre 6) ;
- la cohésion lexicale entre deux pans de texte : par exemple deux segments contigus dont on se demande s'il sont en continuité ou en discontinuité thématique (*cf.* chapitre 5), ou deux segments entretenant une certaine relation

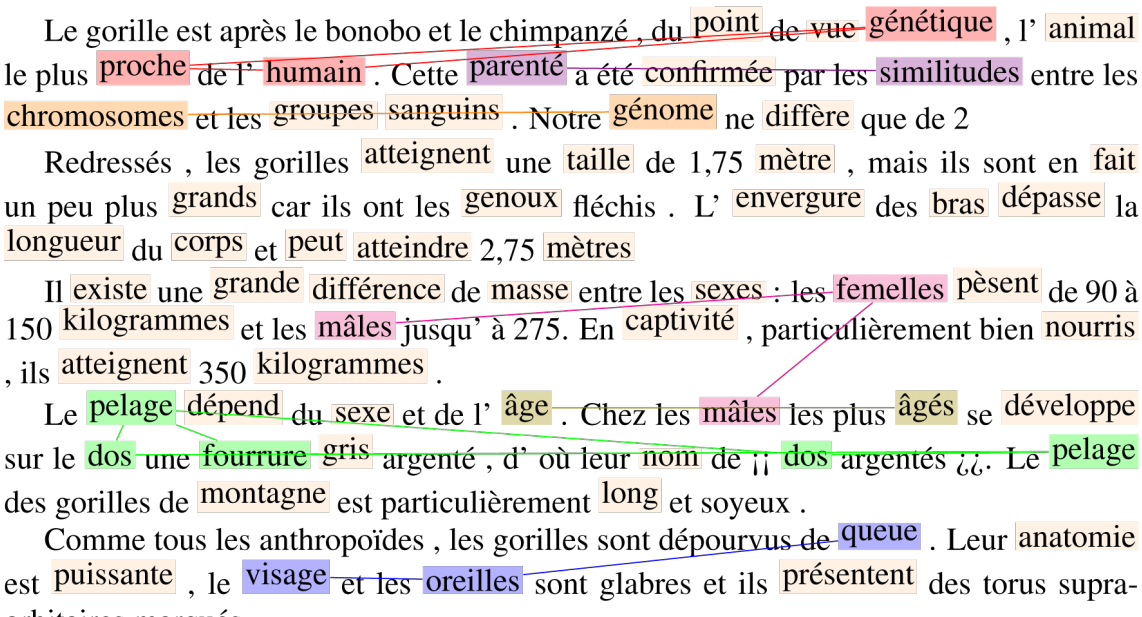


FIGURE 3.16 – Les composantes connexes de taille intermédiaire

de discours (*cf.* chapitre 7).

Le calcul d'un score de cohésion lexicale, étant donné un certain nombre de liens de similarité dotés d'une pondération (score de Lin ou autre appréciation de leur qualité), n'est pas un problème facile et plusieurs approches sont possibles.

(a) La première approche et la plus évidente consiste à compter le nombre de liens, éventuellement pondérés, intervenant à l'intérieur d'un pan de texte ou entre deux pans de textes ; le nombre obtenu est alors normalisé en fonction de la taille des segments. Cette approche est adaptée pour comparer des segments (ou paires de segments) de tailles assez similaires ; elle n'est pas adaptée à la prise en compte d'objets multi-échelles ou de segments très petits (les scores montrant alors une trop forte variabilité). Nous l'avons utilisée dans le cadre de la segmentation thématique, qui repose sur la comparaison de pseudo-phrases de tailles identiques ; la méthode de calcul est exposée dans la section 5.2.1 (page 154).

(b) Pour traiter des segments de très petite taille (typiquement : phrases ou propositions), un score basé sur le nombre de liens de cohésion lexicale n'est pas adapté car ces nombres sont trop faibles. Nous avons fait appel à des mesures d'agrégation de similarités, destinées à calculer des similarités entre phrases à partir des similarités entre mots qui les composent (Mihalcea *et al.*, 2006; Malik *et al.*, 2007). Le principe est que l'on prend en compte un score pour chaque lien (par exemple le Lin et un score de 1 pour les répétitions) et que l'on moyenne ces scores (ou un sous-ensemble de ces scores). Le score final ne dépend alors plus du nombre de liens de cohésion lexicale, mais de leurs forces. Ces mesures

d'agrégations ont été utilisées pour comparer des unités minimales de discours et sont décrites dans la section 7.2.1 (page 215).

(c) Afin de comparer des objets multi-échelles, nous avons proposé une autre approche, consistant à modéliser pour chaque objet le score de cohésion attendu dans le texte où il apparaît en fonction de sa taille. La cohésion est alors exprimée en termes d'écart réduit au score moyen ; plutôt que d'être affectée d'un score compris entre 0 et 1, l'unité considérée reçoit alors un score négatif si elle est moins cohésive que la moyenne, et positif si elle est plus cohésive. Cette méthode a été utilisée pour aborder la cohésion lexicale des structures énumératives, qui apparaissent à différents niveaux de grain et peuvent contenir entre 8 et 8000 mots ; elle est décrite dans la section 6.2.2 (page 191).

3.5 Bilan

Dans ce chapitre, nous avons présenté la projection des voisins en texte depuis un point de vue essentiellement technique : comment procède-t-on ? Qu'est-ce qui caractérise les sorties obtenues ? Quelles informations peut-on extraire de ces sorties ? Nous avons ce faisant soulevé différents problèmes rencontrés, liés à la grande quantité de liens projetés : le problème de la visualisation de ces liens, de la normalisation d'un score de cohésion les faisant intervenir, etc.

D'autres questions d'ordre plus qualitatif, linguistique, et qui sont au coeur du projet VOILADIS, n'ont pas vraiment été évoquées. Quelle est la nature des liens projetés ? Quelle proportion de bruit peut-on y observer ? C'est en répondant à ces questions que l'on peut franchir le pas entre « lien de voisinage distributionnel » et « lien de cohésion lexicale ». C'est pourquoi le prochain chapitre de cette thèse est consacré à l'exploration des liens projetés, et aborde la question de leur filtrage.

Chapitre 4

Du voisinage distributionnel à la cohésion lexicale

Sommaire

Dans ce chapitre nous questionnons la relation entre voisinage distributionnel et cohésion lexicale, à travers une phase d'annotation des liens de voisinage projetés dans les texte (chapitre 3) et la proposition de stratégies de filtrage.

4.1 Préambule

Une question très importante dans le cadre de cette thèse a été la suivante : doit-on prendre la ressource utilisée « telle qu'elle est », ou doit-on questionner cette ressource et tenter de la filtrer avant de l'exploiter ? Les deux positions concurrentes en réponse à cette question ont été envisagées au cours du projet VOILADIS. La première position consiste à assumer que la question du filtrage des voisins se situe hors du champ du projet VOILADIS, en décidant de prendre la ressource telle qu'elle est, et en tentant de montrer, malgré le bruit qu'elle contient, son apport pour la détection de la cohésion lexicale. En effet, on a vu (section 2.1.3) que l'évaluation d'une ressource distributionnelle, sur laquelle repose le principe d'un filtrage, constitue un problème très complexe.

Dans le déroulement chronologique de cette thèse, de premières expériences ont été menées avant de s'attaquer à ce problème du filtrage, l'idée étant de faire le choix de prendre la ressource « telle quelle ». Ces expériences, relatées par la suite, ont donné lieu à des résultats mitigés. Lorsqu'il s'agissait de regarder des phénomènes à petite échelle, en cherchant des couples de voisins dans des positions très ciblées, les résultats se sont avérés concluants. Par contre, lorsqu'il s'agissait de projeter les voisins à plus large échelle, par exemple en repérant tous les couples de voisins dans un texte en vue de le segmenter thématiquement, les résultats se sont montrés bien moins concluants. C'est pourquoi nous abordons finalement dans ce chapitre la question du filtrage des voisins, non pas *a priori* mais de manière intrinsèquement liée à leur projection en texte, en proposant des stratégies de filtrage élaborées à partir d'une annotation en contexte des liens projetés.

La première hypothèse sur laquelle repose notre choix d'annoter des liens en contexte pour un filtrage « en aval » des voisins peut être formulée de la manière suivante : nous pensons que l'annotation en contexte de liens de voisinage est une tâche plus naturelle (donc plus fiable) que l'annotation hors contexte de couples de voisins. Dans la section 4.2, nous décrivons le pré-test que nous avons effectué afin de vérifier cette hypothèse, avant de présenter le dispositif d'annotation mis en place. Se demander comment filtrer les voisins projetés, c'est se demander sur quelles « manettes » on peut jouer pour sélectionner le plus de « bons » liens en rejetant le plus de « mauvais » liens (selon l'annotation effectuée). Dans la section 4.3, nous listons les indices que nous avons dégagés dans l'optique d'évaluer leur incidence sur la qualité d'un lien de voisinage. Parmi ces indices, nous définissons des indices « textuels » ; la seconde hypothèse que nous souhaitons tester est que ces indices calculés sur la base des occurrences particulières des voisins dans un texte donné peuvent être utilisés, au même titre que le score de Lin, pour restreindre

les liens projetés. Lors de l'exploration des indices menée dans la section 4.4, nous montrons l'apport des différents indices définis, ce qui nous permet de confirmer cette seconde hypothèse ; nous évoquons également la possibilité d'une classification automatique des liens projetés. Sur la base de ces analyses, nous proposons enfin différentes stratégies de filtrage dans la section 4.5

4.2 Mise en place d'une annotation manuelle

4.2.1 Motivations

Dans le chapitre 2, nous avons évoqué différentes limites à l'exploitation des *Voisins de Wikipédia*. Il s'agissait, en bref, d'une part du caractère pléthorique de cette ressource, et d'autre part de la grande proportion de bruit qu'elle contient. Il ne nous a pas paru adéquat de chercher des solutions à cette très vaste problématique de l'exploitation des ressources distributionnelles en amont de la projection des *Voisins de Wikipédia* en texte. On peut donc tout naturellement supposer que les inconvénients cités se sont reportés sur les sorties produites. Nous avons ainsi évoqué en 3.3.2 le caractère difficilement maîtrisable des textes annotés produits, dû à la très grande quantité de liens projetés. Dans le préambule de ce chapitre (4.1), nous avons fait état des difficultés auxquelles nous avons été confrontée lorsque nous avons tenté d'exploiter ces sorties directement.

Nous avons donc décidé d'explorer la question du filtrage des liens de voisinage non pas en amont de la projection, mais en aval, en procédant à une annotation des liens projetés. Ce choix n'est pas un choix par défaut, il est au contraire fortement motivé.

1. Au terme de cette annotation, nous souhaitons dégager des critères permettant de prédire la pertinence d'une paire de voisins, et éventuellement utiliser ces critères pour filtrer les liens repérés dans les textes. Or, nous pensons que ces critères peuvent aussi bien dépendre de paramètres de la base de voisins que d'indices liés au texte.

2. Si l'on se situe dans la perspective de fournir un référentiel pour l'évaluation d'une ressource distributionnelle, notre stratégie nous paraît également comporter des avantages. En effet, un des intérêts d'une ressource distributionnelle est qu'elle reflète les relations qui opèrent dans un corpus donné ; il nous semble donc adéquat de replacer ces relations dans les textes de ce corpus pour évaluer ces relations.

3. L'ajout du contexte, outre qu'il permet de lever certaines ambiguïtés, permet de transformer une tâche relativement artificielle (émettre un jugement sur des couples de mots) en une véritable tâche d'annotation d'un phénomène linguistique (la cohésion lexicale) au sein de textes. Ainsi, lors de la phase d'annotation, c'est déjà le problème de la cohésion lexicale qui sera envisagé.

> On ne se demandera pas « qu'est-ce qu'un couple de voisins pertinent ? », mais « qu'est-ce qu'un lien participant à la cohésion lexicale d'un texte ? »

L'annotation *a posteriori* des liens projetés permet ainsi selon nous de mieux circonscrire le problème (*cf.* point 2 et 3), et d'élargir le champ des solutions possibles (*cf.* point 1). Au delà de ces considérations, il nous permet également de « rentabiliser » la phase d'annotation en nous confrontant à des phénomènes faisant l'objet de cette thèse (les voisins distributionnels permettent-ils de mettre au jour la cohésion lexicale d'un texte ?) et non à des problématiques plus générales concernant l'analyse distributionnelle (*cf.* point 3).

Dans la section qui suit, nous montrons qu'une annotation des liens contextualisés est effectivement plus fiable qu'un annotation des couples hors-contexte.

4.2.2 Annoter des liens en contexte : validation de l'approche

Comme nous l'avons mentionné dans la section précédente, nous pensons qu'il est plus adéquat de juger de la pertinence de couples de voisinage en contexte car (a) il s'agit alors d'une tâche plus naturelle (b) cette approche est adaptée aux spécificités d'une ressource distributionnelle et (c) le contexte permet de lever les ambiguïtés.

La réalité des textes offre un cadre, permettant de lever les ambiguïtés liées à la polysémie et de guider l'annotateur. L'exemple (1) présente un cas dans lesquel le contexte influence fortement le jugement porté sur un couple de voisins : le contexte instaure ici une relation – qui est loin d'aller de soi « en langue » – entre les mots "insecte" et "racine", qui désignent tous deux un aliment du hérisson.

(1) Bien que faisant partie des insectivores, les hérissons sont quasiment omnivores. Ils se nourrissent d'**insectes**, [...] de **racines**, de melons et de courges. (Wikipédia, article « Hérisson »)

Au vu de ce premier exemple, on pourrait objecter qu'il sera toujours possible d'interpréter le lien entre deux mots présentés dans un contexte réduit comme une relation sémantique pertinente. L'exemple (2) présente des liens qui selon nous devraient être rejetés. Dans cet exemple, « espace » est voisin avec les mots « animaux » et « majorité » ; toutefois, il nous paraît extrêmement difficile d'établir une relation sémantique pour ces deux paires. Nous verrons dans ce qui suit que plusieurs annotateurs s'accordent effectivement à rejeter certaines paires.

(2) Les impalas sont des **animaux** diurnes ; ils passent donc la **majorité** de la nuit à se reposer et à ruminer et se déplacent le jour afin de trouver de nouveaux **espaces** nourriciers. (Wikipédia, article « Impala »)

Nous présentons dans cette section un pré-test, une première expérience d'annotation à petite échelle, ayant pour objectif de montrer qu'il est effectivement plus pertinent d'annoter des liens en contexte que hors-contexte. Pour atteindre cet objectif, notre stratégie consiste à mettre en place deux annotations de couples extraits

des *Voisins de Wikipédia* ; lors de la première annotation, les couples sont présentés en dehors de tout contexte ; lors de la seconde, les couples sont présentés à l'intérieur d'un paragraphe au sein duquel ils cooccurrent. Ces deux phases d'annotation sont effectuées par trois mêmes annotateurs, ce qui nous permet alors de comparer les accords inter-annotateurs obtenus pour l'annotation dite *hors-contexte* et l'annotation dite *en contexte*. Les trois annotateurs sont des linguistes ; deux d'entre eux (par la suite « Ann1 » et « Ann3 ») connaissent la ressource évaluée et son mode de construction ; le dernier (« Ann2 ») n'est pas expert du domaine. Nous détaillons ci-dessous la mise en place de cette expérience.

4.2.2.1 Sélection des couples à annoter

Pour chaque annotation, 100 couples ont été sélectionnés. Les contraintes posées pour la sélection de ces 100 couples étaient les suivantes :

- pour l'annotation *hors-contexte*, les couples candidats devaient avoir un score de Lin supérieur à $0,2$; $14,1\%$ des couples de voisins répondent à cette contrainte ;
- pour l'annotation *en contexte*, la seule contrainte était que les couples cooccurrent au moins une fois dans un même paragraphe du corpus ayant servi à construire la base de voisins, afin de pouvoir les présenter au sein de ce paragraphe à l'annotateur

Dans un cas comme dans l'autre, les 100 couples finalement présentés aux annotateurs ont été piochés aléatoirement parmi tous les candidats possibles.

4.2.2.2 Consigne aux annotateurs

Afin de ne pas biaiser l'expérience, la consigne donnée était la même dans les deux cas : « Les deux mots proposés présentent-ils un lien de proximité sémantique ? En d'autres termes, existe-t-il une relation sémantique entre eux, qu'elle soit classique (synonymie, antonymie, hyperonymie, co-hyponymie, méronymie, co-méronymie) ou non-classique (la relation peut être glosée mais n'appartient pas aux relations précédemment citées) ? »

Cette consigne est volontairement peu contraignante. Poser une définition trop précise de ce qui constitue un « lien pertinent » (alors même que cette question est loin d'être évidente comme nous l'avons souligné en 2.3.3) aurait cantonné les annotateurs à cette définition, et n'aurait pas permis d'enrichir la réflexion sur ce problème difficile. À l'inverse, en laissant plus de latitude aux annotateurs, nous pouvons :

- vérifier la « consistance » de la notion de pertinence d'un lien de cohésion lexicale : un bon accord inter-annotateurs malgré une consigne peu astreignante est en effet un signe certain de la réalité du phénomène annoté ;
- affiner notre réflexion sur cette notion en nous appuyant sur des données réelles annotées.

4.2.2.3 Résultats

Les tableaux 4.1 et 4.2 présentent les matrices de confusions obtenues par chaque paire d'annotateurs pour les annotations *hors-contexte* (4.1) et *en contexte* (4.2). La classe "1" correspond aux couples jugés pertinents, et la classe "0" aux couples jugés non-pertinents.

		Ann2			Ann3					Ann3		
		1	0	TOT.	1	0	TOT.			1	0	TOT.
Ann1	1	28	16	44	20	24	44	Ann2	1	20	15	35
	0	7	49	56	6	50	56		0	6	59	65
	TOT.	35	65	100	26	74	100		TOT.	26	74	100

TABLEAU 4.1 – Matrices de confusion des annotations hors contexte

		Ann2			Ann3					Ann3		
		1	0	TOT.	1	0	TOT.			1	0	TOT.
Ann1	1	80	7	87	81	6	87	Ann2	1	78	4	82
	0	2	11	13	2	11	13		0	5	13	18
	TOT.	82	18	100	83	17	100		TOT.	83	17	100

TABLEAU 4.2 – Matrices de confusion des annotations en contexte

On constate assez rapidement, un déséquilibre (non contrôlable puisqu'il ne pouvait émerger qu'après annotation) entre les deux ensembles de données à annoter :

- pour les données présentées hors-contexte, la classe majoritaire est, selon tous les annotateurs, la classe "0" (liens non-pertinents) ;
- à l'inverse, les liens présentés en contexte sont jugés très majoritairement pertinents (classe "1").

Ce déséquilibre ne nuit pas à la comparaison entre les deux annotations, dans la mesure où l'on ne cherche pas à interpréter les taux d'accord. Dans le tableau 4.3, nous fournissons en effet, en plus des taux d'accords, les scores obtenus en appliquant le coefficient Kappa de Cohen (1960), qui tient compte de l'accord aléatoire et permet donc la comparaison.

Le coefficient Kappa est défini comme le rapport :

$$\kappa = \frac{p_0 - p_c}{1 - p_c} \tag{4.1}$$

ou p_0 est le taux d'accord entre annotateur et p_c est le taux d'accord attendu par le seul effet du hasard. κ mesure ainsi la proportion d'accord après avoir soustrait l'accord attendu par l'effet du hasard Cohen (1960). Un coefficient $\kappa < 0$ peut ainsi être interprété comme une indication que l'accord entre les annotateurs est plus

Annotateurs	Hors-contexte		En contexte	
	taux d'accord	kappa	taux d'accord	kappa
ann1+ann2	77%	0,52	91%	0,66
ann1+ann3	70%	0,36	92%	0,69
ann2+ann3	79%	0,50	92%	0,69
Moyenne	75,3%	0,46	91,7%	0,68

TABLEAU 4.3 – Accords inter-annotateurs selon le coef. Kappa, en contexte *vs* hors-contexte

mauvais que ce qui aurait été attendu de deux décisions aléatoires. Dans le cas plus intéressant en pratique ou $\kappa > 0$, l'accord entre les annotateurs est meilleur que ce qui aurait été attendu du hasard. κ est borné par la valeur +1.00 en cas d'accord parfait entre les annotateurs. Il est difficile de comparer des scores κ pour des expériences différentes, car les valeurs obtenues dépendront du nombre de catégories. Néanmoins, Landis (1977) ont proposé la gradation reportée dans le tableau 4.4.

κ	Accord
< 0.00	Désaccord
0.00 – 0.20	Marginal
0.21 – 0.40	Modeste
0.41 – 0.60	Modéré
0.61 – 0.80	Fort
0.81 – 1.00	Presque parfait

TABLEAU 4.4 – Échelle d'interprétation du Kappa, (Landis, 1977)

4.2.2.4 Discussion

Notre hypothèse selon laquelle l'annotation de liens en contexte est une tâche plus naturelle et donnant prise à moins d'ambiguïtés est largement confirmée. La comparaison des coefficients kappa fait ressortir une différence moyenne de 0,22 points. Selon l'échelle d'interprétation du kappa proposée par Landis (1977), l'accord *hors-contexte* est *faible* à *modéré*, tandis que le score obtenu *en contexte* correspond à un accord *fort*. On peut également remarquer (bien que cette considération soit à relativiser étant donné le faible nombre d'annotateurs) que la variation d'un couple d'annotateur à l'autre est élevée *hors-contexte*, alors qu'elle est très faible *en contexte*. Cela nous permet d'envisager l'utilisation d'annotations posées en contexte comme référence pour la paramétrisation et l'évaluation d'une ressource distributionnelle. De plus, la qualité de l'accord observé pour l'annotation *en contexte* est pour nous extrêmement encourageante. En effet, comme nous l'avons précédemment

mentionné, un bon accord prouve pour nous la validité de la tâche demandée aux annotateurs.

Il est toutefois important de noter que les jugements posés sur des liens contextualisés ou sur la relation en langue de couples de voisins constituent deux décisions de natures différentes. La pertinence d'un couple de voisins ne peut pas être induite de la pertinence d'une de ses réalisations dans un contexte donné. Ainsi, nous avons mis en évidence que l'annotation de liens en contexte est plus fiable que l'annotation de couples hors-contexte, mais il faut garder à l'esprit que ces deux types d'annotations portent sur des objets différents. La validité des annotations posées en contexte ne peut pas être étendue au-delà de ce contexte.

Une dernière observation, périphérique à l'expérience menée mais essentielle pour la suite de ce chapitre, peut être faite : L'évaluation *hors-contexte* ne fait émerger que 36% de couples de voisins pertinents, malgré une contrainte assez forte sur le score de Lin (seuls $14,1\%$ des voisins dépassaient le seuil posé). Par contre, lors de l'annotation en contexte, alors qu'aucun seuil n'était fixé sur le score de Lin, $82,3\%$ des liens de voisinage sont jugés pertinents en moyenne. Cela signifie que la cooccurrence au sein d'un même paragraphe, qui pourrait paraître être une contrainte assez faible, pourrait constituer un indice plus important de la pertinence d'un couple de voisins qu'un score de Lin élevé. Cela tend à confirmer notre idée selon laquelle le retour aux textes pourrait présenter d'autres avantages qu'une annotation plus sûre : des indices peuvent émerger des textes pour aider à distinguer automatiquement entre couples de voisins pertinents ou non-pertinents. Cette intuition, émise sur la base de deux annotations préliminaires ayant porté sur des objets différents (couples de voisins / liens de voisinage), demande bien sûr à être vérifiée sur la base d'un même dispositif d'annotation. Cela sera développé dans la suite de ce chapitre.

4.2.3 Déroulement de l'annotation

4.2.3.1 Décisions prises sur l'annotation

Quels objets annote-t-on ? Nous avons insisté sur l'importance du contexte dans la décision de l'annotateur, en soulignant qu'un même couple de voisins peut permettre de mettre au jour un lien lexical pertinent dans certains contextes, et non dans d'autres. Il est ainsi difficilement envisageable de proposer d'étendre la validité d'une annotation posée dans un contexte donné à l'ensemble des réalisations possibles d'un couple de voisins.

Faut-il dès lors faire appel au jugement d'un annotateur pour chaque lien de voisinage au sein d'un texte ? Cette solution reviendrait à demander 4 jugements différents sur la paire `atteindre`/`mètre` dans l'exemple (3), que nous reprenons ci-dessous (liens particuliers « atteignent » / « mètre », « atteignent » / « mètres », « atteindre » / « mètre » et « atteindre » / « mètres »). On le voit, cette solution extrêmement fastidieuse peut difficilement être mise en pratique.

(3) Redressés, les gorilles atteignent une taille de 1,75 mètre, mais ils sont en fait un peu plus grands car ils ont les genoux fléchis. L'envergure des bras dépasse la longueur du corpus et peut atteindre 2,75 mètres. (Wikipédia, article « Gorille »)

Nous avons opté pour un compromis entre l'annotation de couples de voisins (c'est-à-dire de la relation en langue) et de liens de voisinage (c'est-à-dire de chaque occurrence particulière d'un couple de voisins) : nous avons fait le choix d'annoter chaque couple une seule fois pour un texte donné, sur la base de l'ensemble de ses occurrences appréhendées simultanément. Nous suivons en cela un postulat bien connu en désambiguïsation lexicale, selon lequel toutes les occurrences d'un mot dans un même texte correspondent au même sens (« One sense per discourse », Gale *et al.*, 1992). Nous supposons que, de la même manière que le sens des mots pris individuellement, la nature de la relation qui unit ces mots reste stable dans un texte donné. Il s'agit ici d'une hypothèse assez forte. Pour s'assurer de sa plausibilité, il faut être attentif à l'accord inter-annotateurss émergeant de ce dispositif d'annotation, qui doit rester comparable à celui du pré-test (section 4.2.2).

Ainsi, le jugement de l'annotateur ne serait sollicité qu'une seule fois pour la paire `atteindre`/`mètre` dans le texte ci-dessus. Par contre, un même couple de voisins peut être annoté plusieurs fois s'il apparaît dans deux textes différents, comme c'est le cas ci-dessous avec les exemples (4) (précédemment commenté dans la section 4.2.2) et (5). Ici, nous pensons que le couple `insecte`/`racine` est pertinent dans le texte « Hérisson » (car insecte et racine font tous deux partie des aliments du hérisson), mais plus difficilement dans le texte « Annélation ».

(4) Bien que faisant partie des insectivores, les hérissons sont quasiment omnivores. Ils se nourrissent d'**insectes**, [...]de **racines**, de melons et de courges. (Wikipédia, article « Hérisson »)

(5) Le mot annélation peut aussi décrire l'écorçage total d'un arbre, d'une branche, d'une **racine** ou d'une tige par un animal (**insecte**, rongeur, grand ou petit herbivore). (Wikipédia, article « Annélation »)

Quel type de jugement doit poser l'annotateur ? De la même manière que dans le pré-test, nous avons choisi de demander à l'annotateur un jugement binaire. Ce jugement, comme nous l'avons signalé, peut être porté en même temps sur plusieurs liens dans le cas où les deux voisins considérés ont plusieurs occurrences dans le texte présenté à l'annotateur. Il pourrait être glosé de la manière suivante :

Les réalisations du couple $\texttt{voisin}_a/\texttt{voisin}_b$, dans le texte T, sont-elles pertinentes, c'est-à-dire participent-elles à la cohésion lexicale de T ?

Que montre-t-on à l'annotateur ? Pour cette phase d'annotation, nous avons choisi de présenter des textes entiers aux annotateurs, pour les raisons suivantes :

– Dans la mesure où le jugement de l'annotateur doit être valide pour la totalité d'un texte donné, il était logique qu'il puisse visualiser ce texte dans son ensemble. Toutes les occurrences des deux voisins dont la relation est évaluée sont mis en surbrillance.
– Même lorsqu'un couple n'a qu'une seule réalisation, il peut s'agir d'un lien connectant deux items très distants, d'où la nécessité d'afficher l'ensemble du texte.

4.2.3.2 Corpus

Pour cette phase d'annotation, nous avons constitué un corpus de 42 textes sélectionnés de manière arbitraire dans WikipédiaFR2007. Nous avons projeté les voisins sur ce corpus en suivant la procédure décrite en 3.2. Aucun filtrage des voisins (par exemple par seuillage sur le score de Lin) n'a été appliqué, dans la mesure où nous souhaitons nous baser sur l'annotation effectuée pour proposer une ou des stratégie(s) de filtrage des voisins. Le tableau 4.5 indique les caractéristiques essentielles du corpus.

Nombre de textes	94307
Nombre de mots	15527
Nombre de liens	3476785
Nombre de couples	801196

TABLEAU 4.5 – Caractéristiques du corpus d'annotation

Le nombre de couples de voisins indiqué correspond non pas au nombre de couples différents dans tout le corpus, mais à l'addition du nombre de couples de chaque texte ; ce sont donc bien ces objets que nous souhaitons annoter, comme nous l'avons développé dans la section précédente. Le but de l'annotation n'est évidemment pas de fournir un jugement sur tous ces couples, et la taille importante du corpus constitué s'explique par le souci d'assurer une variété suffisante des annotations.

Dans la mesure où nous souhaitons nous appuyer sur les résultats de cette annotation pour modifier des paramètres de notre chaîne de traitement, le sous-corpus constitué lors de cette phase d'annotation ne sera plus utilisé pour les expériences relatées dans la suite de cette thèse.

4.2.3.3 Interface développée et modalités d'annotation

Nous avons nous-même développé l'interface d'annotation. Les contraintes que nous souhaitions respecter étaient principalement les suivantes :

– les liens annotés doivent être représentatifs des liens projetés, et couvrir tous les textes du corpus ;
– l'annotation doit être rapide ;
– l'annotateur doit pouvoir s'arrêter à tout moment.

Ces contraintes ont conduit à différents choix lors de la réalisation de l'interface d'annotation.

L'interface a été développée en PhP/MySQL et JavaScript.

Quand un utilisateur accède à l'interface, un texte choisi aléatoirement parmi les 42 textes du corpus est affiché. Dans ce texte, un voisin (qu'on appellera mot-cible) est choisi au hasard, et tous ses voisins apparaissent surlignés dans le texte. Le rôle de l'annotateur est alors de juger de la pertinence de tous les liens dans lesquels est impliqué le mot-cible. Nous avons en effet déterminé empiriquement qu'il était plus efficace de procéder ainsi que d'afficher un nouveau couple après chaque jugement de l'annotateur. La figure 4.1 p. 106 permet de visualiser un texte tel qu'il est présenté à l'annotateur. Dans cette figure, toutes les occurrences du mot-cible, `fonctionnaire`, apparaissent surlignées en bleu, et toutes les occurrences de tous ses voisins apparaissent surlignées en jaune.

L'annotateur peut à tout moment :

- changer de mot-cible ;
- changer de texte ;
- interrompre l'annotation.

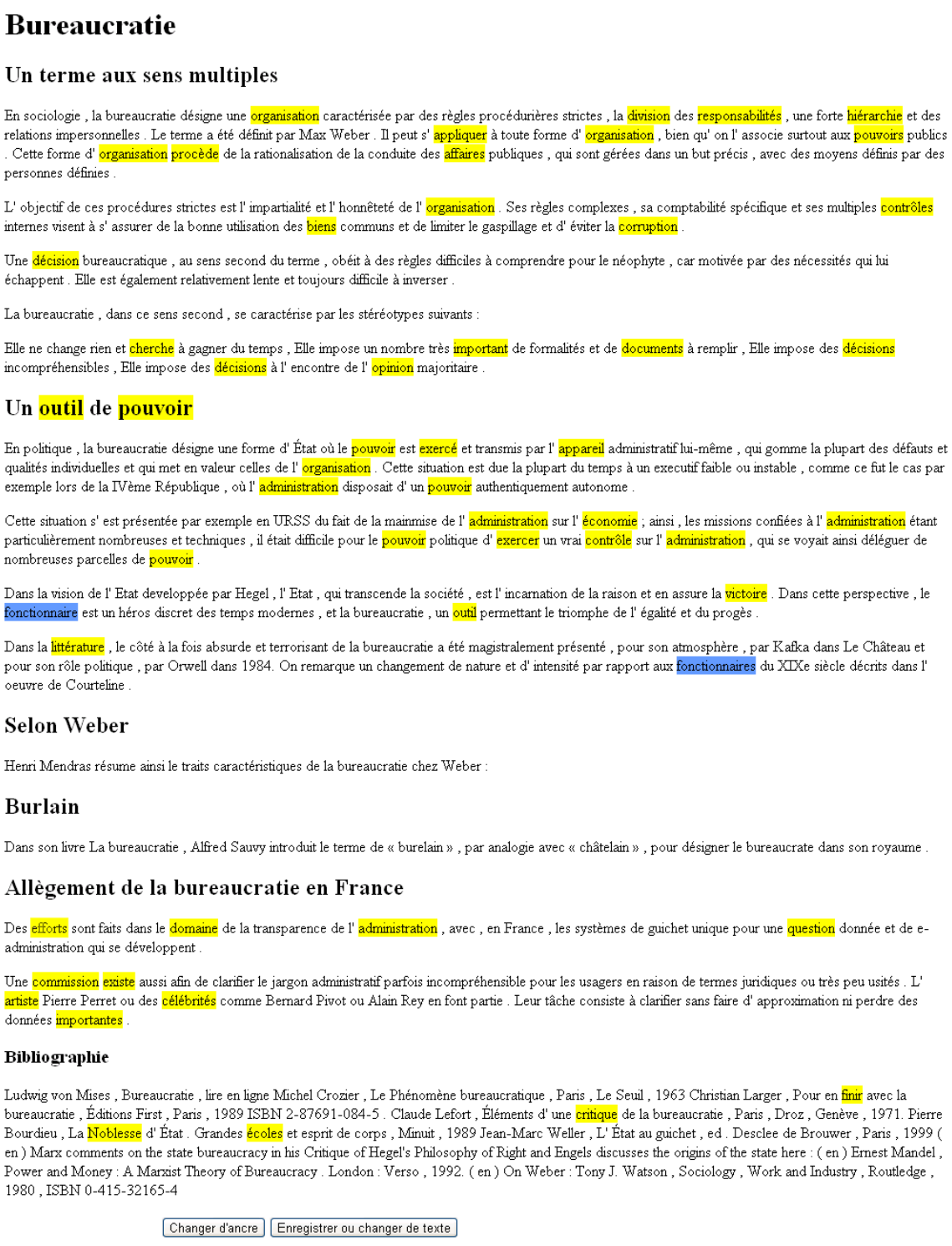
Afin d'améliorer l'efficacité de l'annotation, nous avons veillé à limiter au maximum les rafraichissements de la page. Tant que l'utilisateur reste sur un même texte, la page n'est jamais rafraichie (pas même lorsque le mot-cible change). Les annotations sont mémorisées au fur et à mesure, et la communication avec le serveur se fait uniquement lorsque l'annotateur choisit de s'arrêter ou de changer de texte. C'est à ce moment-là que les annotations sont envoyées à la base de données MySQL.

L'annotation se fait uniquement en cliquant sur les voisins du mot-cible : au premier clic, le voisin apparaît en rouge (pour signifier qu'il est jugé non-pertinent), et au second clic, il apparaît en vert (pour signifier qu'il est jugé pertinent). Le code couleur utilisé est rappelé dans le tableau 4.6. Dès qu'un mot est cliqué, toutes les occurrences du même mot prennent la même couleur. Si un couple a déjà été annoté lors d'une phase précédente, il apparaît tout de même, mais déjà annoté (en vert ou rouge).

bleu	mot-cible
jaune	voisin non-annoté
rouge	voisin annoté non-pertinent
vert	voisin annoté pertinent

TABLEAU 4.6 – Code couleur

La figure 4.2 montre une annotation en cours sur un extrait du texte « Belfast ». Dans cet extrait, le mot-cible est `hockey`. Les couples `hockey`/`glace` et `hockey`/`sport` ont été jugés pertinents dans ce contexte. Le couple `hockey`/`grand` a été rejeté. Les couples `hockey`/`football`, `hockey`/`championnat` et `hockey`/`rugby` doivent encore être annotés.



Bureaucratie

Un terme aux sens multiples

En sociologie , la bureaucratie désigne une organisation caractérisée par des règles procédurières strictes , la division des responsabilités , une forte hiérarchie et des relations impersonnelles . Le terme a été définit par Max Weber . Il peut s' appliquer à toute forme d' organisation , bien qu' on l' associe surtout aux pouvoirs publics . Cette forme d' organisation procède de la rationalisation de la conduite des affaires publiques , qui sont gérées dans un but précis , avec des moyens définis par des personnes définies .

L' objectif de ces procédures strictes est l' impartialité et l' honnêteté de l' organisation . Ses règles complexes , sa comptabilité spécifique et ses multiples contrôles internes visent à s' assurer de la bonne utilisation des biens communs et de limiter le gaspillage et d' éviter la corruption .

Une décision bureaucratique , au sens second du terme , obéit à des règles difficiles à comprendre pour le néophyte , car motivée par des nécessités qui lui échappent . Elle est également relativement lente et toujours difficile à inverser .

La bureaucratie , dans ce sens second , se caractérise par les stéréotypes suivants :

Elle ne change rien et cherche à gagner du temps , Elle impose un nombre très important de formalités et de documents à remplir , Elle impose des décisions incompréhensibles , Elle impose des décisions à l' encontre de l' opinion majoritaire .

Un outil de pouvoir

En politique , la bureaucratie désigne une forme d' État où le pouvoir est exercé et transmis par l' appareil administratif lui-même , qui gomme la plupart des défauts et qualités individuelles et qui met en valeur celles de l' organisation . Cette situation est due la plupart du temps à un executif faible ou instable , comme ce fut le cas par exemple lors de la IVème République , où l' administration disposait d' un pouvoir authentiquement autonome .

Cette situation s' est présentée par exemple en URSS du fait de la mainmise de l' administration sur l' économie ; ainsi , les missions confiées à l' administration étant particulièrement nombreuses et techniques , il était difficile pour le pouvoir politique d' exercer un vrai contrôle sur l' administration , qui se voyait ainsi déléguer de nombreuses parcelles de pouvoir .

Dans la vision de l' Etat developpée par Hegel , l' Etat , qui transcende la société , est l' incarnation de la raison et en assure la victoire . Dans cette perspective , le fonctionnaire est un héros discret des temps modernes , et la bureaucratie , un outil permettant le triomphe de l' égalité et du progès .

Dans la littérature , le côté à la fois absurde et terrorisant de la bureaucratie a été magistralement présenté , pour son atmosphère , par Kafka dans Le Château et pour son rôle politique , par Orwell dans 1984. On remarque un changement de nature et d' intensité par rapport aux fonctionnaires du XIXe siècle décrits dans l' oeuvre de Courteline .

Selon Weber

Henri Mendras résume ainsi le traits caractéristiques de la bureaucratie chez Weber :

Burlain

Dans son livre La bureaucratie , Alfred Sauvy introduit le terme de « burelain » , par analogie avec « châtelain » , pour désigner le bureaucrate dans son royaume .

Allègement de la bureaucratie en France

Des efforts sont faits dans le domaine de la transparence de l' administration , avec , en France , les systèmes de guichet unique pour une question donnée et de e-administration qui se développent .

Une commission existe aussi afin de clarifier le jargon administratif parfois incompréhensible pour les usagers en raison de termes juridiques ou très peu usités . L' artiste Pierre Perret ou des célébrités comme Bernard Pivot ou Alain Rey en font partie . Leur tâche consiste à clarifier sans faire d' approximation ni perdre des données importantes .

Bibliographie

Ludwig von Mises , Bureaucratie , lire en ligne Michel Crozier , Le Phénomène bureaucratique , Paris , Le Seuil , 1963 Christian Larger , Pour en finir avec la bureaucratie , Éditions First , Paris , 1989 ISBN 2-87691-084-5 . Claude Lefort , Éléments d' une critique de la bureaucratie , Paris , Droz , Genève , 1971. Pierre Bourdieu , La Noblesse d' État . Grandes écoles et esprit de corps , Minuit , 1989 Jean-Marc Weller , L' État au guichet , ed . Desclee de Brouwer , Paris , 1999 (en) Marx comments on the state bureaucracy in his Critique of Hegel's Philosophy of Right and Engels discusses the origins of the state here : (en) Ernest Mandel , Power and Money : A Marxist Theory of Bureaucracy . London : Verso , 1992. (en) On Weber : Tony J. Watson , Sociology , Work and Industry , Routledge , 1980 , ISBN 0-415-32165-4

Changer d'ancre | Enregistrer ou changer de texte

FIGURE 4.1 – Interface d'annotation : exemple avec le mot-cible `fonctionnaire`, dans le texte « Bureaucratie »

FIGURE 4.2 – Exemple de texte à annoter

Dans la mesure où cette interface était destinée à un nombre restreint d'annotateurs, nous n'avons pas spécifiquement testé son ergonomie, mais il ne nous a pas été fait part de difficultés de la part des personnes l'ayant utilisée.

4.2.4 Bilan de l'annotation

Deux annotateurs experts (dont nous-même) ont participé à l'annotation. Leur accord a été évalué sur 120 couples répartis dans les différents textes du corpus. Le tableau 4.7 donne la matrice de confusion des deux annotateurs sur ces 120 couples. Le tableau 4.8 indique les taux d'accord et coefficient Kappa correspondants.

	1	**0**	**Tot.**
1	35	6	41
0	5	74	79
Tot.	40	80	120

TABLEAU 4.7 – Matrice de confusion

Taux d'accord	Coefficient Kappa
90.8%	0.80

TABLEAU 4.8 – Taux d'accord et coefficient Kappa

Le coefficient Kappa de 0.80 indique un accord inter-annotateurs encore meilleur que lors du pré-test. Ce meilleur score peut notamment s'expliquer par :

- la quantité supérieure d'informations présentées à l'annotateur : un contexte plus large (tout le texte) et plus d'occurrences des deux voisins impliqués dans le couple à annoter ;
- le plus grand entrainement des annotateurs, qui sont les mêmes que lors du pré-test.

Ce très bon accord inter-annotateurs montre que le compromis décrit en 4.2.3.1 (sur la nature des objets annotés) n'a pas affecté la qualité des jugements posés par les annotateurs. La présentation simultanée de (parfois) plusieurs occurrences, pour un même couple de voisins, a été ressentie par les annotateurs non pas comme une source de confusion, mais comme une indication supplémentaire.

Au total, 9885 couples ont été annotés. Environ 11% de ces couples participent, selon les annotateurs, à la cohésion lexicale des textes dans lesquels ils apparaissent (*cf.* figure 4.3). Cette proportion peut paraître faible de prime abord ; mais il faut rappeler qu'elle est le résultat de la projection de tous les couples rapprochés par l'analyse distributionnelle, sans aucun seuillage de leur score de Lin. Nous avons eu l'occasion de montrer l'énorme quantité de liens projetés en procédant de cette manière (section 3.3.2 page 79), et donc la très forte couverture atteinte. Le fait que plus de 10% des liens projetés soient pertinents avant toute tentative de filtrage est selon nous plutôt encourageant.

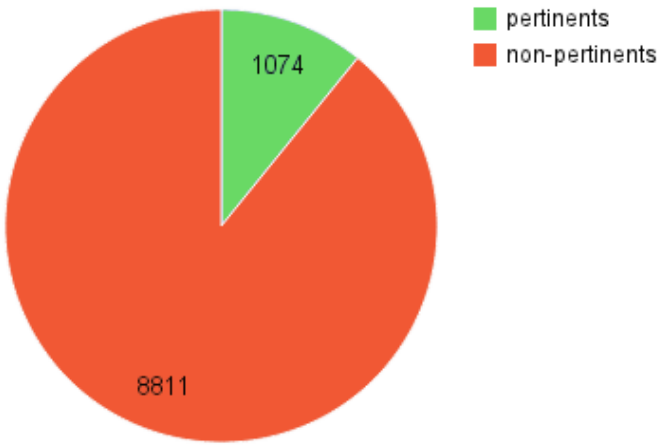


FIGURE 4.3 – Jugements des annotateurs sur les couples de voisins annotés

Au terme de cette annotation, nous avons constitué une référence pour l'évaluation de la projection des voisins en texte. Nous nous sommes également confrontée à grande échelle aux liens de voisinage que le dispositif que nous avons mis en place (*cf.* chapitre 3) permet de projeter dans les textes, et à la question du rapport entre

ces liens et la cohésion lexicale des textes. Dans la suite du chapitre, nous montrons comment les données annotées peuvent être exploitées pour proposer des solutions pour le filtrage des liens projetés.

4.3 Quels indices pour prédire la pertinence d'un lien de voisinage ?

Au terme de l'annotation de liens en contexte décrite dans la section précédente, nous disposons d'une masse relativement importante de données constituées de couples de voisins, et de jugements linguistiques posés sur les réalisations de ces couples dans des textes particuliers. Ces jugements sont assez fiables si l'on se réfère à l'accord inter-annotateurss ; on peut donc envisager d'exploiter ces données pour faire émerger des corrélations entre certaines caractéristiques des couples de voisins, et leur pertinence selon les annotateurs.

Dans cette section, nous listons les caractéristiques que nous avons souhaité confronter aux données annotées. Nous parlons d'indices pour désigner ces caractéristiques, car le but à terme est de s'appuyer sur les caractéristiques les plus saillantes pour prédire la pertinence des couples projetés dans un texte donné, ce qui permettra un filtrage de ces couples. Nous avons classé les indices définis selon leur origine : sont-ils calculés à partir du corpus ? Résultent-ils de la construction de la base distributionnelle ? Ou enfin, émergent-ils après la projection des voisins distributionnels dans le texte ? La figure 4.4 reprend la figure 3.1 (page 72) en montrant où intervient l'extraction de ces différentes catégories d'indices.

Nous avons insisté (section 4.1) sur le fait que nous options pour un filtrage « en aval » des liens projetés. Parmi les indices considérés, certains proviennent d'étapes antérieures à la projection : ceux qui émanent du corpus ou de la base distibutionnelle. Il s'agit là d'indices qu'il est logique d'utiliser dans le contexte de l'exploitation d'une ressource distributionnelle ; pour certains d'entre eux, leur impact a déjà été étudié (*cf.* section 2.1.2.1 page 54). Par contre, l'utilisation d'indices émergeant du texte après la projection des voisins distributionnels est une spécificité de notre approche, rendue possible par le dispositif d'annotation que nous avons choisi d'adopter.

Certains des indices proposés sont le résultat d'une intuition née de l'observation des liens projetés lors de la phase d'annotation ; d'autres indices ont été pris en compte sans être appuyés par une intuition linguistique, mais simplement parce que nous souhaitions explorer une large gamme de caractéristiques. Dans la suite de cette section, nous précisons notre intuition sur les indices proposés lorsque nous en avons une.

Notons également que certains indices sont plus facilement interprétables linguistiquement, car ils cherchent à rendre compte de ce qu'est un lien de cohésion lexicale (par exemple, la cohésion lexicale est peut-être davantage portée par les noms, il y a peut-être plus de liens de cohésion lexicale à proximité qu'à distance,

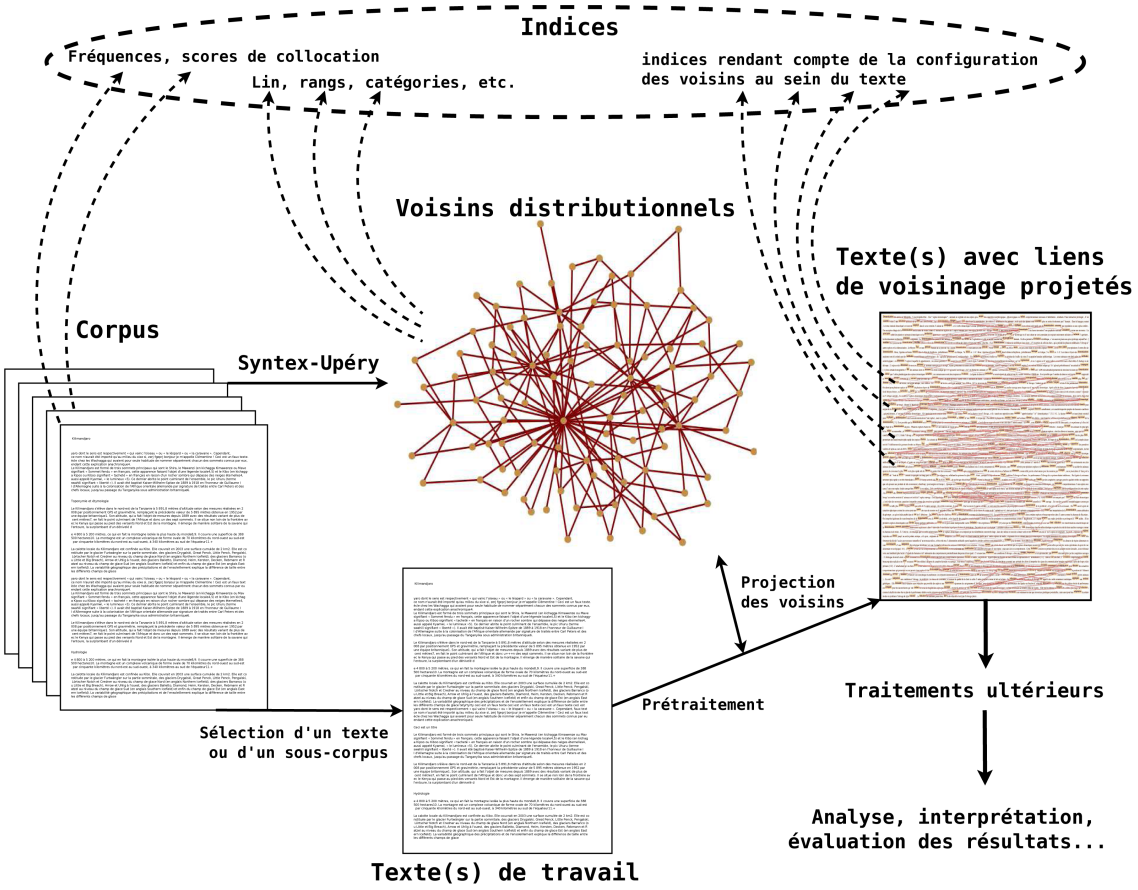


FIGURE 4.4 – Extraction d'indices à différentes étapes de la projection

etc.). D'autres indices cherchent plutôt à évaluer ce qu'est un bon couple de voisins, en faisant intervenir des caractéristiques plus abstraites telles que la productivité des voisins.

4.3.1 Indices émanant du corpus

Pour un couple $\texttt{voisin}_a/\texttt{voisin}_b$ donné, nous avons calculé, à partir du corpus WikipédiaFR2007, deux informations principales.

La première concerne les fréquences de deux voisins dans le corpus. Nous appelons $freq_a$ la fréquence de $\texttt{voisin}_a$ et $freq_b$ la fréquence de $\texttt{voisin}_b$. À partir de ces deux fréquences, trois indices sont définis :

- $freq_{min}$ a pour valeur la plus faible fréquence parmi $freq_a$ et $freq_b$;
- $freq_{max}$ a pour valeur la plus forte fréquence parmi $freq_a$ et $freq_b$;
- $freq_{\times}$ combine les deux fréquences en un seul attribut ; il est défini comme le *log* du produit de $freq_a$ et $freq_b$

Notre intuition est que les mots très fréquents se caractériseront souvent par des

rapprochements distributionnels moins pertinents pour la détection de la cohésion lexicale. Nous suivons en cela une hypothèse déjà exprimée par Halliday et Hasan (1976) :

> There is a third factor influencing the cohesive force between a pair of lexical items in a text, and that is their overall frequency in the system of the language. [...] Words such as *go* or *man* or *know* or *way* can hardly be said to contract significant cohesive relations, because they go with anything at all. Since roughly speaking, words of this kind are also those with high overall frequency in the language, 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)

La seconde information concerne l'association syntagmatique de $\texttt{voisin}_a$ et $\texttt{voisin}_b$. Nous avons vu (section 2.3.2 page 63) que bien que les voisins distributionnels soient rapprochés sur des critères paradigmatiques, ils présentent souvent également une certaine affinité d'ordre syntagmatique. Cette affinité syntagmatique est bien sûre centrale dans le cadre du projet VOILADIS. L'indice que nous utilisons pour en rendre compte est une information mutuelle (im) (Church et Hanks, 1990, p. 23) basée sur la cooccurrence des deux membres du couple dans une fenêtre d'un paragraphe au sein du corpus WikipédiaFR2007. L'information mutuelle met en rapport la probabilité de trouver les deux voisins dans le même paragraphe ($P(a,b)$) et le produit des probabilités indépendantes de trouver chaque voisin dans un paragraphe donné ($P(a) \cdot P(b)$) :

$$IM(a,b) = \log \frac{P(a,b)}{P(a) \cdot P(b)} \tag{4.2}$$

On peut noter que le critère du paragraphe est une contrainte relativement faible ; on espère ainsi une meilleure couverture de l'indice (qui avec cette contrainte peut être calculé pour 70% des couples, *cf.* section 2.3.2). On a par ailleurs constaté lors du pré-test que cette contrainte de cooccurrence au sein d'un même paragraphe constitue un très bon indice de la pertinence d'un couple de voisins (*cf.* section 4.2.2.4).

Ces indices sont résumés dans le tableau 4.9.

Indice	Description	Valeurs possibles
$freq_{\min}$	$\min(freq_a, freq_b)$	$freq_{\min} \in \mathbb{N}^*$
$freq_{\max}$	$\max(freq_a, freq_b)$	$freq_{\max} \in \mathbb{N}^*$
$freq_{\times}$	$\log(freq_a \times freq_b)$	$freq_{\times} \in \mathbb{N}^*$
im	$im = \log \frac{P(a,b)}{P(a) \cdot P(b)}$	$im \in \mathbb{R}$

TABLEAU 4.9 – Indices émanant du corpus

4.3.2 Indices émanant de la base distributionnelle

Utiliser pour filtrer les voisins des informations résultant de leur construction est la solution la plus évidente. Plus particulièrement, c'est précisément la vocation du score de Lin que d'évaluer la qualité d'un rapprochement distributionnel. On a toutefois eu l'occasion de constater la possible insuffisance de ce score : en effet, selon l'annotation menée lors du pré-test (section 4.2.2), un seuil à 0.2 sur le score de Lin rejette plus de 85% des paires de voisins mais ne fait émerger que 36% de couples de voisins pertinents. Nous pensons que de meilleures stratégies de filtrage sont possibles. Dès lors que l'on considère que le score "final" de proximité n'est pas à lui seul suffisant, toutes les informations sur les couples de voisins peuvent alors servir d'indices, et c'est la confrontation avec les données annotées qui devra faire émerger quels indices (ou configurations d'indices) sont les plus saillants.

Outre le score de Lin (*lin*), nous avons donc défini d'autres indices « quantitatifs » qui émergent de la base de voisins distributionnels :

- la productivité des voisins : $prod_a$ est défini comme le nombre de voisins de $\texttt{voisin}_a$ dans les *Voisins de Wikipédia*, et $prod_b$ comme le nombre de voisins de $\texttt{voisin}_b$ dans cette même base. Tout comme la fréquence dans la section précédente, la productivité des voisins est déclinée en trois indices : $prod_{min}$, $prod_{max}$ et $prod_{\times}$. Nous pensons que les voisins trop productifs donnent lieu à beaucoup de bruit.
- les rangs des voisins (qui découlent directement du *lin*) : $rang_{a-b}$ est défini comme le rang de $\texttt{voisin}_a$ parmi tous les voisins de $\texttt{voisin}_b$ triés par score de Lin ; $rang_{b-a}$ est défini réciproquement. Là encore, trois indices sont considérés : $rang_{min}$, $rang_{max}$ et $rang_{\times}$.

Nous avons également défini deux indices catégoriels plus « linguistiques » :

- le premier (*cats*) concerne les catégories morpho-syntaxiques des deux voisins considérés : s'agit-il de noms, de verbes, d'adjectifs ? Lors de la phase d'annotation, il nous a semblé que certaines catégories donnaient plus souvent lieu à des liens pertinents que d'autres : par exemple, les couples NN (nom-nom) nous paraissaient plus souvent pertinents que les couples VV (verbe-verbe) ; nous avons souhaité vérifier cette intuition.
- le second (*predarg*) concerne la distinction prédicat / argument, une des spécificités des *Voisins de Wikipédia* : le couple de voisins considéré résulte-t-il d'un rapprochement entre prédicats, ou entre arguments ? Il s'agit là d'une information disponible dans la base de voisins, que nous ajoutons donc à la liste des indices, mais sur laquelle nous n'avons aucune intution *a priori*.

Le tableau 4.10 résume l'ensemble des indices « distributionnels » utilisés.

Ainsi, nous envisageons plusieurs indices liés à la base distributionnelle et facilement accessible. Nous ne faisons par contre pas varier de paramètres régissant la construction de la base pour examiner leur impact.

Indice	Description	Valeurs possibles
lin	*cf.* section 2.2.3	$0 \leq lin \leq 1$
$rang_{\min}$	$\min(rang_{a-b}, rang_{b-a})$	$rang_{\min} \in \mathbb{N}^*$
$rang_{\max}$	$\max(rang_{a-b}, rang_{b-a})$	$rang_{\max} \in \mathbb{N}^*$
$rang_{\times}$	$\log(rang_{a-b} \times rang_{b-a})$	$rang_{\times} \in \mathbb{N}^*$
$prod_{\min}$	$\min(prod_a, prod_b)$	$prod_{\min} \in \mathbb{N}^*$
$prod_{\max}$	$\max(prod_a, prod_b)$	$prod_{\max} \in \mathbb{N}^*$
$prod_{\times}$	$\log(prod_a \times prod_b)$	$prod_{\times} \in \mathbb{N}^*$
$cats$	catégories des voisins	AA,AN,AV,NN,NV,VV
$predarg$	prédicats ou arguments ?	pred,arg

TABLEAU 4.10 – Indices émanant de la base distributionnelle

4.3.3 Indices émanant du texte après projection

Si l'on considère qu'un couple de voisin peut donner lieu à des réalisations pertinentes (dans certains textes) et à d'autres qui ne le soient pas (dans d'autres textes), il est dès lors nécessaire de développer des indices dépendant du texte pour filtrer ces couples. Nous pensons que de tels indices peuvent compléter efficacement les indices basés sur les propriétés distributionnelles des paires de voisins.

Les indices que nous avons définis visent à appréhender l' « importance » des items considérés au sein du texte, leurs configurations (plus particulièrement *via* la notion de distance entre les deux voisins), ou encore certaines propriétés du sous-graphe des voisins impliqués dans le texte.

Des indices mesurant l'importance d'un lemme au sein du texte La première information prise en compte concerne les fréquences des items au sein du texte, déclinées en trois indices : $freqtxt_{\min}$, $freqtxt_{\max}$ et $freqtxt_{\times}$.

Les éléments situés en position initiale (c'est-à-dire dans la zone préverbale des phrases) jouent un rôle particulier dans l'organisation discursive (Ho-Dac, 2007). Nous avons défini deux indices permettant d'identifier les couples dont un membre apparaît en position initiale de phrase ($posinit_{ph}$) ou de paragraphe ($posinit_{para}$).

Le *tf·idf* (*term frequency · inverse document frequency* Salton *et al.*, 1975), utilisé notamment en recherche d'information, évalue l'importance d'un mot dans un document en se basant sur sa fréquence dans ce document et sa distribution dans une collection de documents : un mot rare dans la collection de documents mais fréquent dans le document considéré sera doté d'un fort *tf·idf*. Ce score a donné lieu à des adaptations pour identifier des mots importants localement, dans une certaine zone de texte, relativement à la totalité du texte. Ainsi, Dias *et al.* (2007) proposent un *tf·isf* (*term frequency · inverse sentence frequency*), utilisé dans le cadre de la tâche de segmentation thématique. Nous avons calculé pour chaque voisin un score basé sur sa fréquence dans le (ou les) paragraphe(s) où il apparaît et sa distribution dans le texte ; l'indice $tfipf$ est défini comme le produit des scores calculés pour

$\texttt{voisin}_a$ et $\texttt{voisin}_b$.

Des indices mesurant la distance entre les deux voisins Lors de la phase d'annotation, nous avons observé que les voisins apparaissant plus proches l'un de l'autre dans le texte semblaient plus souvent entretenir un lien jugé pertinent que les voisins plus éloignés. Nous avons défini différents indices pour rendre compte de cette intuition :

- des indices comptant cette distance dans le texte en nombre de mots. Dans la mesure où chaque voisin d'un couple peut avoir plusieurs occurrences dans le texte, nous calculons :
  - une distance minimale (*ppd* pour « plus petite distance ») définie comme le nombre de mots séparant les deux occurrences les plus proches de $\texttt{voisin}_a$ et $\texttt{voisin}_b$;
  - une distance maximale (*pgd* pour « plus grande distance ») définie comme le nombre de mots séparant les deux occurrences les plus éloignées de $\texttt{voisin}_a$ et $\texttt{voisin}_b$;
  - une distance moyenne (*md*) ;
- des indices booléens indiquant si oui ou non $\texttt{voisin}_a$ et $\texttt{voisin}_b$ sont, au moins une fois dans le texte, co-présents au sein de la même phrase ($copr_{ph}$) ou du même paragraphe ($copr_{para}$).

Ces évaluations de distance ou de proximité n'épuisent pas la notion de « configuration » des deux voisins ; on aurait par exemple pu imaginer des indices rendant compte de l'ordre d'apparition de $\texttt{voisin}_a$ et $\texttt{voisin}_b$.

Des indices liés au sous-graphe de voisinage associé au texte À partir des sous-graphes des voisins impliqués dans chaque texte, nous relevons les informations suivantes :

- la productivité de chaque voisin dans le texte, c'est-à-dire le nombre de couples différents dont il fait partie pour le texte considéré (indices $prodtxt_{\min}$, $prodtxt_{\max}$ et $prodtxt_{\times}$) ;
- l'appartenance du couple à une composante connexe de taille réduite (indice booléen *cc*), dont la définition et le mode de calcul ont été détaillés dans la section 3.4.1.2 page 89 (on peut également se référer à l'annexe A.2 page 287 pour un exemple). Nous avons observé que les liens extraits de cette manière paraissaient souvent pertinents.

Ces deux informations sont liées : *via* les composantes connexes, nous repérons des groupes de voisins ayant une très faible productivité.

Les indices « textuels » définis sont résumés dans le tableau 4.11.

Indice	Description	Valeurs possibles
$freqtxt_{\min}$	$\min(freqtxt_a, freqtxt_b)$	$freqtxt_{\min} \in \mathbb{N}^*$
$freqtxt_{\max}$	$\max(freqtxt_a, freqtxt_b)$	$freqtxt_{\max} \in \mathbb{N}^*$
$freqtxt_{\times}$	$\log(freqtxt_a \times freqstxt_b)$	$freqstxt \in \mathbb{N}^*$
$tfipf$	$tf{\cdot}ipf(\texttt{voisin}_a) \cdot tf{\cdot}ipf(\texttt{voisin}_b)$	$0 \leq tfipf \leq 1$
$posinit_{sec}$	$\texttt{voisin}_a$ ou $\texttt{voisin}_b$ en pos. init. de section	booléen
$posinit_{para}$	$\texttt{voisin}_a$ ou $\texttt{voisin}_b$ en pos. init. de paragraphe	booléen
$copr_{ph}$	co-présence dans une même phrase	booléen
$copr_{para}$	co-présence dans un même paragraphe	booléen
ppd	plus petite distance entre $\texttt{voisin}_a$ et $\texttt{voisin}_b$	$ppd \in \mathbb{N}^*$
pgd	plus grande distance entre $\texttt{voisin}_a$ et $\texttt{voisin}_b$	$pgd \in \mathbb{N}^*$
md	distance moyenne entre $\texttt{voisin}_a$ et $\texttt{voisin}_b$	$md \in \mathbb{N}^*$
$prodtxt_{\min}$	$\min(prod_a, prod_b)$	$prod_{\min} \in \mathbb{N}^*$
$prodtxt_{\max}$	$\max(prod_a, prod_b)$	$prod_{\max} \in \mathbb{N}^*$
$prodtxt_{\times}$	$\log(prod_a \times prod_b)$	$prods \in \mathbb{N}^*$
cc	*cf.* section 3.4.1.2	booléen

TABLEAU 4.11 – Indices émanant du texte

4.4 Exploration des indices définis

Dans cette section, nous investiguons le lien de chaque indice pris individuellement avec la pertinence des couples annotés. La combinaison de ces indices pour le filtrage des liens projetés sera envisagée dans la section 4.5. Pour certains indices, nous serons amenée à noter qu'ils devraient être affinés, ou qu'un nouvel indice devrait être envisagé. Nous fournissons en effet ici une photographie à un moment donné d'un processus qui pourrait être poursuivi sur plusieurs itérations.

4.4.1 Méthode d'exploration

Dans un premier temps, nous avons calculé pour chaque indice son association avec la pertinence des couples de voisins (parfois notée *pert*), en utilisant la corrélation de Pearson. Pour n mesures de deux variables x et y, le coefficient de corrélation r_{xy} est défini comme :

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{4.3}$$

où $\bar{x}$ et $\bar{y}$ sont les valeurs moyennes respectives de x et y. Dans les résultats présentés ci-dessous, les variables binaires telles que la pertinence ont été considérées comme des variables discrètes, pouvant prendre les valeurs 0 ou 1. Le seul indice pour lequel cette analyse n'a pas pu être menée est le couple de catégories grammaticales des voisins, qui ne se prête pas à une interprétation numérique, puisqu'il faudrait dans ce cas introduire un classement des différentes paires de catégories. La *p-value* est

la probabilité d'obtenir la même valeur de corrélation (ou une valeur encore plus extrême) dans le cas ou la corrélation est en fait nulle. Une *p-value* faible assure ainsi de la significativité des résultats. Concernant la significativité des corrélations calculées, nous utilisons la notation classique résumée dans le tableau 4.12. Les

*	$p\text{-}value < 0.05$
**	$p\text{-}value < 0.01$
***	$p\text{-}value < 0.001$

TABLEAU 4.12 – Notation utilisée pour la significativité des corrélations présentées

corrélations calculées sont résumées par le tableau 4.13 et la figure 4.5.

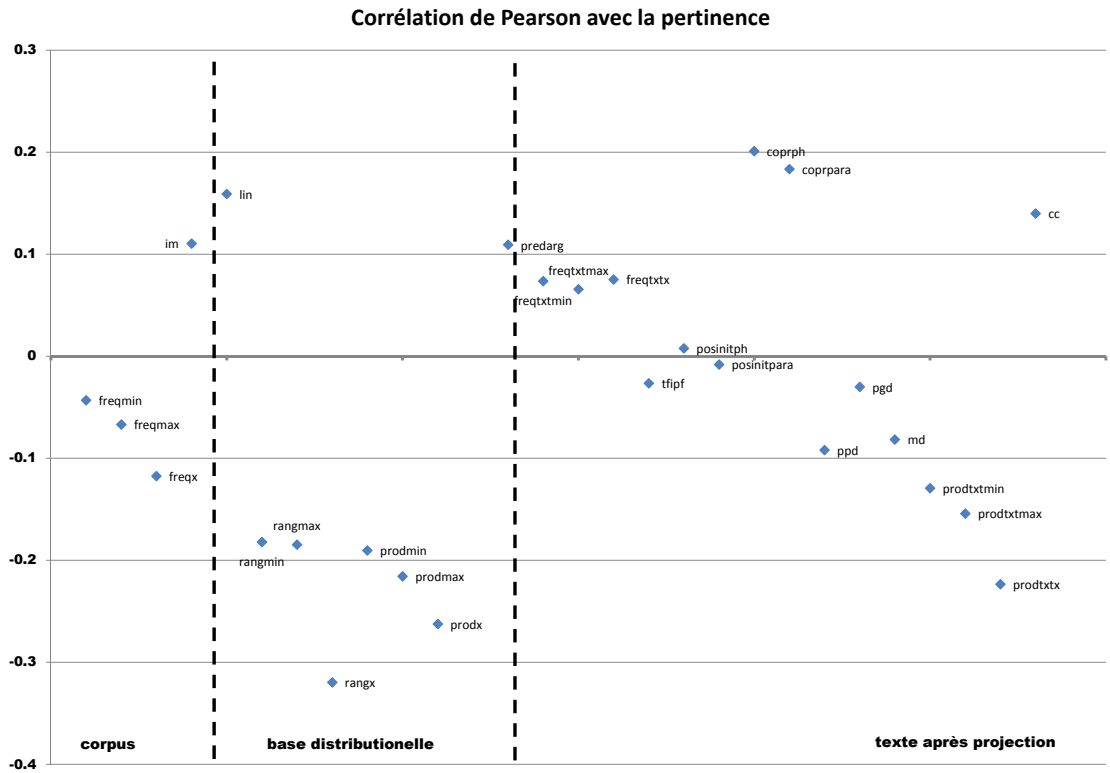


FIGURE 4.5 – Corrélation des indices avec la pertinence

La corrélation de Pearson permet, pour chaque indice, de déterminer s'il est positivement ou négativement associé à la pertinence d'un couple de voisins. Par contre, elle n'apparait pas très adaptée pour procéder à un classement des indices entre eux. En effet, les indices considérés sont de natures hétérogènes : certains indices sont numériques, d'autres sont binaires, etc. C'est pourquoi, dans un second temps, nous avons utilisé l'outil *Weka Explorer* (Hall *et al.*, 2009) pour calculer le gain d'information (*info gain*, Hall, 2011) pour chacun des indices. Le gain d'information d'un

Indice	Corrélation
$freq_{\min}$	-0.043^{***}
$freq_{\max}$	-0.067^{***}
$freq_{\times}$	-0.118^{***}
im	$+0.110^{***}$

Indice	Corrélation
lin	$+0.159^{***}$
$rang_{\min}$	-0.182^{***}
$rang_{\max}$	-0.185^{***}
$rang_{\times}$	-0.320^{***}
$prod_{\min}$	-0.190^{***}
$prod_{\max}$	-0.216^{***}
$prod_{\times}$	-0.263^{***}
$predarg$	$+0.109^{***}$

Indice	Corrélation
$freqtxt_{\min}$	$+0.074^{***}$
$freqtxt_{\max}$	$+0.066^{***}$
$freqtxt_{\times}$	$+0.075^{***}$
$tfipf$	-0.027^{*}
$posinit_{sec}$	$+0.008$
$posinit_{para}$	-0.008
$copr_{ph}$	$+0.201^{***}$
$copr_{para}$	$+0.183^{***}$
ppd	-0.092^{***}
pgd	-0.030^{***}
md	-0.082^{***}
$prodtxt_{\min}$	-0.129^{***}
$prodtxt_{\max}$	-0.154^{***}
$prodtxt_{\times}$	-0.223^{***}
cc	$+0.140^{***}$

TABLEAU 4.13 – Corrélations de Pearson entre les indices définis et la pertinence

indice y est défini comme la difference d'entropie (Shannon, 1948) mesurée pour la pertinence (X) selon que l'on connait ou non l'indice y :

$$IG(X, y) = H(X) - H(X|y) \tag{4.4}$$

Cette différence d'entropie indique dans quelle mesure un indice aide à discriminer entre différentes classes. Le gain d'information est une mesure d'informativité classique, et sert de base à certaines méthodes, notamment les arbres de décision. Néanmoins, le gain d'information est une mesure dont l'interprétation doit être faite de manière prudente, car les valeur obtenues peuvent varier d'un indice à un autre selon la distribution des valeurs possibles pour cette indice. En particulier, un trait catégoriel avec beaucoup de valeurs différentes aura tendance à présenter un gain d'information supérieur à un trait avec peu de valeurs différentes.

Le tableau 4.6 et la figure 4.14 résument les gains d'information calculés.

À partir de ces données, nous discutons de la pertinence des indices issus de la base distributionnelle (section 4.4.2), puis du corpus et du texte annoté (section 4.4.3). Nous discutons ensuite de la complémentarité entre différentes grandes classes d'indices (section 4.4.4).

4.4.2 Indices issus de la base distributionnelle

Score de Lin Comme nous l'avons mentionné, la stratégie la plus naturelle pour filtrer les voisins consiste à poser un seuil sur le score de Lin, qui est l'agent principal du rapprochement distributionnel. C'est pourquoi nous examinons en premier lieu

Rang	Indice	InfoGain	Rang	Indice	InfoGain
1	$rang_{\times}$	0.0604	15	lin	0.0154
2	$rang_{\max}$	0.0589	16	md	0.0132
3	$rang_{\min}$	0.0541	17	$freq_{\max}$	0.0123
4	im	0.0427	18	$freq_{\times}$	0.0118
5	$prod_{\max}$	0.0413	19	$freq_{\min}$	0.0111
6	$prod_{\times}$	0.0395	20	pgd	0.0104
7	$prod_{\min}$	0.0362	21	cc	0.0088
8	$prodtxt_{\times}$	0.0351	22	$predarg$	0.0083
9	$prodtxt_{\max}$	0.0319	23	$freqtxt_{\times}$	0.0047
10	$prodtxt_{\min}$	0.0310	24	$freqtxt_{\max}$	0.0047
11	ppd	0.0267	25	$tfipf$	0.0034
12	$cats$	0.0214	26	$freqtxt_{\min}$	0.0032
13	$copr_{ph}$	0.0208	27	$posinit_{para}$	0.0000
14	$copr_{para}$	0.0202	28	$posinit_{ph}$	0.0000

TABLEAU 4.14 – Gain d'information (InfoGain) vers la pertinence pour les indices émanant du texte

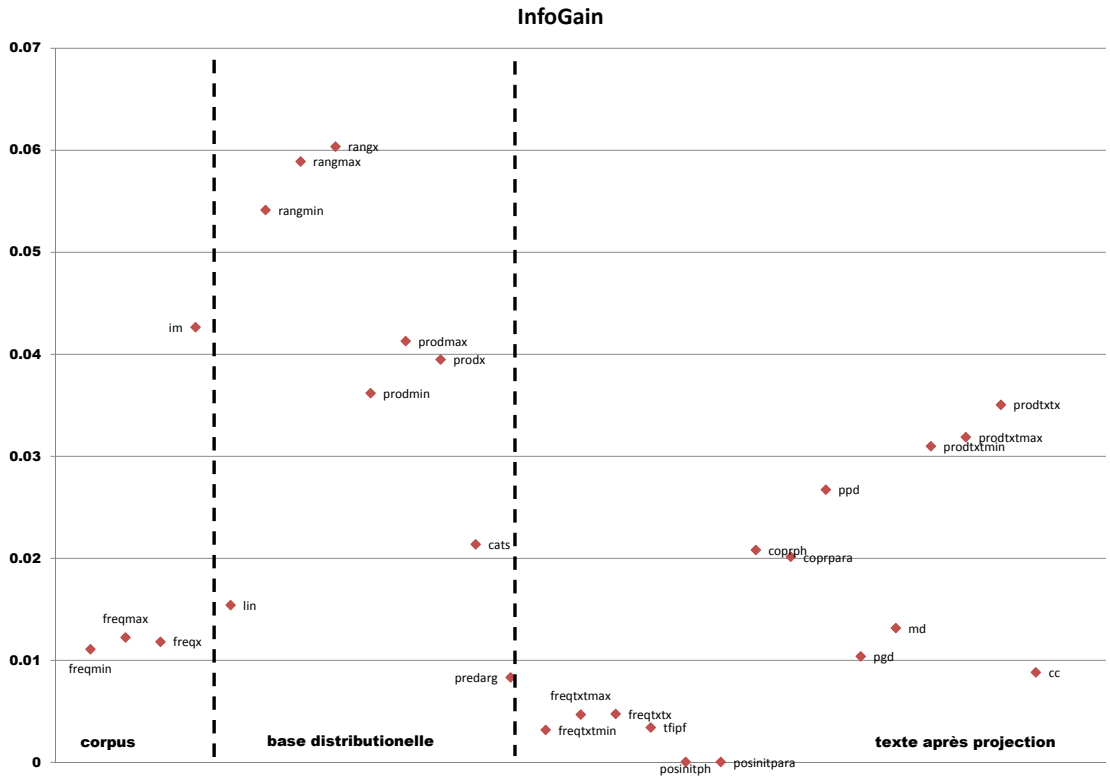


FIGURE 4.6 – Gain d'information (InfoGain) vers la pertinence pour chacun des indices

l'impact de cet indice sur la pertinence des liens projetés. Selon le tableau 4.13, le score de Lin apparaît effectivement corrélé à la pertinence des couples de voisins.

À partir des données annotées, nous pouvons observer l'effet d'un seuillage sur le score de Lin. La figure 4.7 permet de visualiser le nombre de couples acceptés, pertinents ou non, en fonction du seuil posé sur *lin*. On peut observer que le nombre de couples acceptés décroit très rapidement ; parmi eux, le nombre de couples pertinents décroit moins rapidement, ce qui se signifie une augmentation de la précision.

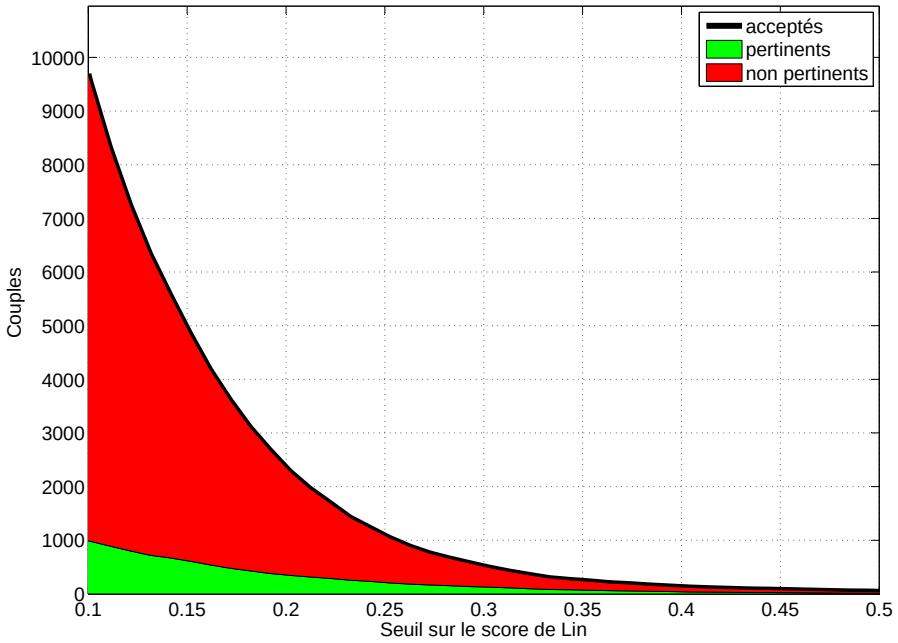


FIGURE 4.7 – Nombre de liens acceptés et proportion de liens pertinents en fonction du seuil sur le score de Lin

La figure 4.8 montre l'évolution des courbes de précision et de rappel, définis de la manière suivante :

- la précision est la proportion de couples pertinents parmi tous les couples acceptés ;
- le rappel est la proportion de couples acceptés parmi tous les couples pertinents.

Cette courbe peut être utilisée pour guider le choix d'un seuil, selon que l'on veuille favoriser la précision (accepter moins de couples, mais de meilleure qualité) ou le rappel (viser une forte couverture mais en acceptant beaucoup de bruit). Si l'on ne veut favoriser ni la précision ni le rappel, un seuil sur le point de croisement des courbes est approprié : avec un seuil à environ 0.24 sur le score de Lin, on obtient un rappel et une précision légèrement inférieurs à 25%. Nous pensons qu'un meilleur filtrage est possible en prenant en considération les autres indices définis.

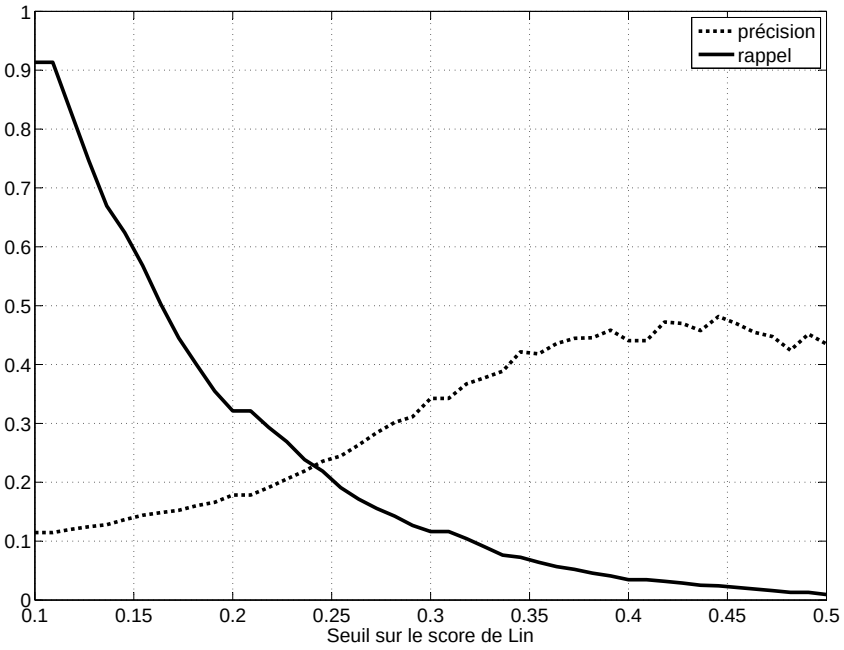


FIGURE 4.8 – Précision et rappel en fonction du seuil sur le score de Lin

Productivité et rangs Nous nous intéressons à présent à l'ensemble des indices distributionnels quantitatifs : *lin*, $rang_{\min}$, $rang_{\max}$, $rang_{\times}$, $prod_{\min}$, $prod_{\max}$, et $prod_{\times}$. Les différents indices liés au rang et à la productivité des voisins apparaissent, de manière nette, négativement corrélés à la pertinence. Cela signifie que :

- plus un voisin est productif, moins les couples dans lesquels il apparaît auront tendance à être pertinents ;
- plus un item apparaît tard parmi les voisins d'un autre item, moins leur relation aura tendance à être jugée pertinente.

Parmi $rang_{\min}$, $rang_{\max}$ et $rang_{\times}$, c'est ce dernier (qui combine les deux autres) qui a la plus forte valeur informationnelle. Parmi $prod_{\min}$, $prod_{\max}$, et $prod_{\times}$, c'est $prod_{\max}$ qui a la plus forte valeur informationnelle ; ainsi, il suffit que l'un des deux voisins d'un couple ait une forte productivité pour diminuer les chances que ce couple soit pertinent. Dans ce qui suit, pour plus de clarté, nous ne présentons, pour rendre compte des rangs et de la productivité des voisins, que ces deux indices $rang_{\times}$ et *prod*max.

Ces indices distributionnels ont tous pour but de rendre compte d'une même information : la qualité du rapprochement distributionnel. C'est pourquoi ces indices apparaissent globalement assez fortement corrélés entre eux (positivement ou négativement), comme le montre la matrice de corrélations 4.15. Toutefois, ces indices sont différents du point de vue de leur gain informationnel : c'est $rang_{\times}$ qui détient le meilleur gain informationnel (0.0604), suivi de $prod_{\max}$ (0.0413) puis de *lin* (0.0154).

On peut observer que *lin* ne défavorise pas assez les voisins trop productifs, alors que cette productivité apparaît fortement corrélée à la pertinence des couples qu'ils forment. Les indices basés sur le rang cumulent les deux informations : ils découlent directement du classement induit par *lin*, mais sont fortement influencés par la productivité des voisins. En effet, pour un couple $\texttt{voisin}_a/\texttt{voisin}_b$, le rang de $\texttt{voisin}_b$ parmi les voisins de $\texttt{voisin}_a$ tendra à être supérieur, pour un même score de Lin, si $\texttt{voisin}_a$ a plus de voisins. Cela est confirmé par la matrice de corrélation : alors que *lin* et *prod*max sont très faiblement corrélés entre eux, *rang*× apparaît bien corrélé à la fois avec *lin* et avec *prod*max. On comprend donc pourquoi le rang constitue un indice plus informatif que le score de Lin.

$rang_\times$	$prod_{\max}$	
-0.694^{***}	$+0.039^{***}$	*lin*
	$+0.457^{***}$	$rang_\times$

TABLEAU 4.15 – Matrice de corrélation pour les indices *lin*, $rang_\times$, et $prod_{\max}$

Les figures 4.9 et 4.10 permettent de visualiser l'évolution des couples acceptés selon la limite posée sur $rang_\times$. Elles sont construites de manière similaire aux figures 4.7 et 4.8 ci-dessus. Ces figures confirment la supériorité de $rang_\times$ sur *lin* si l'on ne veut se baser que sur un indice pour filtrer les voisins. Un bon point de comparaison est le croisement des courbes de précision et de rappel : proche de 35% lorsque l'on fait varier la limite sur $rang_\times$, inférieur à 24% lorsque l'on fait varier le seuil sur *lin*. Ainsi, selon nos données, en utilisant le seul indice $rang_\times$ sur lequel on pose une limite supérieure, il est possible de capter plus d'un tiers des couples pertinents en acceptant une erreur de 65% environ.

Caractéristiques morpho-syntaxiques des voisins Le tableau 4.13 ne contient pas d'information sur la corrélation entre *cats* et *pert*. Afin d'évaluer cette corrélation, nous avons développé l'indice *cats* en 6 indices booléens correspondant à ses valeurs possibles (précédemment énumérées dans le tableau 4.10) : les paires de catégories morpho-syntaxiques AA, AN, AV, NN, NV, VV. Le tableau 4.16 liste les corrélations calculées. Par ailleurs, l'histogramme 4.11 montre, pour chaque paire de catégories, le nombre de couples, pertinents ou non, qui relèvent de ces catégories.

Il ressort de ces données que les couples Nom / Nom, les plus fréquents, présentent la plus forte tendance à porter la cohésion lexicale des textes annotés. Les couples Adj. / Adj. présentent également une très faible corrélation positive avec *pert*. Les autres paires de catégories, par contraste, présentent une légère corrélation négative avec *pert*. Aucune tendance significative ne peut être observée pour les couples Adj. / Nom ou Adj. / Verbe, très rares.

En ce qui concerne l'indice *predarg*, une corrélation légère apparaît avec *pert* en faveur de la valeur « argument » : les couples d'arguments sont plus souvent pertinents que les couples de « prédicats ». Il apparaît que cette corrélation est uni-

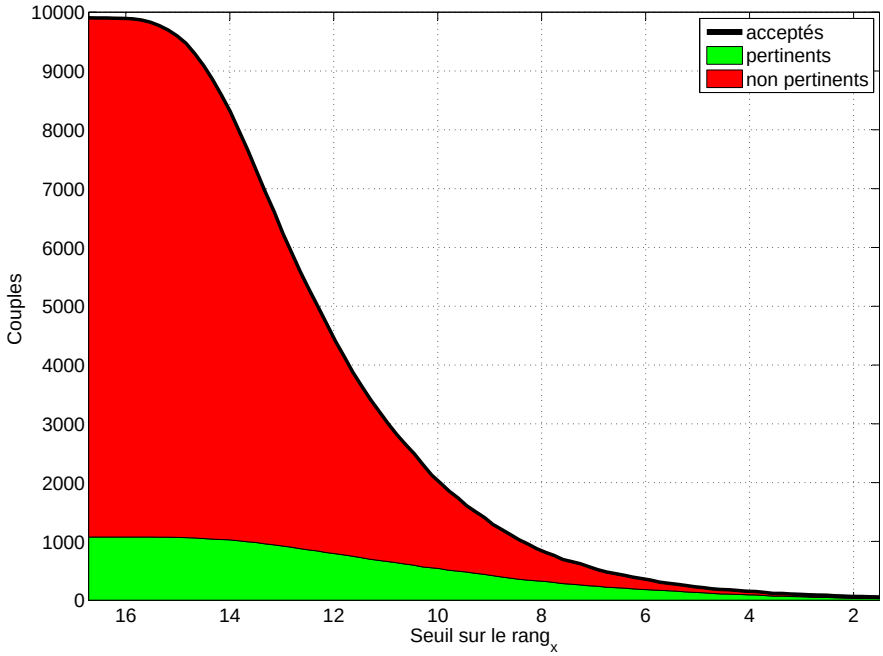


FIGURE 4.9 – Nombre de liens acceptés et proportion de liens pertinents en fonction de la limite sur $rang_{\times}$

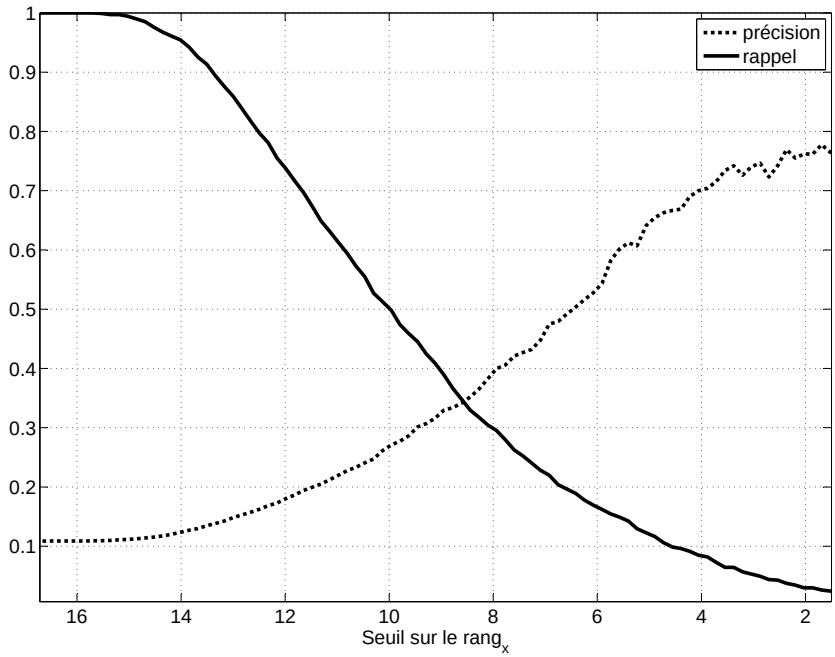


FIGURE 4.10 – Précision et rappel en fonction de la limite sur $rang_{\times}$

Indice	Corrélation
AA	0.0660^{***}
AN	0.0019
AV	-0.0172
NN	0.1256^{***}
NV	-0.0864^{***}
VV	-0.1053^{***}

TABLEAU 4.16 – Corrélations entre les différentes valeurs de *cats* et la pertinence

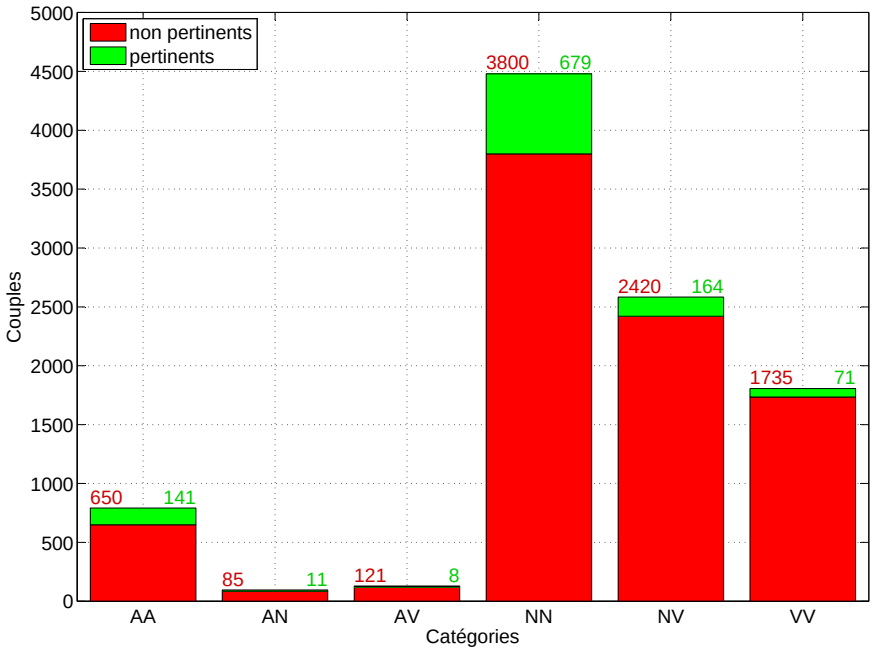


FIGURE 4.11 – Couples pertinents / non pertinents en fonction des catégories morpho-syntaxiques

quement le résultat des différences observées selon les catégories morpho-syntaxiques des voisins : les couples VV et NV sont toujours des couples de prédicats, tandis que les couples AA sont typiquement des arguments (*cf.* section 2.3.1 page 62). Les couples NN peuvent être argumentaux (environ 65% des cas) ou prédicatifs (environ 35%) des cas. Si l'on se base cette fois sur ces seuls couples NN pour calculer l'association *predarg* avec *pert*, il n'émerge aucune corrélation significative. Le rôle de prédicat ou d'argument n'est donc pas lié à la pertinence des couples de voisins.

Au terme de cette exploration des indices distributionnels, il est important de noter que ces derniers se placent logiquement comme les indices les plus informatifs

par rapport à la pertinence des voisins projetés, avec notamment la prévalence des indices liés aux rangs. Toutefois, parmi les autres indices définis, certains s'avèrent également très pertinents. Nous passons en revue l'ensemble des indices émanant du corpus ou des textes particuliers dans la section qui suit.

4.4.3 Indices émanant du corpus et du texte

Fréquence en corpus et information mutuelle Le principal « *outsider* » parmi les indices n'émanant pas directement de la base distributionnelle est l'information mutuelle (im), qui se place juste après les indices liés aux rangs en termes de gain informationnel (*cf.* tableau 4.14). Ainsi, ce score de collocation contribue fortement, dans la problématique de la détection de la cohésion lexicale, à faire émerger des rapprochements distributionnels pertinents. Nous discutons de cette complémentarité entre indices syntagmatiques et indices paradigmatiques dans la section 4.4.4.

Contrairement aux caractéristiques distributionnelles des couples, l'information mutuelle n'a pas pu être calculée pour tous les couples : on a vu (section 4.3.1) qu'elle pouvait être calculée pour 70% des voisins : les 30% restants n'apparaissent jamais dans un même paragraphe au sein de WikipédiaFR2007. Si l'on s'en tient aux couples projetés dans les textes et non à l'ensemble des couples de la base distributionnelle, cette proportion passe à 22%. Si une information mutuelle faible est fortement indicatrice d'un couple non-pertinent, par contre, le fait qu'im n'ait pas pu être calculé ne signifie rien concernant la pertinence des couples concernés. On peut donc réellement considérer que ces 22% de couples ne sont pas couverts par cet indice, dans le sens où l'absence de score n'est pas elle-même porteuse d'information. Ces observations sont appuyées par la figure 4.12. Le label « IM faible » est affecté aux couples dont l'information mutuelle est inférieure ou égale à la valeur du premier quartile (si l'on considère l'ensemble des *Voisins de Wikipédia* pour lesquels ce score a pu être calculé) et le label « IM forte » aux couples dont l'information mutuelle est supérieure au troisième quartile.

En ce qui concerne la fréquence en corpus des voisins (dont rendent comptent les indices $freq_{\min}$, $freq_{\max}$ et $freq_{\times}$), elle est corrélée négativement à la pertinence : les couples impliquant des mots très fréquents ont une plus faible tendance à mettre au jour des liens pertinents.

Fréquence en texte et autres indices de saillance La fréquence des voisins dans les textes où ils ont été projetés (exprimée par les indices $freqtxt_{\min}$, $freqtxt_{\max}$ et $freqtxt_{\times}$) apparaît faiblement positivement corrélée à la pertinence, à l'inverse de la fréquence en corpus. Les liens impliquant un mot fréquent dans le texte ont (très légèrement) plus tendance à être pertinents. Il aurait ainsi été intéressant de calculer un *tf·idf* sensible à ces deux informations (forte fréquence dans le document et faible fréquence dans le corpus).

En ce qui concerne le *tf·ipf* ainsi que les deux autres indices permettant d'identifier des items saillants au sein du texte ($posinit_{ph}$ et $posinit_{para}$), ils ne sont pas

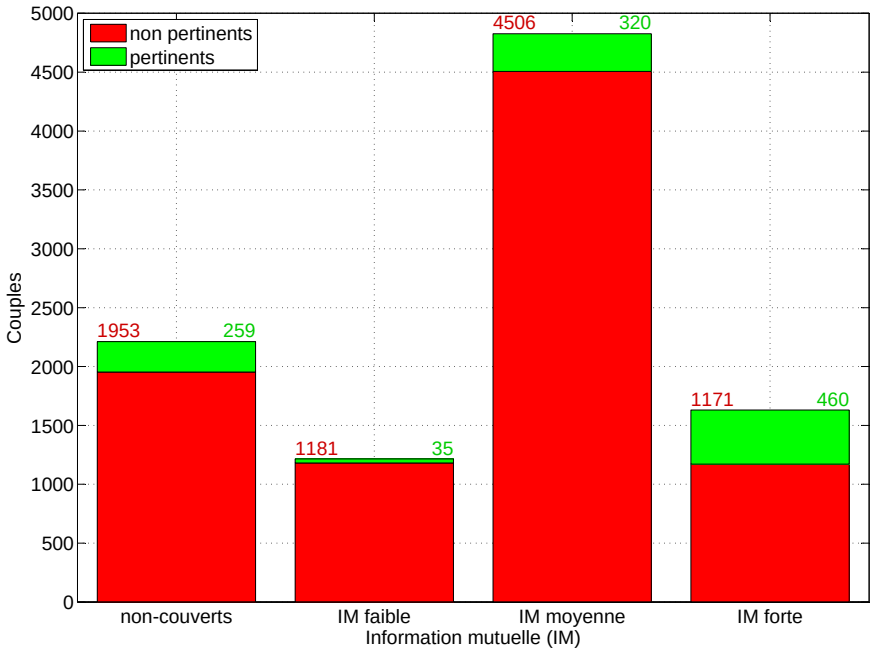


FIGURE 4.12 – Couples pertinents / non-pertinents en fonction de l'information mutuelle

corrélés avec la pertinence des couples de voisins. Ainsi, bien que ce *tf·idf* « local » semble adapté pour faire émerger des liens lexicaux utiles pour la segmentation thématique (Dias *et al.*, 2007) (donc supposément représentatifs de différents thèmes au sein d'un texte), cela ne veut pas dire qu'il aide à distinguer entre liens de cohésion lexicale et liens ne relevant pas de la cohésion lexicale. Peut-être que, de la même manière, les indices $posinit_{ph}$ et $posinit_{para}$ pourraient permettre de mettre au jour, parmi des liens de cohésion lexicale, certains liens ayant un rôle particulier.

Distance des items dans le texte Si les indices ayant pour but de rendre compte de la saillance des items au sein du texte se révèlent peu voire pas informatifs, par contre, les indices rendant compte de leur proximité apparaissent très pertinents.

Les indices $copr_{para}$ et $copr_{ph}$ apparaissent fortement corrélés à *pert*. Ces indices booléens permettent de sélectionner des sous-ensembles de couples qui sont pertinents dans 24% des cas pour $copr_{para}$ et dans 34% des cas pour $copr_{ph}$. Le problème de ces indices est bien sûr la taille des sous-ensembles sélectionnés, qui reflète leur très faible couverture : seuls 17% des couples de voisins projetés dans un texte apparaissent, au moins une fois, en co-présence dans le même paragraphe, et 7% dans la même phrase. Ces informations peuvent être visualisées à partir des figures 4.13 et 4.14.

Les indices de distance (*ppd*, *pgd* et *md*) présentent quant à eux l'avantage de

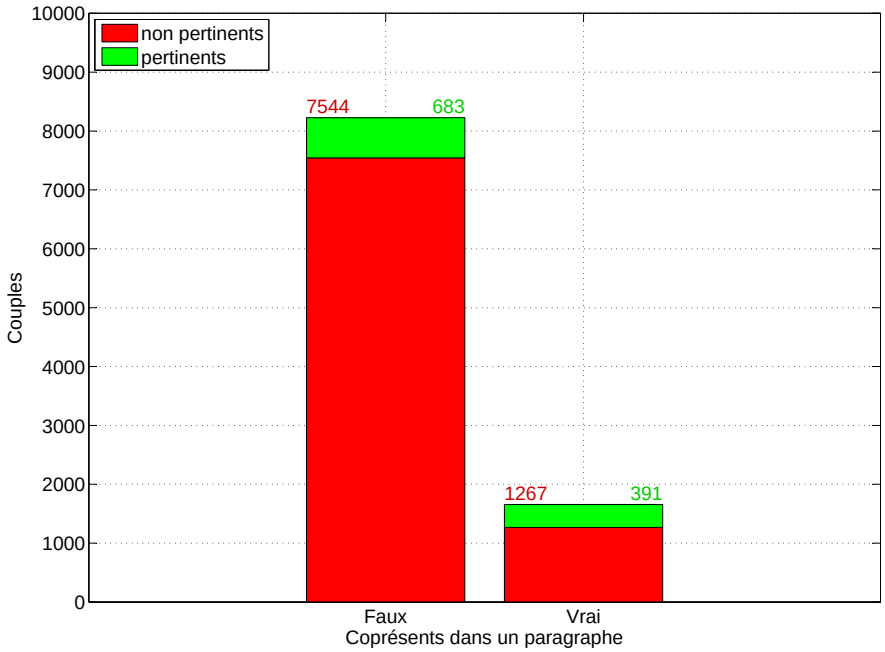


FIGURE 4.13 – Couples pertinents / non-pertinents en fonction de la valeur de $copr_{para}$

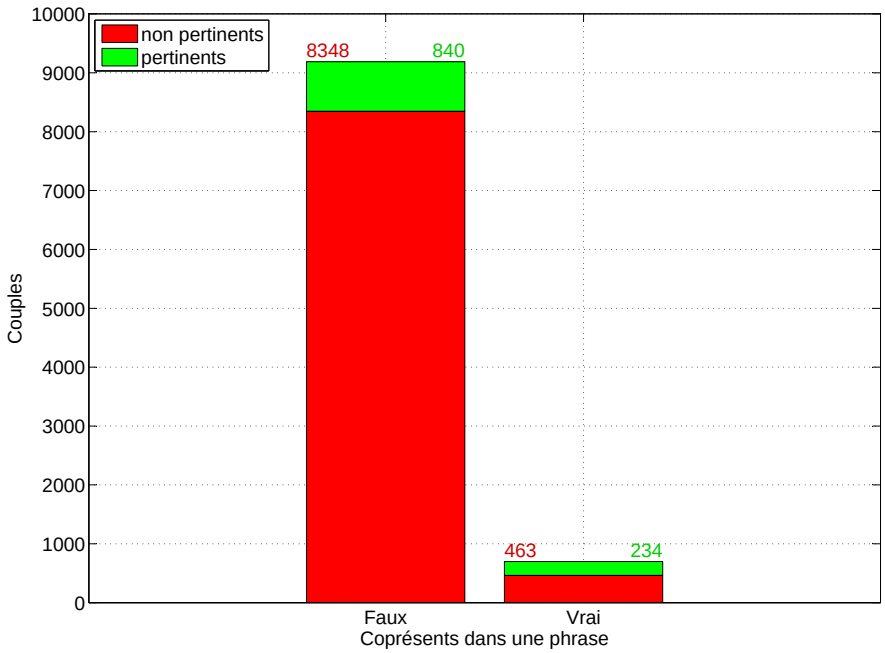


FIGURE 4.14 – Couples pertinents / non-pertinents en fonction de la valeur de $copr_{ph}$

couvrir l'ensemble des données. Parmi eux, c'est l'indice *ppd* qui rend le mieux compte de l'observation, effectuée lors de la phase d'annotation, selon laquelle deux items proches dans le texte sont plus souvent pertinents. Cet indice présente un des meilleurs gain d'information, se classant juste après les différents indices sur le rang, la productivité et l'information mutuelle.

La figure 4.15 permet d'apprécier le nombre de couples pertinents ou non-pertinents détectés en fonction d'un maximum sur la valeur de l'indice *ppd*. La figure 4.16 montre l'évolution des précision et rappel en fonction d'une limite placée sur *ppd*. On voit qu'à très petite distance dans les textes (quelques items), près de 60% des liens de voisinage sont considérés pertinents. De même que l'approche consistant à placer une limite sur les rangs des voisins, l'approche consistant à les filtrer en fonction de la distance minimale qui les sépare dans le texte permet des résultats plus satisfaisants que le seuillage sur le score de Lin.

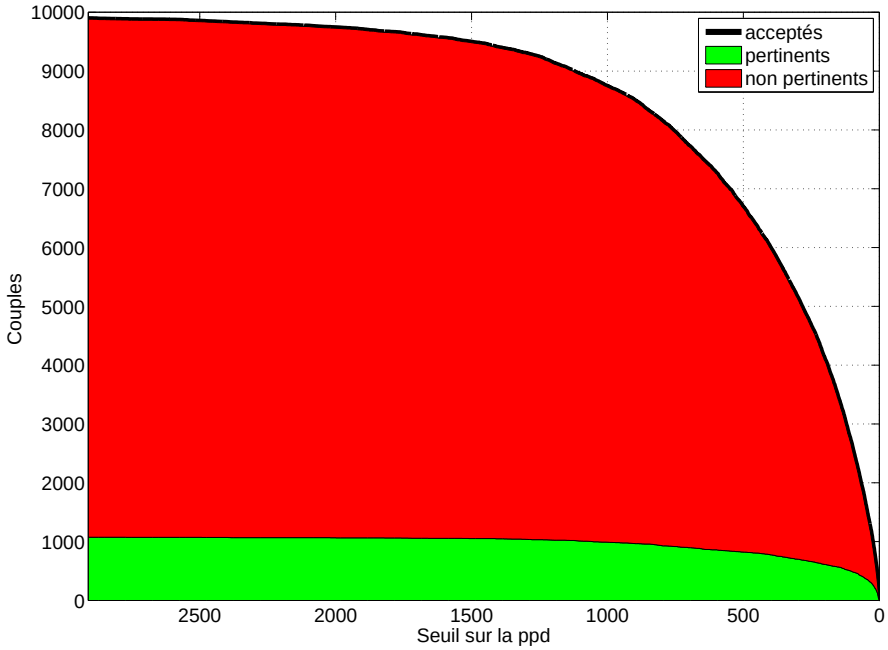


FIGURE 4.15 – Nombre de liens acceptés et proportion de liens pertinents en fonction de *ppd*

Nous avions déjà montré (dans la section 3.3.2 page 79) que lorsque la distance dans le texte est plus petite, un nombre plus important de liens est identifié par la projection des voisins ; ici, nous voyons que ces liens plus nombreux présentent également une bien plus forte tendance à être pertinents pour la détection de la cohésion lexicale. L'effet de la proximité textuelle est donc très important.

Les indices « textuels », bien que très simples, s'avèrent ainsi très informatifs. À ce stade, il faut bien rappeler que ce qu'on évalue, c'est la pertinence d'un indice

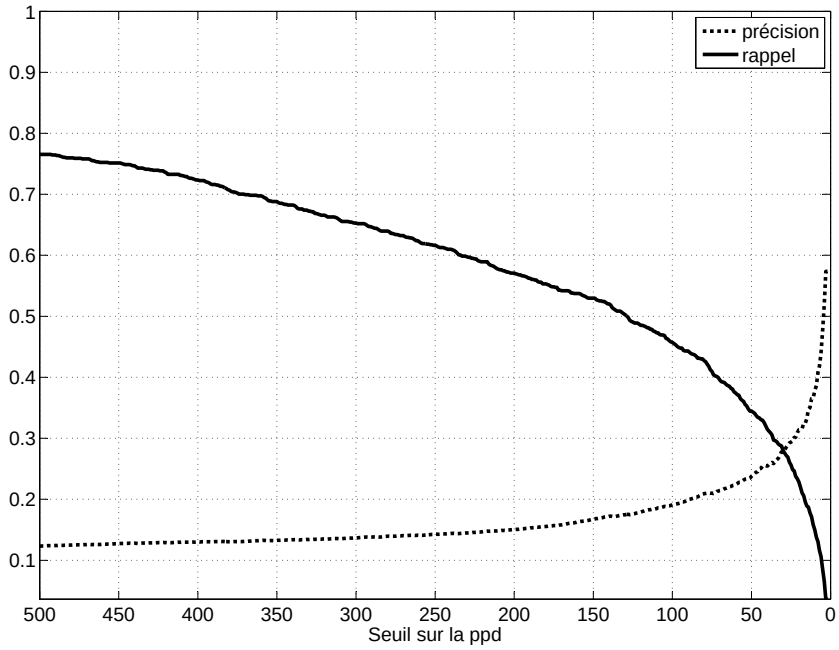


FIGURE 4.16 – Précision et rappel en fonction de *ppd*

donné pour faire émerger des liens de voisinage participant à la cohésion lexicale d'un texte, et pas directement pour détecter la cohésion lexicale. Ainsi, le fait qu'un couple de voisin apparaissant au sein de la même phrase soit jugé pertinent dans 34% des cas ne signifie absolument pas qu'un tiers des paires de mots co-présents dans la même phrase tissent un lien de cohésion lexicale, mais bien que deux mots qui non seulement sont voisins mais en plus apparaissent dans une même phrase ont de bonnes chances d'établir un lien de cohésion lexicale.

Indices issus du sous-graphe des voisins projetés dans le texte Les indices mesurant la productivité des voisins au sein du texte ($prodtxt_{\min}$, $prodtxt_{\max}$ et $prodtxt_{\times}$) apparaissent fortement liés à *pert*. Mais une observation plus approfondie montre que ces indices ne présentent pas vraiment d'autonomie par rapport aux indices $prod_{\min}$, $prod_{\max}$ et $prod_{\times}$, puisqu'ils leur sont fortement corrélés avec des coeficients respectifs de 0.43, 0.45 et 0.59. Ainsi, ils sont surtout le reflet, moins informatif selon le classement par gain d'information, de la productivité des voisins dans la base distributionnelle. Un indice faisant le rapport entre la productivité des voisins dans le texte et leur productivité dans la base distributionnelle serait sans doute bien plus intéressant, car il permettrait de mettre au jour des voisins qui, bien qu'ayant une faible productivité, apparaissent fortement connectés dans un texte donné.

L'appartenance à une composante connexe de petite taille est un indice très fort

(près de 50% de couples pertinents), mais à très faible couverture (seuls 154 couples sont concernés parmi les 9885 annotés). La valeur informationnelle de cet indice est ainsi assez faible. Néanmoins, nous pensons que cette piste est à creuser. Notamment, l'introduction de méthodes plus complexes de clustering pourrait faire émerger des groupes de voisins ne répondant pas à la définition stricte d'une composante connexe, mais présentant des propriétés similaires (groupes de voisins fortement connectés entre eux, et peu connectés au reste du sous-graphe).

4.4.4 Complémentarité entre indices « paradigmatiques » et indices « syntagmatiques »

L'intérêt de définir de nouveaux indices est d'autant plus grand que ces indices sont peu corrélés aux précédents (on parle d'« orthogonalité » de ces indices), et offrent ainsi une bonne complémentarité. Nous nous intéressons dans cette section aux quinze indices apportant le meilleur gain d'information selon le tableau 4.14. On retrouve dans ces quinze indices l'ensemble des indices émanant de la base distributionnelle (à l'exception de *predarg* dont nous avons noté qu'il n'était pas pertinent), qui rendent compte d'une relation paradigmatique entre deux mots ($\texttt{voisin}_a$ et $\texttt{voisin}_b$). En dehors de ces indices distributionnels, on trouve les indices suivants, par gain informationnel décroissant : *im*, *ppd*, $copr_{ph}$ et $copr_{para}$. Ces indices captent des informations qui sont plutôt de nature syntagmatique, liées à la tendance de $\texttt{voisin}_a$ et $\texttt{voisin}_b$ à apparaître ensemble, au sein du corpus ou d'un texte particulier. Ce que nous souhaitons montrer, c'est que ces indices « syntagmatiques » sont assez peu corrélés aux indices « paradigmatiques ». La matrice 4.17 montre les corrélations entre les différents indices syntagmatiques, et la matrice 4.18 entre ces indices syntagmatiques et les indices paradigmatiques.

$copr_{para}$	*ppd*	*im*	
$+0.614^{***}$	-0.248^{***}	$+0.076^{***}$	$copr_{ph}$
	-0.376^{***}	$+0.082^{***}$	$copr_{para}$
		-0.119^{***}	*ppd*

TABLEAU 4.17 – Corrélation entre indices syntagmatiques

De ces données, il émerge que :
- tout comme les indices paradigmatiques étaient assez corrélés entre eux (*cf.* matrice 4.16), les indices $copr_{ph}$, $copr_{para}$ et *ppd* sont corrélés avec des coefficients dont la valeur absolue est comprise entre 0.24 et 0.62 ;
- par contre, ces indices sont peu corrélés aux indices dits paradigmatiques, avec des coefficients dont la valeur absolue est comprise entre 0.004 et 0.100 ;
- l'indice *im* forme une classe à part, puisqu'il présente des corrélations relativement faibles à la fois avec les deux catégories d'indices.

		indices paradigmatiques		
		lin	$rang_{\times}$	$prod_{\max}$
ind. synt.	$copr_{ph}$	$+0.071^{***}$	-0.100^{***}	-0.041^{***}
	$copr_{para}$	$+0.062^{***}$	-0.089^{***}	-0.040^{***}
	ppd	-0.062^{***}	$+0.041^{***}$	-0.004
	im	$+0.131^{***}$	-0.130^{***}	-0.070^{***}

TABLEAU 4.18 – Corrélation entre indices paradigmatiques et indices syntagmatiques

En ce qui concerne l'indice *cat*, nous précisons, sans détailler les données (l'analyse nécessitant de binariser chacune de ses valeurs), que cet indice apparaît également peu corrélé avec l'ensemble des autres indices (mais d'avantage corrélé avec les indices paradigmatiques, notamment $rang_{\times}$).

Ainsi, il émerge différentes classes d'indices :

- des indices paradigmatiques calculés en corpus pour chaque paire de voisins (indices émanant de la base distributionnelle) ;
- un indice syntagmatique calculé statistiquement sur tout le corpus pour chaque paire de voisins le permettant (*im*) ;
- des indices évaluant la proximité syntagmatique de deux voisins au sein du texte considéré ($copr_{ph}$, $copr_{para}$ et *ppd*).

Ainsi, la pertinence des indices contextuels que nous avons proposés (calculés à l'échelle du corpus ou du texte) se voit renforcée par le fait qu'ils permettent de capter des informations différentes, certainement liées à la possible instanciation de relations lexicales en contexte (*cf.* section 1.3.1 page 37) ; ils complètent ainsi de manière convaincante les indices distributionnels.

La combinaison effective des différents indices pour le filtrage de la ressource est envisagée dans la section suivante.

4.5 Exploitation des indices pour le filtrage des liens projetés

Au terme de l'exploration des indices définis, nous en savons plus sur la relation entre la pertinence des couples de voisins pour détecter la cohésion lexicale et certaines de leurs caractéristiques. Dans cette section, nous discutons de l'exploitation de ces résultats pour le filtrage des liens projetés. Comme le montre la figure 4.17, la phase de filtrage prend place après la projection des liens en texte, sur la base des indices extraits aux différentes étapes.

Le prolongement naturel de la démarche menée jusqu'ici consiste à mettre en place un dispositif d'apprentissage automatique supervisé à partir des données annotées et des indices définis, afin de proposer une classification automatique des

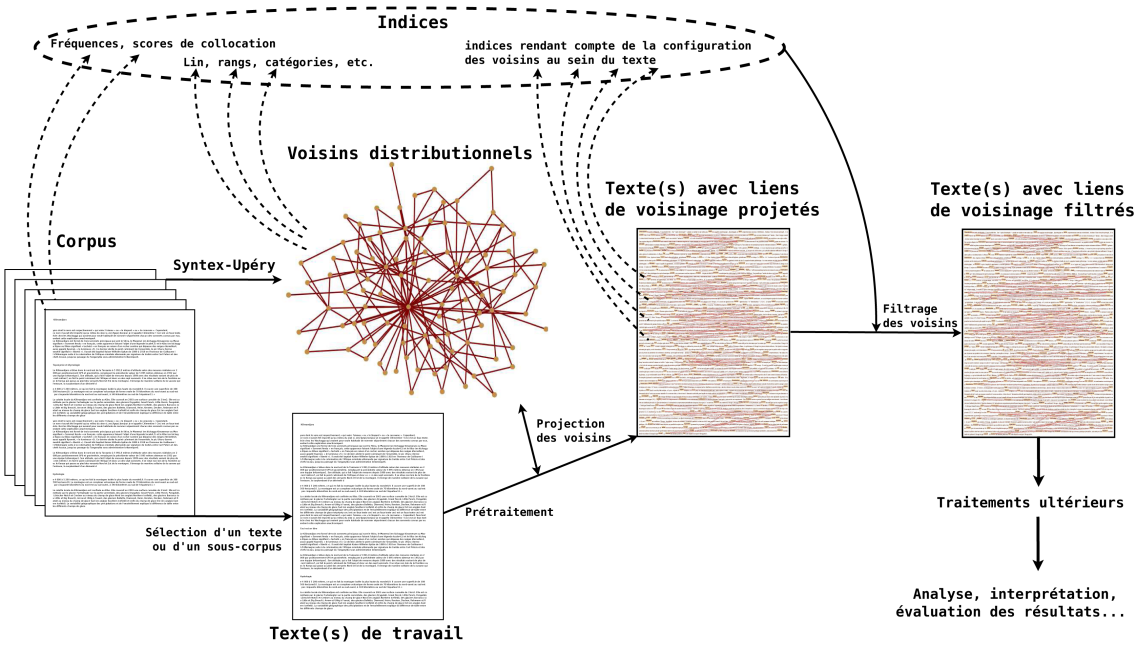


FIGURE 4.17 – Filtrage des liens projetés

liens projetés. Dans la section 4.5.1, nous montrons quels pourraient être les résultats d'une telle classification. Dans la section 4.5.2, nous expliquons pourquoi nous avons fait appel à d'autres stratégies de filtrage lors des expérimentations décrites dans la III^{e} partie de ce document, et nous décrivons ces stratégies en relation avec les grandes catégories d'objectifs auxquels elles répondent.

4.5.1 Vers une classification automatique des liens de voisinage ?

L'apprentissage automatique est un champ d'étude de l'intelligence artificielle ; l'objectif de l'apprentissage automatique dit supervisé de faire émerger, à partir d'un ensemble de données annotées, un modèle d'explication de ces données qui pourra être utilisé pour faire des prédictions sur de nouvelles données non-annotées.

Algorithme d'apprentissage Nous avons utilisé sur nos données l'implémentation proposée par Weka de l'algorithme d'apprentissage Random Forest. Les méthodes Random Forest, proposées par Breiman (2001), visent à rendre plus robuste une classification par arbres de décision en intégrant une composante probabiliste lors de l'apprentissage. Chaque arbre est construit de manière indépendante, et la classification finale résulte du vote de tous les arbres. Lors de la construction des arbres, un ensemble aléatoire de m indices parmi les M indices disponibles est utilisé pour la décision à chaque noeud de l'arbre, un nouveau tirage ayant lieu pour cha-

cun des noeuds. Weka propose la valeur par défaut $m = log\,(M + 1)$. De plus, une seconde option pour l'augmentation de la robustesse est d'utiliser une stratégie de *bagging* (Breiman, 2001, section 4), où l'apprentissage à chaque noeud est basé sur un sous-ensemble de n données d'apprentissage parmi les N données disponibles.

Méthode d'évaluation Pour évaluer les résultats obtenus, nous utilisons une validation croisée (*cross-validation*) (Witten *et al.*, 2011, pp. 152-154).

Évaluer les résultats d'un système d'apprentissage automatique suppose de partitionner les données annotées en deux sous-ensembles :

- les données d'apprentissage, utilisées par le système pour construire un modèle ;
- les données de test : le système utilise le modèle construit pour classer ces données ; la classification automatique est alors comparée aux annotations de référence pour évaluer le système.

Le principe de la validation croisée consiste à diviser les données annotées en un nombre n d'échantillons (typiquement 10, on parle de *10-fold cross-validation*). L'évaluation se fait alors en n itérations ; à chaque itération, on sélectionne un échantillon différent pour jouer de rôle de données de test, en utilisant les $n-1$ échantillons restants comme données d'apprentissage. Ainsi, le système est testé successivement sur l'ensemble des données.

Résultats La matrice de confusion 4.19 montre les résultats obtenus.

	non-pert.	pert.	← Auto.
non-pert.	8701	110	
pert.	813	261	

↑ Référence

Exactitude : 90.7%

TABLEAU 4.19 – Matrice de confusion pour la classification automatique

Le taux d'exactitude (c'est-à-dire la proportion d'instances correctement classifiées) est très élevé du fait de l'important déséquilibre entre les deux classes (*pertinent* et *non-pertinent*) : la *baseline* consistant à considérer que tous les liens sont non-pertinents permet déjà d'atteindre un taux d'exactitude de 89.1%. Si l'on reprend les notions de précision et de rappel telles que posées précédemment (c'est-à-dire la précision et le rappel pour la classe *pertinent*, *cf.* section 4.4.2), cette classification correspond à une précision de 70.4% (261 liens pertinents sur les 371 liens acceptés) pour un rappel de 25.3% (261 liens acceptés sur les 1074 liens pertinents). Si l'on se réfère à la figure 4.10, on voit que, pour atteindre la même précision en posant un seuil sur $rang_{\times}$, il faudrait accepter un rappel de 8.7% seulement.

Impact des différentes classes d'indices Le tableau 4.20 permet de mesurer l'impact des différentes classes d'indices discutées dans la section 4.4.4. Il a été

obtenu en répétant la phase d'apprentissage en soustrayant à chaque fois un certain nombre d'indices (précisés dans la première colonne).

Indices	Précision	Rappel	F-mesure	Exactitude
Tous	70.4%	24.3%	36.1%	90.7%
$-im$	64.0%	20.7%	31.2% (-4.9)	90.1%
$-(lin, rangs, prod, prodtxt)$	54.9%	18.9%	28.1% (-8.0)	89.5%
$-cats$	65.5%	23.4%	34.5% (-1.6)	90.3%
$-(ppd, pgd, md, copr_{ph}, copr_{para})$	65.0%	23.0%	34.0% (-2.1)	90.3%

TABLEAU 4.20 – Impact des différentes classes d'indices

Ces résultats confirment ce que nous avions observé dans la phase d'exploration des indices, puisque les indices distributionnels sont ceux qui présentent le plus fort apport, suivis de l'information mutuelle et des indices syntagmatiques.

Perspectives Les résultats que nous avons présentés ici sont encourageants. Eu égard au caractère extrêmement pléthorique et fortement bruité de la ressource projetée, souligné dans les chapitres 2 et 3 et confirmé par les résultats de l'annotation, atteindre une précision de 70% avec un rappel de 25% est au-dessus de nos espérances. C'est en tout cas bien meilleur que ce qu'une démarche naturelle, consistant à poser un seuil sur le score de Lin, aurait permis (*cf.* figure 4.8 : pour un rappel de 25%, la précision est de 21%).

Toutefois, nous considérons qu'il s'agit de résultats préliminaires, indiquant que cette direction de recherche constitue une piste à poursuivre. Les perspectives que nous envisageons sont de deux ordres.

(a) D'une part, les résultats pourraient être améliorés en affinant certains des indices décrits et en en proposant de nouveaux. Dans la section 4.4, nous avons évoqué d'autres indices dont nous pensons qu'ils pourraient se révéler très pertinents : le *tf·idf*, une productivité « relative » (définie comme *prodtxt/prod*), ainsi que d'autres indices faisant intervenir des objets issus de la théorie des graphes : cliques, composantes connexes, clusters, etc.

(b) D'autre part, poursuivre ce travail impliquerait selon nous de discuter de la démarche à adopter avec un expert de l'apprentissage automatique. Nous pensons par exemple que la mise en place d'une seconde phase d'annotation après un premier pré-filtrage (permettant d'exclure les liens les moins probablement pertinents) pourrait constituer une approche intéressante, car elle permettrait de disposer plus rapidement de plus d'instances « positives » pour l'apprentissage. Ce pré-filtrage pourrait se baser sur l'entraînement d'un classificateur spécialisé, qui privilégierait le rappel sur les instances pertinentes au détriment de la précision globale, par exemple au moyen d'une matrice de coût qui pénaliserait plus fortement les erreurs sur la classe voulue.

4.5.2 Différentes stratégies de filtrage pour différents objectifs

Nous avons vu dans la section précédente qu'une phase d'apprentissage automatique – prenant en entrée un texte annoté et l'ensemble des indices calculés et fournissant en sortie le texte annoté avec seulement les liens catégorisés comme pertinents – était envisageable. Toutefois, ce n'est pas la solution qui a été adoptée dans toutes les expériences relatées dans la suite de cette thèse, pour différentes raisons.

Tout d'abord, comme nous l'avons mentionné, nous pensons que le travail décrit dans ce chapitre pourrait être prolongé avant d'aboutir à un dispositif de classification automatique des liens projetés intégré dans notre chaîne de traitement.

D'autre part, il est important de rappeler que cette réflexion sur le filtrage des voisins a été menée en interaction avec les expérimentations visant à exploiter les voisins dans différentes approches du discours : nous avons dans un premier temps tenté d'exploiter les *Voisins de Wikipédia* tels quels, puis nous avons adopté des heuristiques de filtrage fixées de manière empirique, ce qui nous a naturellement amenée à nous interroger sur la validité de ces heuristiques et à réaliser le travail décrit dans ce chapitre, qui a permis d'affiner les méthodes de filtrage utilisées.

Le fruit de ces interactions est la définition de différentes stratégies de filtrage correspondant à différents besoins rencontrés au cours du projet VOILADIS. Ces différents besoins couvrent selon nous les principaux objectifs qui peuvent régir l'exploitation de la cohésion lexicale pour l'étude de phénomènes discursifs. Pour définir ces objectifs, nous classons les situations d'exploitation selon deux grands axes :

(a) Dans quel type d'approche se trouve-t-on ? On a vu dans le chapitre 3 que le fait de se trouver dans une approche globale ou locale (ciblée) avait une influence sur la méthode de projection (*cf.* section 3.2.1). Nous verrons qu'une approche locale suppose une exploitation plus fortement contrainte par des paramètres contextuels, et donc une importance moindre accordée au filtrage.

(b) Pour qui détecte-t-on la cohésion lexicale : pour la machine (exploitation automatique) ou pour l'analyste (exploitation qualitative) ? Dans le cas de l'étude de phénomènes locaux, à petite échelle, la réponse à cette question n'implique pas forcément de différences dans le filtrage des liens à projeter. Par contre, pour une exploitation à plus grande échelle, une exploitation qualitative suppose de sélectionner davantage l'information fournie en sortie, alors qu'une exploitation automatique peut manipuler une information beaucoup plus riche.

Dans la suite de cette section, nous discutons des différentes stratégies de filtrage adoptées pour satisfaire ces différents objectifs. Nous faisons référence aux situations concrètes dans lesquelles elles ont été mises en œuvre, lors des expérimentations décrites dans la IIIe partie de ce document de thèse.

4.5.2.1 Filtrer pour la machine : stratégie pour un filtrage « robuste »

Exemple de situation Dans certains cas, les liens projetés ont surtout vocation a être traités automatiquement, par exemple dans le cadre d'une exploitation pour la tâche de segmentation thématique (chapitre 5), ou pour le calcul automatique de la force de la cohésion lexicale d'un objet linguistique tel que les structures énumératives (chapitre 5).

Stratégie de filtrage Dans une telle situation, nous avons fait appel à un filtrage peu important, dans le souci de garder une forte couverture.

Pour ce faire, nous exploitons le modèle d'apprentissage automatique construit (section 4.5.1). Mais nous ne l'utilisons pas directement pour classifier les couples de voisins des textes traités : nous utilisons la sortie intermédiaire qui précise, pour chaque couple, sa probabilité d'être pertinent (c'est-à-dire, avec le modèle RandomForest, la proportion d'arbres de décisions votant pour la classe *pert*). Afin de favoriser davantage le rappel par rapport à la précision, nous posons un seuil à 0.3 sur cette probabilité ; en effet, la validation croisée menée dans la section 4.5.1 a montré qu'un seuillage normal à 0.5 conduit à des scores de précision et de rappel de respectivement 70.4% et 25.3%. Nous pondérons ensuite chaque lien par la valeur de probabilité, qui présente l'avantage d'être plus fiable que le score de Lin et de couvrir toute la gamme de scores possibles entre 0.3 et 1 (alors que les valeurs de *lin* se concentrent majoritairement entre 0.1 et 0.2).

4.5.2.2 Filtrer pour l'analyste : stratégies pour un filtrage sélectif

Exemple de situation Lorsque l'on se situe dans une perspective plus qualitative et que l'on souhaite visualiser les liens associés à certaines structures (notamment les structures énumératives dans le chapitre 6), la problématique est légèrement différente de celle du cas précédent.

Stratégie de filtrage Dans la mesure où l'important ici est la visualisation des liens (*cf.* 3.3.3 page 82), la pondération des liens est ici inutile (le lien est tracé ou non). Pour pouvoir interpréter les liens projetés, il est nécessaire de filtrer plus fortement, en favorisant cette fois la précision sur le rappel. L'utilisation « telle quelle » du classifieur présenté dans la section 4.5.1 paraît ici une bonne option. Néanmoins, elle n'a pas toujours été la solution mise en œuvre pour les raisons citées en préambule.

D'autre part, l'utilisation de composantes connexes ou cliques, qui font émerger des objets de plus grande taille qu'un lien unique (*cf.* section3.4.1 page 86) – et constituent ainsi une autre méthode de sélection de l'information – peut s'avérer pertinente, surtout si elle est menée en interaction avec les structures discursives étudiées. Par exemple, dans la section 6.3.1 (page 199), nous proposons une visualisation du contenu lexical des structures énumératives basée sur l'extraction de

composantes connexes couvrant l'intégralité des éléments (items) des SE.

4.5.2.3 Filtrer pour une exploitation locale : stratégie pour un filtrage minimal

Exemple de situation Dans la section 3.2.1 page 71, nous avons discuté des différentes modalités d'exploitation des voisins. Nous avons évoqué une exploitation dans une approche locale, par interrogation de la base de voisins sur des items particuliers, repérés par exemple par l'application de patrons (*cf.* figure 3.2 page 73). Ainsi, dans le chapitre 8 consacré à la relation discursive d'élaboration, nous serons amenée à rechercher des phrases contenant un syntagme gérondif, et dont les verbes de la proposition principale et du syntagme gérondif soient voisins (ce motif ayant été identifié comme un possible marqueur de la relation d'élaboration, *cf.* section 8.3). On se trouve ici dans un cas d'exploitation fortement contrainte : plutôt que de projeter les voisins « tous azimuts », on identifie d'abord des positions entre lesquelles la présence d'un lien de cohésion lexicale jouerait un rôle particulier, avant d'interroger la base de voisins distributionnels pour déterminer si un tel lien est réalisé.

Stratégie de filtrage Les contraintes posées dans ce type d'exploitation des voisins agissent déjà partiellement comme des filtres : on a par exemple vu que des voisins établissant un lien non pertinent ont moins de chances d'apparaître ensemble dans une fenêtre réduite (dans notre exemple ci-dessus, la phrase). Ainsi, ce type d'exploitation rend le filtrage des voisins moins critique ; un filtrage léger s'avère suffisant.

La projection des voisins est, dans une telle situation, déjà régie par des contraintes d'ordre syntagmatique : contraintes sur les positions syntaxiques des items de voisinage et sur leur proximité. On a montré l'orthogonalité entre indices « syntagmatiques » et indices « paradigmatiques » pour la prédiction de la pertinence d'un lien de voisinage (section 4.4.4) ; il paraît donc judicieux de compléter ces contraintes syntagmatiques par des indices rendant compte de la qualité du rapprochement paradigmatique des voisins, comme *lin* et $rang_{\times}$. Ainsi, dans ce type de situations (rencontrées dans le cadre d'une approche locale du discours), nous avons choisi de ne filtrer les voisins que par des seuils posés sur le score de Lin ou le nombre de voisins pris en compte pour chaque item de la base des voisins. L'exploration des indices a montré que $rang_{\times}$ possède une plus grande valeur informationnelle que *lin*. Nous discutons donc dans la suite de l'utilisation de cet indice, en montrant qu'une stratégie adaptée aux contraintes d'exploitation permet de filtrer les voisins de manière plus satisfaisante qu'une stratégie « tout-venant » de classification automatique.

Nous avons évoqué les résultats pouvant être obtenus en posant une limite sur $rang_{\times}$ (figure 4.10), mais nous n'avons pas évoqué la combinaison de $rang_{\times}$ avec d'autres indices ou contraintes.

Premièrement, on peut adapter cette limite en fonction des catégories morpho-syntaxiques des items considérés. La figure 4.18 montre la distribution des couples pertinents et non-pertinents en fonction des catégories morpho-syntaxiques. Des différences apparaissent clairement ; par exemple, les couples composés d'un adjectif et d'un nom ou d'un verbe connaissent des valeurs de $rang_\times$ bien plus élevées en moyenne que les couples composés de deux adjectifs. Cela signifie que les premiers voisins d'un adjectif, dans un classement par score de Lin, sont généralement des adjectifs et que les autres catégories n'arrivent que plus tardivement. La répercussion de cette observation est que, selon les catégories des items à propos desquels on interroge la base, la limite sur $rang_\times$ devra être différente.

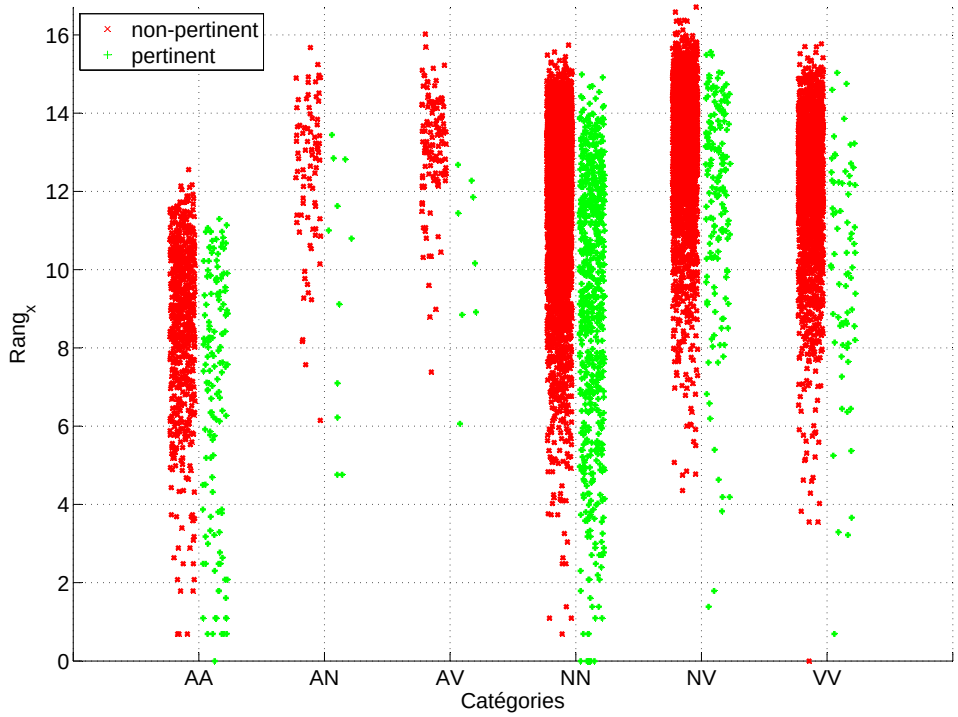


FIGURE 4.18 – Distribution des couples pertinents / non pertinents en fonction des catégories morpho-syntaxiques

Les figures 4.19 et 4.20 montrent les courbes de précision et de rappel pour deux valeurs de *cats* (NN et VV). L'aspect accidenté des courbes est dû à la plus faible quantité de données. Ces figures permettent toutefois d'apprécier les différences de choix qui pourront être faits ; par exemple, pour atteindre une précision de 20%, une limite sur $rang_\times$ aux alentours de 13 suffit pour les couples NN, tandis qu'elle sera plutôt d'environ 8 pour les couples VV.

L'autre contrainte dont nous discutons la prise en compte est la proximité entre les items. Nous avons vu que cette proximité est fortement corrélée avec la pertinence

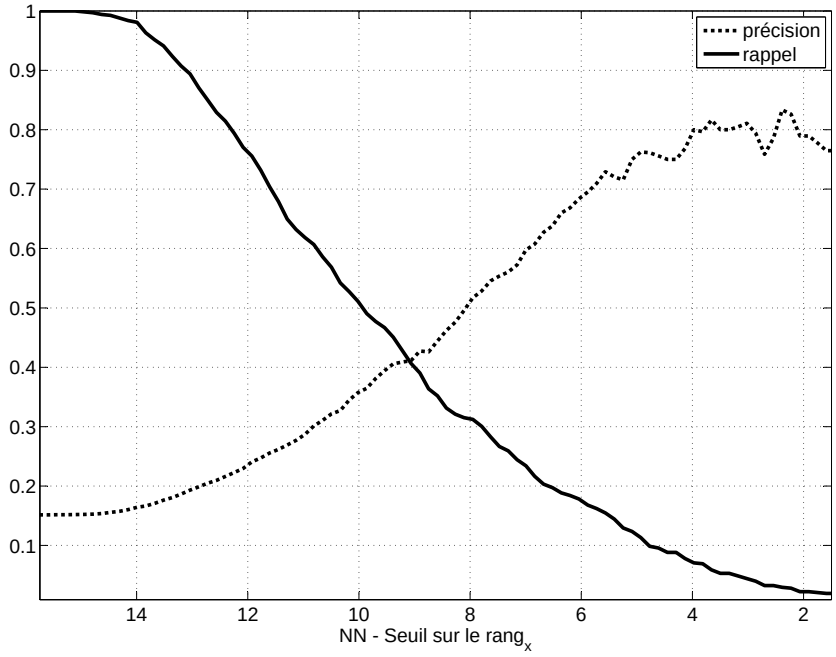


FIGURE 4.19 – Précision et rappel en fonction de $rang_\times$ pour les couples NN

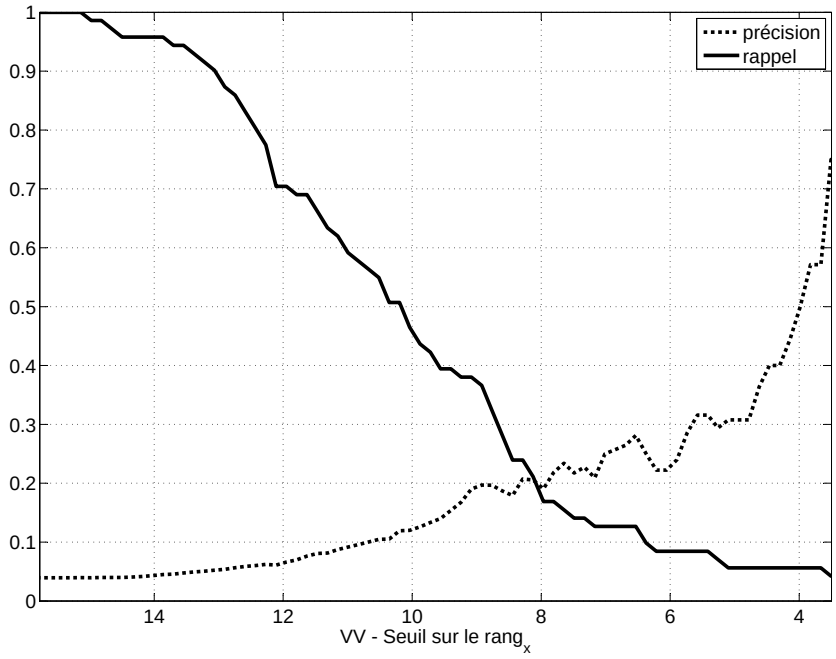


FIGURE 4.20 – Précision et rappel en fonction de $rang_\times$ pour les couples VV

des liens de voisinage. Ainsi, la présence d'une contrainte sur la proximité des items amène un relâchement de la contrainte nécessaire sur $rang_{\times}$. Or, dans une approche locale du discours, on s'intéresse logiquement à des liens entre items proches dans le texte ; dans notre exemple de situation, il s'agit même de liens intra-phrastiques. Les figures 4.21 et 4.22 montrent l'évolution de la précision et du rappel en fonction de la valeur de $rang_{\times}$ en ne considérant que les liens internes à un paragraphe ($copr_{para}$ = vrai), puis intra-phrastiques ($copr_{ph}$ = vrai). En comparant ces graphiques au graphique 4.10, on observe clairement un relâchement de la limite nécessaire sur $rang_{\times}$ au fur et à mesure que l'on augmente la contrainte de proximité.

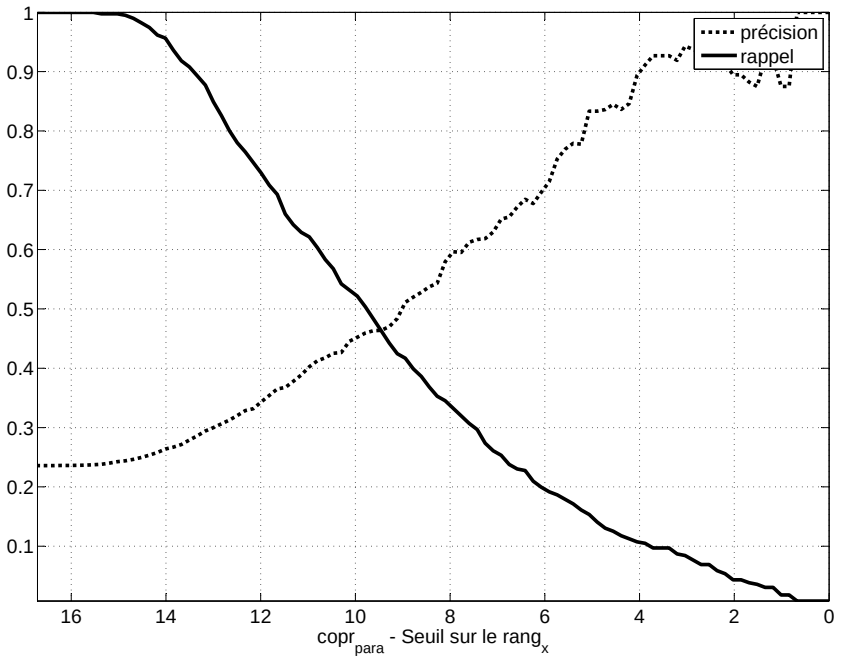


FIGURE 4.21 – Précision et rappel en fonction de $rang_{\times}$ pour les couples internes à un paragraphe

Finalement, dans une approche locale du discours, les contraintes de proximité imposées par l'approche sont suffisamment fortes pour rendre la question du filtrage des liens projetés plus triviale. Nous avons opté pour un filtrage simple basé uniquement sur les rangs des voisins, de manière concertée avec les catégories morpho-syntaxiques des items considérés.

4.6 Bilan et perspectives

En choisissant de traiter ici la question du filtrage des voisins, nous avons rejeté l'idée de prendre les *Voisins de Wikipédia* « tels quels », en simple entrée de nos

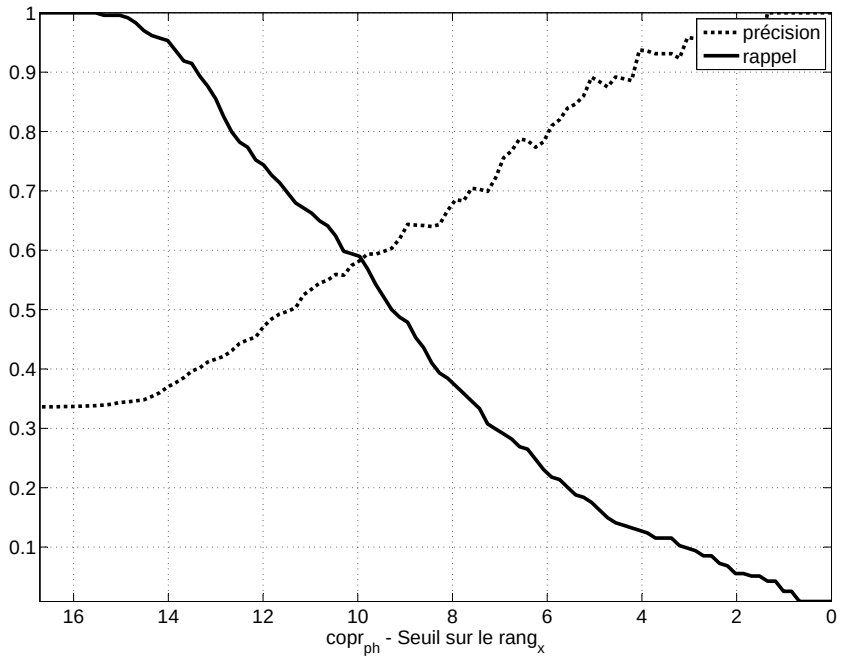


FIGURE 4.22 – Précision et rappel en fonction de $rang_{\times}$ pour les couples intra-phrastiques

traitements. Nous avons considéré que le bruit contenu dans les *Voisins de Wikipédia* était un trop grand obstacle à leur exploitation, et avons donc pris position en faveur d'un filtrage. Nous avons mis en place le dispositif permettant d'examiner l'impact de différents indices sur la qualité de ce filtrage. Nous avons enfin couvert les différents objectifs qui peuvent régir la projection des *Voisins de Wikipédia* et exposé les stratégies de filtrage qui nous paraissaient adaptées dans ces différentes situations.

Conséquences des travaux effectués en termes de retour sur la ressource Les travaux relatés dans ce chapitre apportent une pierre à la question cruciale de l'évaluation et de la description linguistique d'une ressource distributionnelle. Nous avons souligné dans la section 2.1.3 (page 56) les difficultés auxquelles est confrontée l'évaluation d'une ressource distributionnelle : l'évaluation par comparaison à un *gold standard* paraît peu appropriée (puisque l'intérêt de l'analyse distributionnelle est de faire émerger de nouvelles catégories de relations sémantiques) ; l'évaluation par la tâche reste tributaire de la tâche utilisée, et ne permet pas réellement d'enrichir la description linguistique de la ressource évaluée. L'évaluation par confrontation à des jugements humains paraît ainsi une bonne option.

Dans ce sens, l'expérience relatée dans la section 4.2.2 montre qu'une annotation de liens lexicaux en contexte est bien plus adéquate qu'une annotation de couples de mots en dehors de tout contexte. Ainsi, notre démarche inspirée des objectifs du projet VOILADIS paraît également une démarche cohérente si l'on se place dans un objectif de mise en place d'un dispositif permettant l'évaluation d'une ressource distributionnelle et la création d'une référence pour en explorer les caractéristiques linguistiques, ainsi que l'impact des différents paramètres liées à sa construction. L'évaluation par confrontation à des jugements humains posés en contexte apparaît ainsi :

- plus adaptée à la nature d'une ressource distributionnelle qu'une évaluation par comparaison à un *gold standard* ;
- plus générique qu'une évaluation par la tâche (la contextualisation des liens par les textes étant un parti pris moins fort que la « contextualisation » par la tâche) ;
- plus fiable et plus basée sur des réalités linguistiques qu'une évaluation par confrontation à des jugements humains posés hors-contexte.

La prolongation de cette direction de travail impliquerait notamment de revérifier les Kappa obtenus sur une plus large palette d'annotateurs. Pour l'heure, nous avons déjà fourni de premières observations sur l'étude de différentes caractéristiques associées aux couples de voisins, et, *via* l'interface d'annotation programmée, créé une fenêtre d'observation en texte des voisins, qui pourrait être utilisée en complément de l'interface d'interrogation existant déjà (*cf.* section 2.2 page 57).

Des perspectives nombreuses L'entreprise que nous avons lancée – en nous attaquant à la question de la relation entre liens de voisinage et liens de cohésion

lexicale – est vaste. Elle a finalement abouti à des décisions sur le filtrage des liens de voisinage variables en fonction des objectifs. Cela est également dû au fait que cette étape n'a pas été réalisée dans une phase préliminaire, mais que ces objectifs se sont définis dans des aller-retour entre exploitation des liens projetés dans des approches de l'organisation textuelle (partie III) et réflexion sur la ressource et le processus de sa projection en texte.

Mais nous pensons que le chantier que nous avons ouvert permet d'envisager des perspectives plus ambitieuses. Nous avons défini des modalités pour l'annotation des liens lexicaux en contexte, développé une interface d'annotation performante, et mis en place le dispositif permettant de définir des indices pour le filtrage des voisins et de les tester. À partir de cette base, l'annotation de plus de liens et l'exploration de nouveaux indices pourraient aboutir à un filtrage plus satisfaisant et plus indépendant de l'exploitation, réellement intégré à notre chaîne de traitement. Il faudrait alors évaluer plus précisément le rapport entre liens projetés (au terme de cette chaîne de traitement) et cohésion lexicale. Une telle évaluation suppose de mesurer la précision (ce qui peut être fait grâce au dispositif mis en place) mais aussi le rappel :

- pas uniquement par rapport à la performance maximale pouvant être atteinte avec la ressource projetée,
- mais également par rapport à une annotation humaine de liens de cohésion lexicale, ce qui suppose une tâche différente d'annotation (à petite échelle puisque menée uniquement à des fins d'évaluation).

Troisième partie

De la cohésion lexicale à l'organisation textuelle : éclairages

Chapitre 5

La segmentation thématique

Sommaire

And now for something completely different.
Monty Python's Flying Circus

Dans ce chapitre, nous traitons de l'utilisation des *Voisins de Wikipédia* pour la segmentation thématique. Il s'agit d'un chapitre charnière dans cette thèse ; il nous permet d'effectuer la jonction entre la question de la projection en texte des voisins, et celle de leur exploitation pour différentes approches du discours. La tâche de segmentation thématique nous permet en effet de :

(a) proposer une évaluation "par la tâche" de la méthodologie de projection des voisins décrite dans le chapitre précédent ;

(b) aborder une première approche du discours, certes très rudimentaire, mais qui présente l'avantage d'être essentiellement basée sur le lexique.

5.1 État de l'art et motivations

5.1.1 Définition de la tâche

La tâche de segmentation thématique a pour but de découper un texte en segments consécutifs censés présenter une homogénéité du point de vue de leurs thèmes. Cette tâche est parmi les plus tributaires des phénomènes de cohésion lexicale : les zones thématiques sont principalement définies par l'idée qu'elles se singularisent par le recours à un vocabulaire suffisamment spécifique pour être distinguées des autres. Le recours à la segmentation thématique peut permettre d'améliorer les performances de différentes applications du TAL, notamment la recherche d'information – (Callan *et al.*, 1992), entre autres, ont montré qu'un système de recherche d'information gagne à indexer des unités inférieures au document –, le résumé automatique (Boguraev et Neff, 2000; Brunn *et al.*, 2001; Châar *et al.*, 2004), l'extraction d'informations (Prince et Labadié, 2007), la navigation intra-documentaire, etc.

La segmentation thématique est une approche très empirique du discours, qui présuppose une structuration très rudimentaire des textes, basée sur la juxtaposition linéaire de segments contigus, séparés par des ruptures strictes. Cette vision de l'organisation textuelle, essentiellement dictée par une logique de traitement automatique du discours, est très limitée ; les travaux menés sur le discours laissent entrevoir une organisation plus complexe, supposant différents plans de cohérence (et non le seul plan thématique), et des relations entre segments pouvant être hiérarchiques (et non uniquement linéaires) (Péry-Woodley et Scott, 2006). Ainsi, la segmentation thématique ne prend pas vraiment ses sources dans des travaux sur le discours, mais résulte plutôt de la disponibilité de l'information lexicale dans tous les types de documents, et de l'hypothèse selon laquelle la cohésion lexicale est un fort indice de cohérence thématique (notamment formulée dans Masson (1995); Salton *et al.* (1996); Hearst (1997)).

On peut noter à ce stade que certaines approches de segmentation thématique utilisent, seuls ou en complément de la cohésion lexicale, des marqueurs de changement de rupture thématique (Passonneau et Litman, 1997; Couto *et al.*, 2004) ; ces approches sont toutefois minoritaires. Dans la mesure où nous nous intéressons très spécifiquement à l'apport de la cohésion lexicale, nous nous concentrons ici uniquement sur les approches basées sur cet indice.

De nombreux algorithmes ont été développés pour la segmentation thématique, que l'on peut *grosso modo* regrouper en deux familles (Hernandez, 2004) : (a) ceux qui parcourent linéairement le texte selon une fenêtre d'observation glissante, et procèdent donc de manière ascendante et (b) ceux qui calculent une matrice de similarité pour l'ensemble des unités du texte avant de décider où placer les ruptures, procédant donc de manière descendante (Malioutov et Barzilay, 2006). Les systèmes les plus connus représentant ces deux familles sont d'une part l'algorithme TextTiling de Hearst (1994, 1997), et d'autre part l'algorithme C99 de Choi (2000).

Une alternative à ces approches est de supposer des thèmes sous-jacents qu'il s'agit de détecter : chaque unité du texte considéré est rapportée à un ou plusieurs thèmes, et la segmentation consiste à trouver ces thèmes (Chen *et al.*, 2009; Ferret, 2007). Les thèmes peuvent être prédits par des « *topic models* » (Chen *et al.*, 2009) qui sont associés à des distributions lexicales différentes, ou bien par des associations lexicales calculées à partir des textes, par exemple par un *clustering* en amont (Ferret, 2007). Une fois les thèmes identifiés pour chaque unité de texte, les segments correspondent aux blocs d'unités contiguës partageant le même thème.

Dans la section 5.2.1, nous présentons en détail l'approche de type TextTiling, qui est la plus utilisée et à laquelle nous faisons nous aussi appel.

5.1.2 Informations lexicales utilisées

Plus que les différences entre approches, ce qui nous intéresse ici est la variété des informations lexicales utilisées dans le cadre de la segmentation thématique. Pour mesurer la « force » de la cohésion lexicale entre deux pans de texte, on se base sur le nombre de liens (éventuellement pondérés) qu'entretiennent les unités lexicales qu'ils contiennent. Ces liens peuvent être de natures diverses.

Beaucoup de systèmes de segmentation thématique se cantonnent aux liens de répétition lexicale, c'est-à-dire aux répétitions de formes, de formes tronquées ou de lemmes (Hearst, 1997; Choi, 2000) ; une extension consiste à prendre en compte les répétitions de n-grammes, en leur attribuant un poids plus important (Beeferman *et al.*, 1997). L'inconvénient de ces approches est que les scores sont alors basés sur un nombre très restreint d'occurrences (ce qui provoque des fossés dans la courbe de cohésion lexicale des textes) alors que beaucoup de liens participant à la cohésion sont ignorés.

Pour pallier ce problème, la stratégie la plus couramment employée consiste à faire appel à des connaissances extérieures. Il peut notamment s'agir de :

- relations définies par un thésaurus (Morris et Hirst, 1991; Jobbins et Evett, 1998) ;
- similarités sémantiques calculées à partir d'un dictionnaire (Kozima, 1993) ;
- collocations extraites de corpus (Ferret, 1998, 2002) ;
- relations du réseau sémantique *WordNet* (Stokes *et al.*, 2002) ;
- similarités calculées par « analyse sémantique latente » (*Latent Semantic Analysis*, Landauer *et al.*, 1998) (Choi *et al.*, 2001; Bestgen et Piérard, 2006).

Une autre stratégie consiste à lisser ou enrichir l'information lexicale utilisée de manière endogène :

- en jouant sur la pondération des liens, par exemple par l'utilisation de mesures de *tf·idf* locales (Malioutov et Barzilay, 2006; Dias *et al.*, 2007) ;
- en calculant des associations lexicales à partir du document à segmenter, par exemple par similarité distributionnelle (Ferret, 2006b).

Enfin, des différences peuvent être soulignées dans la façon dont l'information lexicale est modélisée. Plus particulièrement, certaines méthodes de segmentation thématique utilisent, plutôt que de simples liens lexicaux, la construction de chaînes lexicales (Morris et Hirst, 1991; Stokes, 2003; Sitbon et Bellot, 2005), qui permettent de représenter plus directement le thème. À un moment donné du texte, il est alors possible de compter, plutôt que le nombre de liens reliant deux blocs, le nombre de chaînes lexicales actives (*cf.* section 1.3.2 page 38).

5.1.3 Évaluation de la segmentation thématique

Nous avons vu qu'il existe une grande diversité d'approches pour la segmentation thématique. Cette diversité rend plus prégnants encore les problèmes liés à l'évaluation de cette tâche, les nouveaux algorithmes devant être comparés à leurs prédécesseurs. Évaluer un système de segmentation thématique est délicat. De nombreux problèmes sont soulevés, et peuvent *grosso modo* être ramenés à deux questions :

(a) À quelle référence comparer la segmentation effectuée par le système ?

(b) Quel score d'évaluation appliquer pour estimer la différence avec cette segmentation de référence ?

Avant de revenir en détail sur ces deux problèmes, nous notons toutefois l'existence de solutions autres pour l'évaluation de la segmentation thématique. Notamment, certaines recherches évaluent les systèmes de segmentation thématique à travers l'apport qu'ils fournissent aux applications pour lesquelles ils ont été conçus (Bellot et El-Beze, 2001; Prince et Labadié, 2007). Nous avons fait le choix d'évaluer nos propositions pour la segmentation thématique de façon consensuelle, par comparaison avec une segmentation de référence. La segmentation thématique nous offrant elle-même l'occasion d'évaluer « par la tâche » l'adéquation de notre ressource pour la détection de la cohésion lexicale, nous souhaitons l'évaluer pour elle-même et non à travers son apport à une autre tâche. Nous présentons donc les approches

pour la construction d'un matériel de référence (5.1.3.1), ainsi que les différentes mesures existantes (5.1.3.2) dans les deux sous-sections qui suivent.

5.1.3.1 Quelle référence pour la segmentation des textes ?

Jugements humains La première stratégie pour établir une référence pour la segmentation thématique, mise en œuvre par Kozima (1993); Hearst (1997); Passonneau et Litman (1997); Klavans *et al.* (1998); Bestgen et Piérard (2006), consiste à demander à des annotateurs d'effectuer la même tâche que la machine, c'est-à-dire de placer des ruptures thématiques dans des textes qui peuvent être de différentes origines. Cette stratégie est coûteuse et s'avère peu satisfaisante, dans la mesure où tous les auteurs font état d'accords inter-annotateurs très faibles (avec généralement un score Kappa autour de 0.4).

Pour remédier à ce problème, plusieurs travaux ont proposé de déterminer une segmentation de référence à partir d'une concordance suffisante de jugements humains (Passonneau et Litman, 1997; Klavans *et al.*, 1998; Bestgen et Piérard, 2006). Ainsi, Bestgen et Piérard (2006) proposent un « indice global de segmentation », dérivé des annotations d'un ensemble de juges (il correspond à la proportion de juges ayant placé une rupture à un point donné), et montrent que cet indice est plus fiable que chaque juge pris individuellement. Mais Bestgen et Piérard (2006) montrent également que les algorithmes de segmentation ne sont pas capables de reproduire la référence créée par concordance de jugements humains, et notent qu'il est illusoire de demander aux algorithmes de surpasser les juges.

Concaténations de textes La seconde stratégie consiste à fabriquer un matériel d'évaluation en créant un texte artificiel à partir d'extraits de différents textes concaténés entre eux. Cette stratégie a été mise en œuvre par Choi (2000), qui a mis à disposition son matériel de référence, constitué à partir d'une sélection aléatoire de segments de textes appartenant au *Brown corpus* (Kucera et Francis, 1967). Ce matériel de référence a été très utilisé.

Cette stratégie pose un problème évident de circularité puisque, en vue d'évaluer la segmentation thématique, on crée l'objet postulé par cette tâche. Bestgen et Piérard (2006) notent que cette procédure semble justifiée lorsque l'algorithme à évaluer a été développé dans le but de segmenter des séquences continues de textes brefs (c'est notamment le cas de Ponte et Croft (1997); Allan *et al.* (1998), ou encore de Guinaudeau *et al.* (2012) qui travaillent sur des transcriptions d'émissions télévisuelles), mais apparaît beaucoup plus discutable si l'objectif du système est d'identifier des changements de thèmes à l'intérieur des textes. Plusieurs auteurs ont montré que ce matériel artificiel d'évaluation facilite la tâche des systèmes de segmentation thématique, dans la mesure où les séquences de textes accolées appartiennent à des domaines radicalement différents (Moens et Busser, 2001; Ji et Zha, 2003). Par ailleurs, Georgescul *et al.* (2006) montrent que le classement obtenu

en comparant différents algorithmes de segmentation thématique est modifié si l'on utilise des données réelles plutôt que le matériel artificiel créé par Choi (2000).

Malgré ces limites, cette stratégie d'évaluation de la segmentation thématique s'est imposée. Des variations ont toutefois été proposées dans la constitution du matériel de référence, afin de pallier certains défauts de celui proposé par Choi (2000). Ji et Zha (2003) utilisent des séquences de textes extraites de différentes sections d'un même roman ; Ferret (2006a) crée chaque document d'évaluation en entrelaçant des séquences provenant de seulement deux documents ; Dias *et al.* (2007) concatènent des articles de presse traitant tous d'un même objet (le football), etc.

Utiliser la structure préexistante du texte Enfin, une dernière approche très peu utilisée consiste à s'appuyer sur la structuration des textes en sections, en considérant les différentes sections comme les zones thématiques à identifier (approche notamment utilisée par Chen *et al.*, 2009). C'est cette stratégie que nous avons choisie, nous la discutons donc par la suite (section 5.2.2).

5.1.3.2 Quelles mesures pour évaluer la segmentation thématique ?

La figure 5.1 présente un exemple d'une segmentation de référence comparée à 3 hypothèses. Cet exemple permet d'illustrer les cas de :

- faux positif (ajout d'une frontière) : en position 16 pour les hypothèses 1 à 3, ainsi qu'en position 2 pour l'hypothèse 2 et en position 10 pour l'hypothèse 3.
- faux négatif (oubli d'une frontière) : en position 14 pour les hypothèses 1 à 3.
- erreurs légères (frontière de l'hypothèse très proche de la frontière de la référence) : en position 11 pour les hypothèses 1 à 3, ainsi qu'en position 3 pour les hypothèses 1 et 3.

num.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
ref.	|			|			|				|		|			|
hyp. 1	|		|				|			|					|	|
hyp. 2	|	|					|			|					|	|
hyp. 3	|		|				|		|	|					|	|

FIGURE 5.1 – Exemple de segmentation de référence *vs* 3 segmentations hypothétiques à évaluer

Classer nos trois hypothèses fictives n'est pas évident. On peut toutefois affirmer que la première est la meilleure, puisque sa seule différence avec l'hypothèse 2 est une rupture placée en position 3 plutôt que 2 (ce qui se rapproche davantage de la référence), et qu'elle ne se distingue de l'hypothèse 3 que par l'absence du faux positif en position 10. Les mesures utilisées pour la segmentation thématique devraient idéalement permettre de traduire ces différences en des différences de score. Les prochains paragraphes présentent quatre scores utilisés dans ce contexte.

Rappel et Précision Le rappel et la précision, mesures issues de la recherche d'information, sont des standards pour l'évaluation de tâches du traitement automatique des langues. C'est donc assez naturellement qu'elles ont été utilisées pour évaluer les premiers systèmes de segmentation thématique, entre autres Hearst (1994) et Beeferman *et al.* (1997), avec les définitions suivantes :

$$\text{Précision} = \frac{1}{N_{hyp}} \sum_{i=1}^{N_{hyp}} \delta_{ref}\,(i) \tag{5.1}$$

$$\text{Rappel} = \frac{1}{N_{ref}} \sum_{j=1}^{N_{ref}} \delta_{hyp}\,(j) \tag{5.2}$$

ou N_{ref}, respectivement N_{hyp} indiquent le total des segmentations pour la référence et l'hypothèse, et où la fonction $\delta_{ref}\,(i)$, respectivement $\delta_{hyp}\,(j)$ indique la présence d'une rupture à la position i dans la référence, respectivement l'hypothèse :

$$\delta_{ref}\,(i) = \begin{cases} 1 & \text{si } i \in \text{ref} \\ 0 & \text{sinon} \end{cases}$$

Mais l'inadéquation de ces scores pour l'évaluation de la segmentation thématique a aussitôt été soulignée : le rappel et la précision traitent de la même manière les erreurs légères (frontière de l'hypothèse très proche de la frontière de la référence) et les erreurs importantes (ajout ou oubli d'une frontière) (Beeferman *et al.*, 1999).

C'est pourquoi des mesures de comparaison dédiées à la tâche ont été proposées. Dans cette perspective, l'indice P_k de Beeferman *et al.* (1999) s'est longtemps imposé ; en 2002, Pevzner et Hearst (2002) ont proposé un nouvel indice, *WindowDiff*, qui est une adaptation de P_k palliant certains de ses inconvénients. Enfin, Bestgen (2009) suggère d'utiliser pour l'évaluation de la segmentation la distance de Hamming généralisée (Bookstein *et al.*, 2002). Nous décrivons ces différentes mesures ci-après.

Indice Pk Le score P_k est défini comme une réalisation particulière de l'indice P_D introduit par Beeferman *et al.* (1999). Cet indice mesure la probabilité que deux unités de segmentation (phrase ou paragraphe) à une distance k dans le texte soient classifiés différemment, c'est-à-dire appartiennent au même segment uniquement dans la référence ou uniquement dans l'hypothèse :

$$P_k\,(\text{ref}, \text{hyp}) = \sum_{1 \le i \le j \le n} \Pi_k(i-j)\,(\delta_{ref}(i,j) \overline{\oplus} \delta_{ref}(i,j)) \tag{5.3}$$

Dans l'expression ci-dessus, la fonction Π_k représente une fenêtre de largeur k :

$$\Pi_k(\ell) = \begin{cases} 1/k & \text{si } 0 \le \ell \le k \\ 0 & \text{sinon} \end{cases} \tag{5.4}$$

l'expression $\delta_{ref}(i,j)$ indique si deux unités de segmentation i et j appartiennent au même segment :

$$\delta_{ref}(i,j) = \begin{cases} 1 & \text{si } i,j \in \text{ même segment} \\ 0 & \text{sinon} \end{cases}$$

et le symbole $\overline{\oplus}$ dénote l'opération logique XNOR (non-ou exclusif) : son résultat est 1 si et seulement si $\delta_{ref}(i,j) \neq \delta_{hyp}(i,j)$. Le choix classiquement fait pour la largeur k de la fenêtre est la moitié de la longueur moyenne des segments dans le texte.

Comme l'ont fait remarquer Pevzner et Hearst (2002), ce score présente le désavantage de ne pas être symétrique pour les erreurs de faux positifs et de faux négatifs, et ne traite pas forcément de façon adéquate les segments plus courts que k. De plus, ce score est fortement sensible à une variation de la longueur retenue pour k. Ces observations ont servi de motivation au développement du score *WindowDiff* présenté ci-dessous.

WindowDiff Le score *WindowDiff* introduit par Pevzner et Hearst (2002) s'exprime comme :

$$\text{WindowDiff}(\text{ref},\text{hyp}) = \frac{1}{N-k}\sum_{i=1}^{N-k}\left(\left|b_k^{ref}(i) - b_k^{hyp}(i)\right| > 0\right) \tag{5.5}$$

ou la fonction $b_k^{ref}(i)$ compte le nombre de ruptures dans la fenêtre de longueur k débutant en i pour la référence, tandis que la fonction $b_k^{hyp}(i)$ effectue le même compte pour l'hypothèse. *WindowDiff* ne souffre pas des défauts du score P_k concernant les petits segments ou les faux négatifs. Toutefois, Pevzner et Hearst (2002) notent que *WindowDiff* ne résout pas totalement les problèmes du score P_k, en particulier en ce qui concerne la forte dépendance à une variation de la longueur retenue pour k. Dans la pratique, *WindowDiff* est souvent identique à P_k, sauf pour les hypohtèses contenant beaucoup de faux positifs. Ceci conduit Bestgen (2009) à proposer un indice basé sur la distance de Hamming généralisée.

Distance de Hamming généralisée Bestgen (2009) propose d'employer la distance de Hamming généralisée, une forme spécialisée de distance d'édition, pour mesurer la qualité des hypothèses de segmentation. Afin de correspondre à cette tâche, les coût des opérations d'édition sont définis de la façon suivante :

- déplacement : coût 1
- insertion : coût $k/2$
- suppression : coût $k/2$

Ce choix de coûts permet en particulier d'assurer que les déplacements sont préférés aux insertions ou suppressions pour des distances inférieures à k, ce qui permet à ce score de moins pénaliser les erreurs légères que les mesures classiques que sont

précision et rappel. Ce score est aussi symétrique en ce qui concerne les pénalités apportées par chaque faux positif et faux négatif.

Le tableau 5.1 illustre les scores introduits ci-dessus pour les hypothèses de segmentation de l'exemple de la figure 5.1.

Score	hyp. 1	hyp. 2	hyp. 3
Précision	0.25	0.25	0.20
Rappel	0.25	0.25	0.25
P_k	0.40	0.40	0.47
WindowDiff	0.40	0.40	0.47
Distance de Hamming généralisée (*edit*)	1.56	1.68	1.92

TABLEAU 5.1 – Scores pour les hypothèses de segmentation de l'exemple de la figure 5.1

Sur notre exemple, les mesures P_k et *WindowDiff* aboutissent à des résultats identiques. Les trois mesures sont relativement cohérentes entre elles, avec le classement hyp. 1 $\leq$ hyp. 2 $<$ hyp. 3. La distance de Hamming généralisée est la seule à exprimer la meilleure adéquation de l'hypothèse 1 avec la référence.

5.1.4 Motivations pour travailler sur la segmentation thématique

Comme nous l'avons souligné, le projet VOILADIS a pour objectif premier d'explorer le niveau lexical dans les structures discursives identifiées lors du projet ANNODIS, en utilisant les *Voisins de Wikipédia*. L'organisation thématique des documents se situe en dehors de ce premier objectif, puisqu'elle ne fait pas partie des problématiques du projet ANNODIS. Ainsi, nos motivations pour aborder cette « approche du discours » ne découlent pas de la disponibilité de données annotées, comme pour les structures étudiées dans les chapitres suivants.

Si nous nous sommes intéressée à la segmentation thématique, c'est parce que cette tâche, comme nous l'avons vu, est essentiellement basée sur le lexique ; il s'agit de la seule approche automatique de l'organisation textuelle qui puisse reposer uniquement sur la prise en compte d'indices lexicaux ; la problématique de l'organisation thématique telle que posée par la segmentation thématique apparaît ainsi comme un bon point d'entrée – voire comme un passage obligé – pour aborder la question de l'exploitation de la cohésion lexicale. D'autre part, il s'agit d'une tâche de TAL assez bien définie et facile à mettre en œuvre, dans la mesure où les algorithmes les plus couramment utilisés sont relativement simples et clairement décrits dans la littérature.

Ainsi, mettre en place un système de segmentation thématique capable d'utiliser différents types de ressources lexicales extérieures apparaît comme une solution efficace pour mener une évaluation par la tâche de la détection de la cohésion lexicale

résultant de la projection de notre ressource distributionnelle. Cette évaluation par la tâche repose sur l'hypothèse suivante : meilleure est la détection de la cohésion lexicale, meilleures sont les performances du système de segmentation thématique [1]. Il est alors possible de mesurer l'effet du filtrage des voisins proposé dans le chapitre 4, de comparer les résultats obtenus grâce aux voisins distributionnels avec ceux obtenus par la projection d'une autre ressource, etc.

Nous abordons donc la segmentation thématique non pas pour cette tâche en elle-même (nous n'avons pas pour objectif de créer un système de segmentation thématique plus performant que l'état de l'art), mais dans la perspective des objectifs du projet VOILADIS : notre but est de montrer l'apport d'une base de voisins distributionnels à la détection de l'organisation textuelle à travers cette tâche. Cette perspective particulière a des conséquences sur la façon dont nous avons appréhendé la tâche de segmentation thématique. Notamment, nous avons pris le parti de ne pas mener d'évaluation sur un matériel artificiel obtenu par concaténation de textes, car ce choix aurait évacué la problématique de l'« organisation textuelle ».

Dans la section 5.2, nous décrivons le système de segmentation thématique développé (5.2.1) et les choix que nous avons faits concernant l'évaluation de la segmentation automatique (5.2.2). Nous présentons ensuite une première expérience.

5.2 Aborder la segmentation thématique : système développé et expériences préliminaires

Dans cette section, nous décrivons le système de segmentation thématique que nous avons développé et utilisé pour les différentes expériences réalisées et décrites dans ce chapitre (section 5.2.1). Nous discutons ensuite des choix que nous avons faits concernant l'évaluation de la segmentation thématique (section 5.2.2), notamment concernant le choix du corpus et de la segmentation de référence. Enfin, nous présentons une première expérience de segmentation, qui nous a permis de montrer la sensibilité de la tâche de segmentation thématique aux types de textes (section 5.2.3).

5.2.1 Système de segmentation thématique développé

Le système de segmentation thématique que nous avons développé utilise l'algorithme TextTiling décrit par (Hearst, 1997). Cette approche n'est pas forcément la plus performante, mais nous l'avons préférée en raison de sa simplicité d'implémentation ; en effet, nous ne poursuivons pas ici l'efficacité, mais l'exploration de l'impact de l'utilisation des voisins distributionnels sur la tâche de segmentation thématique.

1. Nous verrons au terme de ce chapitre que cette hypothèse doit être au moins mitigée.

Nous donnons ci-dessous le détail de la chaîne de traitement :

1. Les liens de voisinage, éventuellement pondérés, sont projetés sur les textes ; les répétitions sont également repérées, et dotées d'un score de 1.
2. Le texte est parcouru par une fenêtre glissante, afin de calculer localement des scores de cohésion. L'unité de segmentation, ainsi que la taille de la fenêtre en nombre d'unités, sont paramétrables. Les unités de segmentation possibles sont : (i) la phrase ; (ii) le bloc de mots de taille fixe. Par exemple, si l'unité choisie est la phrase, et que la taille de la fenêtre est fixée à 6, on calculera à la fin de chaque phrase un score basé sur le nombre de liens entretenus par le groupe de trois phrases qui précède, et le groupe de trois phrases qui suit (fig. 5.2).

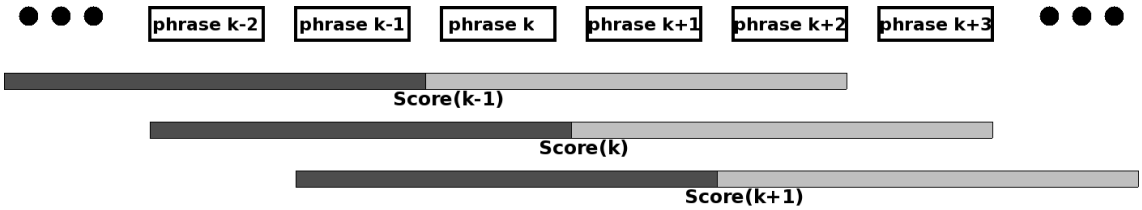


FIGURE 5.2 – Représentation de la fenêtre glissante avec -fen=3 et -unit=phrase

3. Le score est calculé pour chaque fenêtre. Soit w la taille de la fenêtre utilisée (en nombre de mots). Cette fenêtre se décompose en deux sous-fenêtres w_g et w_d telles que $w_g + w_d = w$. Soit N_{gd} le nombre de liens entre les sous-fenêtres gauche et droite. Le score S est calculé comme le rapport de ce nombre de liens au nombre maximum de liens possibles $w_g \cdot w_d$. Afin de garder une dynamique plus restreinte, une fonction log est appliquée à ce ratio :

$$S = \log\left(\frac{N_{gd}}{w_g \cdot w_d}\right) \tag{5.6}$$

4. La courbe des scores obtenue est lissée : la courbe brute étant fortement bruitée (présentant de fortes variations rapides et non-significatives), un filtrage est opéré afin d'éliminer les variations aberrantes ; nous avons opté pour un lissage gaussien (Wells, 1986), avec deux paramètres ajustables : le nombre d'itérations et le degré du lissage. La figure 5.3 présente les courbes brute et lissée pour l'article Wikipédia « Bulgarie ». Les barres verticales indiquent les positions de titres de section, que nous considérerons par la suite comme les ruptures de référence (*cf.* section 5.2.2).

5. Dans la courbe de cohésion lexicale, toutes les vallées sont repérées, et leurs profondeurs calculées. On appelle vallée un point de la courbe qui est entouré par des points de valeurs plus élevées (c'est-à-dire un minimum local). Pour déterminer la profondeur d'une vallée, on remonte de part et d'autre du point considéré

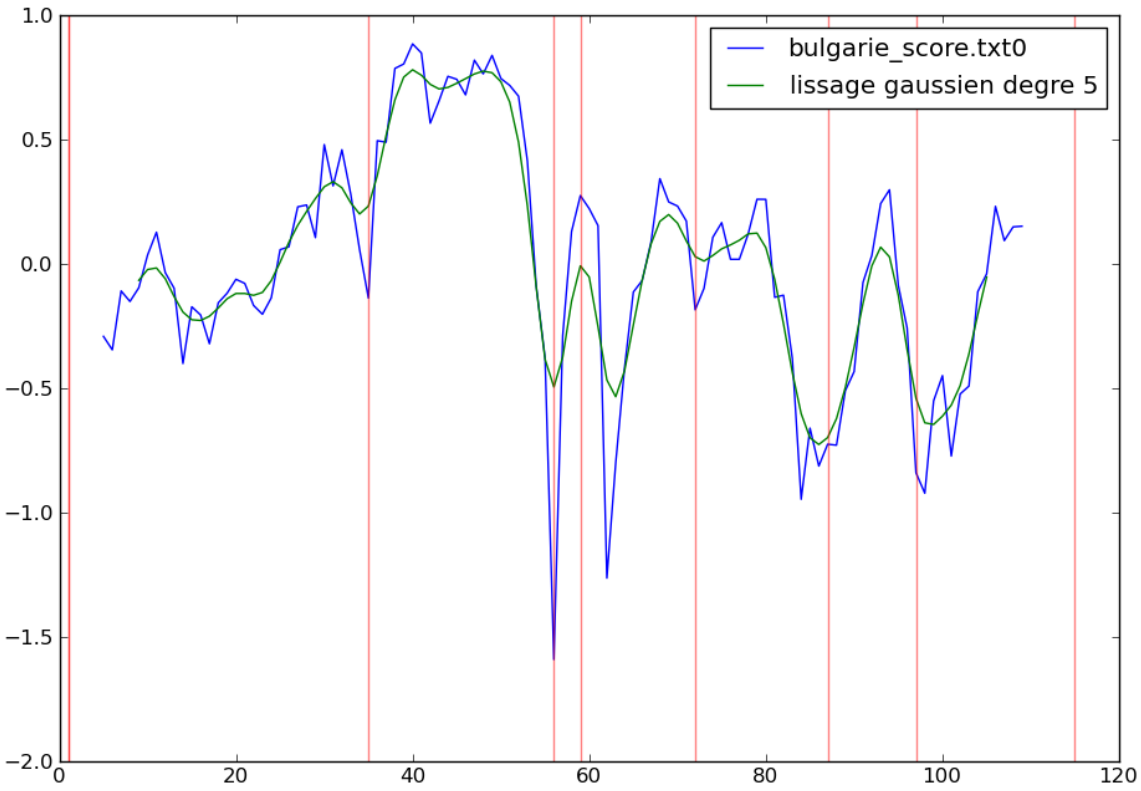


FIGURE 5.3 – Exemple de courbe avec lissage

tant que l'on rencontre des valeurs plus élevées ; la profondeur de la vallée est calculée en faisant la moyenne des deux différences calculées (à gauche et à droite du point). Les vallées dont la profondeur dépasse l'écart type à la moyenne sont considérées comme correspondant aux ruptures du texte.

5. Les ruptures, selon l'unité choisie, se trouvent dans le meilleur des cas à la frontière d'une phrase, mais peuvent également intervenir en plein milieu d'une phrase, puisque les blocs de mots comparés ne tiennent pas forcément compte de la structure des textes. De plus, elles ne sont pas posées avec précision, en raison notamment du lissage de la courbe. C'est pourquoi les points de ruptures sont ajustés pour correspondre à des unités interprétables des textes. Les ruptures calculées sont ainsi ramenées à la frontière de paragraphe la plus proche, qui peut se situer avant ou après le point de rupture, comme préconisé par Hearst (1997). Le paragraphe est en effet généralement considéré comme une unité thématiquement homogène (Masson, 1995; Salton *et al.*, 1996) ; d'autre part, dans le référentiel que nous utilisons pour évaluer la segmentation thématique, les ruptures interviennent toujours entre deux paragraphes. Pour une discussion du rapport entre paragraphe, cohésion lexicale et cohérence thématique, on peut se reporter à Hernandez (2004, pp. 194-197).

On constate que plusieurs paramètres sont ajustables dans notre système ; nous

les récapitulons dans le tableau 5.2. Leur optimisation sur un corpus de développement est décrite dans la section 5.2.3.

Paramètre	Description	Valeurs
unité	unité de segmentation	phrase / bloc
bloc	taille du bloc	<nb mots>
fenêtre	taille de la demi-fenêtre glissante	<nb blocs>
itérations	nombre d'itérations du lissage	<nb>
degré	degré du lissage	<nb>

TABLEAU 5.2 – Paramètres de notre système de ST

5.2.2 Décisions prises pour l'évaluation

Pour évaluer notre système de segmentation thématique, nous avons souhaité respecter les contraintes suivantes :

- l'utilisation de textes réels et non un matériel d'évaluation artificiel ; en effet, nous nous intéressons à la segmentation thématique dans la mesure où elle représente une approche de l'organisation textuelle exploitant des indices lexicaux ; cette perspective nous fait exclure une évaluation basée sur des pseudo-textes ;
- l'utilisation de textes extraits du corpus WikipédiaFR2007, pour des raisons déjà mentionnées (dans la section 1.5.2.2 page 47) : nous avons, dans le cadre du projet VOILADIS, préféré circonscrire l'exploitation des voisins distributionnels au corpus à partir duquel ils ont été calculés.

À ces deux contraintes théoriques s'ajoute une contrainte pratique : la mise en place d'une annotation manuelle est très coûteuse, surtout si l'on considère qu'une annotation fiable ne peut être établie qu'à partir de la concordance d'un nombre important de jugements humains (par exemple, Bestgen et Piérard (2006) fait appel aux jugements de 15 annotateurs différents).

En conséquence de ces contraintes, nous avons décidé d'exploiter le fait que les textes issus de l'encyclopédie Wikipédia ont généralement une structure explicite, marquée par des titres de section, en faisant le choix de considérer les positions des titres comme les ruptures thématiques de référence.

Lors d'une première expérience relatée dans Adam et Morlane-Hondère (2009), nous avons constitué un sous-corpus d'évaluation en sélectionnant dans WikipédiaFR2007 des articles ayant pour sujet des lieux – pays (par exemple l'article « Danemark ») ou villes (par exemple « Salzbourg »). En effet, nos observations montraient que dans cette catégorie d'articles, les différentes sections correspondaient généralement à différents « thèmes » (« histoire », « géographie », « culture », etc.), ce qui permettait d'appuyer le choix des titres comme ruptures de référence. Travailler sur des textes qui se prêtent intuitivement à la segmentation thématique est

courant (par exemple, Chen *et al.* (2009) utilisent également des articles encyclopédiques sur les villes), sans qu'on cherche pour autant à déterminer la nature des textes qui sont adaptés à la tâche.

En effet, il ne va pas du tout de soi que l'on puisse aborder par les mêmes méthodes de segmentation des textes organisés thématiquement, rhétoriquement, temporellement, voire par une combinaison de ces modes, et que les indices lexicaux soient toujours discriminants pour placer des ruptures entre segments textuels. Les tableaux 5.3 et 5.4 donnent des exemples de textes tirés de Wikipédia de manière à illustrer cette diversité des modes d'organisation. La liste des titres de section de premier niveau donne un bon aperçu de la façon dont le texte s'organise. Le tableau 5.3 montre des textes dont l'organisation thématique est manifeste.

Le Malawi	Le panda géant
- Histoire	- Historique
- Politique	- Légende
- Subdivisions	- Alimentation
- Géographie	- Reproduction
- Économie	- Protection
- Démographie	
- Culture	

TABLEAU 5.3 – Exemples d'organisation textuelle thématique

Le tableau 5.4 montre d'abord un exemple d'organisation temporelle, typique des biographies, qui se clôt par une partie bilan. Les deux textes montrés par le tableau 5.5 (« leadership » et « mythe ») illustrent un mode de progression rhétorique qui permet dans ces deux cas d'organiser la présentation d'une notion selon un schéma argumentatif similaire : d'abord définir la notion, puis présenter une typologie, enfin détailler certaines de ses instances.

Laurent Truguet
- Jeunesse jusqu'à la Révolution
- sous la Révolution
- L'Empire
- sous la Monarchie
- Le bilan

TABLEAU 5.4 – Exemples d'organisation textuelle non thématique (temporelle)

Dans la section suivante, nous reprenons une expérience décrite dans un article intitulé « Une évaluation de l'impact des types de texte sur la tâche de segmentation thématique », présenté à la conférence TALN 2010 (Adam *et al.*, 2010). Le but de cette expérience est de montrer, à partir de textes de Wikipédia (appartenant donc à un même genre : l'article encyclopédique), que la segmentation thématique ne peut

Leadership	Mythe
- Terminologie	- Définition
- Types de leadership	- Aspects du mythe
- Caractéristiques du leadership	- Typologie et éléments du mythe
- Le leadership de droit et de fait	- Postérité du mythe
- Le paradigme des leaderships multiples	

TABLEAU 5.5 – Exemples d'organisation textuelle non thématique (rhétorique)

pas s'appliquer à n'importe quel type de texte. Elle permet ainsi de nous guider dans la sélection de textes pour la constitution d'un corpus dédié à la segmentation thématique.

À travers cette expérience préliminaire, outre la sensibilité de la segmentation thématique aux types de textes, nous avons également testé l'apport des voisins distributionnels à la tâche, et les différentes mesures d'évaluation de la segmentation thématique. Nous expliquons également comment nous optimisons les paramètres de notre système à partir d'un corpus de développement.

5.2.3 Choix du corpus : une évaluation de l'impact des types de texte sur la tâche de segmentation thématique

L'objectif de l'expérience relatée dans cette section est d'intégrer le paramètre du type de textes dans la tâche de segmentation en comparant les performances d'un système de segmentation thématique sur deux groupes de textes, déterminés selon leur propension à s'organiser plutôt thématiquement ou à obéir à d'autres principes de présentation - rhétorique, temporelle. L'hypothèse que nous souhaitons valider par cette expérience est que le recours à la segmentation thématique se justifie pour des textes dont la structuration est effectivement ressentie comme thématique, mais n'est pas motivé pour aborder d'autres modes d'organisation textuelle. Partant, nous voulons inciter à mieux définir quel peut être l'objet de la tâche de ST, et à ne pas appliquer cette tâche sans précaution à des textes tout-venant.

5.2.3.1 Procédure

Caractérisation du corpus utilisé Nous avons pour cette expérience constitué deux sous-corpus à partir de la version d'avril 2007 de l'encyclopédie en ligne Wikipédia. Les articles ont été extraits de manière automatique, sur la base de critères fixés par nous. Les critères de sélection sont les suivants : pour être retenu, un article doit avoir au minimum 1000 mots, au moins 4 titres de sections (qui fournissent la segmentation de référence), et un maximum de 2 niveaux de profondeur de titres (une profondeur trop importante aurait amené à faire des choix délicats quant aux titres retenus pour la segmentation de référence) ; il doit également appartenir à une liste de catégories thématiques établie. Nous avons en effet pris le niveau des caté-

gories définies dans l'encyclopédie comme critère de répartition des textes dans les deux sous-corpus. Le sous-corpus à organisation thématique forte (corpus THEM) rassemble des textes consacrés à la description de pays, de villes et d'animaux, dont on sait qu'ils se prêtent généralement bien à une organisation thématique. Le sous-corpus à organisation thématique faible (corpus NON-THEM) réunit des biographies, dont l'organisation est typiquement temporelle, et des textes présentant des notions abstraites, des concepts, pour lesquels nous avons montré (tab. 5.4) que l'approche thématique est généralement mal adaptée. L'intervention humaine se concentre donc en amont de la constitution du corpus, par la définition des critères de sélection et de répartition dans les sous-corpus. Aucun traitement n'est effectué en aval (post-sélection, nettoyage, etc.). La caractérisation des corpus obtenus est donnée dans le tableau 5.6. Nous précisons pour chaque sous-corpus le nombre de paragraphes, qui correspond au nombre de segments potentiels pour notre système de ST et le nombre de titres de premier niveau, c'est-à-dire le nombre de segments effectifs de notre segmentation de référence.

	corpus THEM	corpus NON-THEM
nb textes	344	210
nb mots	578346	387941
nb para.	10953	6454
nb seg. (titre niv=1)	3051	1182
nb seg/txt	8.869	5.629
nb seg/para	0.279	0.183

TABLEAU 5.6 – Caractérisation des corpus THEM et NON-THEM

Corpus de développement et optimisation des paramètres La procédure de segmentation décrite section 5.2.1 dépend de nombreux paramètres dont les conséquences ne sont pas toujours prédictibles *a priori*. Beaucoup d'auteurs fixent des paramètres similaires selon des critères empiriques pas toujours explicites. Nous avons choisi d'isoler une partie du corpus de départ pour l'utiliser comme un corpus de développement, sur lequel nous avons fait varier un certain nombre de paramètres afin d'ajuster la segmentation. Pour cela, nous avons extrait au hasard un peu moins de 10% du corpus rassemblé initialement, en prenant autant (*ie* 21) de textes des sous-corpus THEM et NON-THEM (rappelons que le corpus initial n'est pas tout à fait équilibré ; nous avons équilibré celui de développement pour ne pas biaiser vers la classe majoritaire). Les variations faites sur les 5 paramètres ont généré plus de 1000 configurations, dont certaines donnaient des résultats très proches. Nous avons conservé la configuration ayant obtenu les meilleurs résultats selon l'indice *WindowDiff*, le plus couramment utilisé ; elle est donnée dans le tableau 5.7.

unité	bloc	fenêtre	itérations	degré
bloc	10 mots	10 blocs	2	3

TABLEAU 5.7 – Configuration de paramètres retenue

5.2.3.2 Résultats

Nous avons appliqué notre système de ST, avec la configuration de paramètres optimisée sur le corpus de développement, sur les deux sous-corpus THEM et NON-THEM. Les tables 5.8 et 5.9 synthétisent les résultats, avec pour chaque mesure l'indication de l'écart type. Nous donnons deux résultats pour notre système, selon les types de liens de cohésion lexicale pris en compte : répétitions simples de lemmes, ou voisinage distributionnel. Nous testons ainsi une première fois l'apport du voisinage distributionnel à la tâche. Dans cette expérience, nous avons pondéré les paires de voisins par la valeur de leur Lin, en attribuant par ailleurs un score de 1 aux liens de répétition.

Nous avons généré deux *baselines* simplifiées qui permettent de donner une idée des écarts que l'on peut avoir sur les mesures P_k et *WindowDiff*, qui ne sont pas nécessairement simples à interpréter. La première *baseline* (nommée plus bas « hasard exact ») place des ruptures au hasard, mais en nombre correspondant à la référence. Elle permet de contrôler la facilité qu'il y a à se rapprocher des vraies bornes par rapport au nombre moyen de segments rapporté à la taille du texte. Une deuxième *baseline* proche consiste à perturber le nombre exact de ruptures, en le faisant varier au hasard dans un intervalle de 30% du vrai nombre de ruptures.

Une autre indication de la représentativité des scores peut être prise dans la littérature, même si la variété des approches, des entrées et des évaluations (vrais textes ou concaténations artificielles) doit inciter à la prudence. Si l'on se réfère à Chen *et al.* (2009), qui opère sur un corpus similaire à une partie du nôtre (les articles de villes dans le Wikipédia anglais), l'état de l'art précédent représenté par Eisenstein et Barzilay (2008) atteint un P_k de 0.317 et un *WindowDiff* de 0.376 sur ce corpus, avec la connaissance du nombre de segments ; l'approche de Chen *et al.* (2009) à base de *topic models* enrichis de contraintes globales atteint quant à elle sur leur meilleure configuration les très bons scores de 0.28 pour le P_k et de 0.25 pour le *WindowDiff*, sans connaissance du nombre de segments, mais en posant une borne supérieure sur le nombre de thèmes présents dans tout le corpus (fixée à 10 ou 20 thèmes), ce qui limite un peu la généralisation.

Pour évaluer les différences entre méthodes, peu d'auteurs reportent des résultats de significativité statistique. Nous avons pour notre part fait le test des rangs signés de Wilcoxon entre les séries de scores des *baselines* et des méthodes par cohésion, appariées par texte, test qui ne suppose rien sur la distribution *a priori* des scores.

On constate que l'hypothèse globale d'une différence entre les deux types de textes THEM et NON-THEM se vérifie assez nettement, quelle que soit la métrique considérée, et que les algorithmes de segmentation choisis sont meilleurs sur les textes

Méthode	P_k	*WD*	*edit*	nb seg/txt
référence	0	0	0	7.89
baseline "hasard bruité"	0.3659	0.3738	1.6492	9.46
baseline "hasard exact"	0.3417	0.3452	**1.5789**	7.89
répétitions	0.3114	0.3144	1.5907	4.93
voisins	**0.3091**	**0.3129**	1.5837	5.09

TABLEAU 5.8 – Résultats pour le sous-corpus THEM

Méthode	P_k	*WD*	*edit*	nb seg/txt
référence	0	0	0	8.07
baseline "hasard bruité"	0.3569	0.3616	1.8032	6.68
baseline "hasard exact"	**0.3149**	**0.3181**	**1.5645**	8.07
répétitions	0.3612	0.3662	1.8846	5.08
voisins	0.3613	0.3676	1.9291	5.16

TABLEAU 5.9 – Résultats pour le sous-corpus NON-THEM

du sous-corpus THEM, même si les variances (non rapportées dans le tableau) sont importantes. Par ailleurs, alors qu'on observe des valeurs de $p<0.01$ pour la différence entre les *baselines* et les algorithmes de segmentation sur l'expérience THEM, la différence n'est pas significative pour les textes du sous-corpus NON-THEM (et on constate que les variances sont plus fortes).

Étant donnée la forte variance que nous avons observée sur les résultats, y compris sur le sous-corpus THEM, nous avons évalué les résultats par catégories Wikipédia à l'intérieur des corpus THEM et NON-THEM. Le récapitulatif de la meilleure méthode par catégories de textes se trouve table 5.10. On constate que les résultats

thème	P_k	*WD*	*edit*	nb par./seg.	nb par.	nb textes
animaux	0.2794	0.2803	1.4076	0.1651	25.5724	145
pays	0.3443	0.3472	1.7695	0.1223	40.7353	136
concepts	0.3488	0.3510	1.8377	0.1632	30.4706	68
villes	0.3508	0.3541	1.7610	0.1348	31.1250	32
personnes	0.3738	0.3777	1.9062	0.1778	26.6132	106
autres NON-THEM	0.4041	0.4112	1.7337	0.1655	31.7333	15

TABLEAU 5.10 – Résultats par sous-catégories par P_k / *WindowDiff* croissant

des deux sous-corpus se reportent sur les catégories qui les composent, à l'exception de la catégorie *concepts* qui obtient des résultats légèrement meilleurs que ceux de la catégorie *villes*. Encore une fois les variances sont fortes. Il s'avère que notre découpage volontairement grossier *a priori* (dans un souci de ne pas trop biaiser l'étude) pourrait s'affiner – à condition de poser clairement les paramètres de ce que nous avons appelé pour l'instant le caractère thématique fort ou faible des textes –, mais

qu'il semble valide.

Rôle des voisins distributionnels Cette expérience, au delà de la question des types de texte, nous a permis de tester une première fois l'apport des voisins distributionnels à la tâche de segmentation thématique. Pour cette expérience, les *Voisins de Wikipédia* n'ont pas été filtrés, mais chaque lien a été pondéré en fonction des scores de Lin. Concernant cet aspect, on ne peut que constater la proximité des scores sur les deux sous-corpus (et le test de Wilcoxon n'indique pas de différence significative). Dans la section suivante, nous testons plus spécifiquement le rôle des voisins distributionnels pour cette tâche, en appliquant notamment la stratégie de filtrage présentée dans la section 4.5.2.

5.3 Voisinage distributionnel, cohésion lexicale et segmentation thématique

Cette section est consacrée à l'exploration des relations entre voisinage distributionnel, cohésion lexicale et segmentation thématique. Nous y interrogeons la capacité des voisins à détecter des liens de cohésion lexicale, et la relation entre la cohésion lexicale détectée et la cohérence thématique des textes.

5.3.1 Méthodologie

Afin d'évaluer l'apport des voisins distributionnels à la tâche de segmentation thématique, nous utilisons le corpus THEM présenté dans la section 5.2.3.1. Le système de segmentation thématique est configuré suivant les paramètres listés dans le tableau 5.7, fixés après une phase d'optimisation sur un corpus de développement.

Dans un premier temps (section 5.3.2), nous proposons une évaluation de la stratégie de filtrage des voisins distributionnels proposée dans la section 4.5 page 130. Dans un second temps, nous proposons une évaluation de l'apport des voisins distributionnels par rapport à d'autres types d'informations lexicales (section 5.3.3.2). La figure 5.4 résume les différentes étapes menant à la segmentation des textes : extraction automatique du corpus THEM (sur des critères structurels et basés sur les catégories Wikipédia), prétraitement et projection des voisins, et enfin segmentation en suivant la procédure décrite dans la section 5.2.1.

Dans les deux cas, nous présentons des exemples de projection de liens lexicaux illustrant les différentes méthodes avant de présenter les performances obtenues par le système de segmentation thématique. En effet, ces différences de performances, d'ailleurs relativement faibles, restent difficiles à interpréter et ne permettent pas vraiment d'apprécier les différences induites par les diverses approches de la cohésion lexicale : quantité de liens lexicaux repérés, variété de ces liens, importance du bruit, etc.

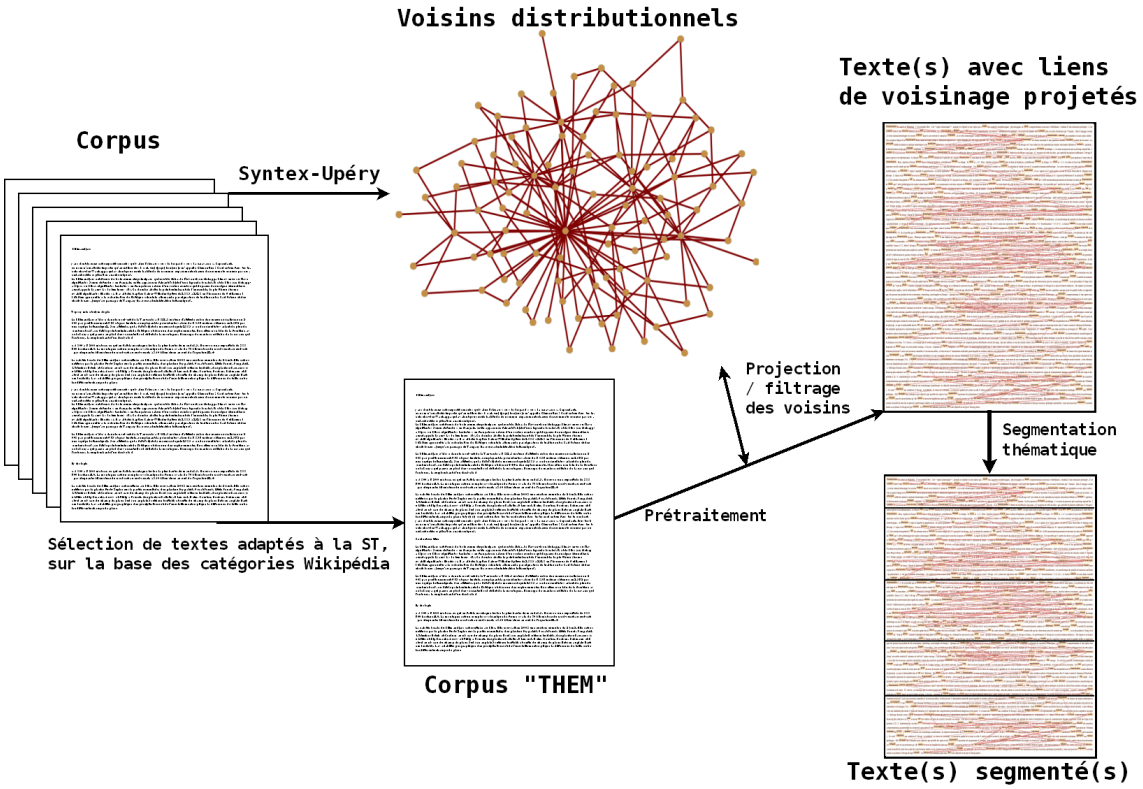


FIGURE 5.4 – Projection des voisins pour la segmentation thématique

Enfin, nous apportons quelques éléments de mise en perspective des résultats, en discutant notamment de l'hypothèse selon laquelle une meilleure détection de la cohésion lexicale devrait conduire à de meilleures performances pour un système de segmentation thématique (section 5.3.4).

5.3.2 Évaluation de la stratégie de filtrage des voisins distributionnels

Dans cette section, nous comparons les liens lexicaux mis au jour par la projection des voisins selon qu'elle est suivie ou non d'une phase de filtrage telle que décrite en 4.5.2.1 : nous utilisons les probabilités que les liens soient pertinents calculées par un système d'apprentissage automatique ; lorsque celle-ci dépasse 0.3 le lien est conservé et pondéré par sa probabilité.

5.3.2.1 Filtrage des voisins et détection de la cohésion lexicale : exemples

Nous présentons ci-dessous deux exemples d'extraits d'articles de Wikipédia afin d'apprécier l'effet du filtrage mis en place sur les liens de voisinage projetés. Les figures 5.5 et 5.6 page 167 sont construites à partir de l'exemple (1), tiré de l'article

« Girafe ». Les figures 5.7 et 5.8 page 168 sont quant à elles construites à partir de l'exemple (2), tiré de l'article « Albanie ». Le tableau 5.11 page 166 résume les liens rejetés par le filtrage dans le cadre de ces deux exemples.

Ces exemples sont tous deux relativement longs (plusieurs paragraphes). En effet, on a vu que la pertinence d'un couple de voisins, selon la référence ayant servi à mettre en place le système de filtrage, est fortement corrélée avec la proximité de ses réalisations dans le texte. Il est ainsi difficile d'apprécier l'effet du filtrage sur des exemples très courts, puisque le nombre de liens filtrés est alors très faible. Dans les exemples présentés, la proportion de paires rejetées suite au filtrage reste bien inférieure à la proportion observée à l'échelle d'un texte entier. Par ailleurs, dans la mesure où le système de segmentation thématique que nous utilisons prend comme unité thématique minimale le paragraphe, il est plus intéressant d'observer les liens projetés à l'échelle de plusieurs paragraphes.

(1) Pour la NASA, étant le plus « haut » des animaux, elle est le modèle idéal pour étudier l'effet de la gravité sur la circulation. Les phlébologues de la NASA ont copié son réseau sanguin pour réaliser la combinaison anti-G des pilotes de chasse et astronautes.
Son cœur de 11 kg pompe 60 litres de sang par minute. Dans les artères du cou, tout un réseau de muscles annulaires aident à hisser le sang jusqu'au cerveau. Dans les veines, les valvules, véritables soupapes, orientent le sang vers le cœur.
Lorsque l'animal baisse la tête au sol, les valves de la jugulaire sont fonctionnelles et empêchent le sang de retomber vers le cerveau (ce qui conduirait à un « voile rouge »).
La veine jugulaire de la girafe est la plus longue et la plus droite du monde animal et possède 9 valves. En 1993, à Vincennes, son endoscopie confirma que les constituants anatomiques d'une veine sont orientés en fonction de son axe d'aplatissement et donc qu'une veine a bien deux faces et deux bords. En bas des jambes où la pression est énorme, un système de capillaires sanguins très résistants, comparables aux nôtres, empêche l' œdème fatal. (Wikipédia, article « Girafe »)

(2) L'Albanie est un pays montagneux (70 %), dont le point culminant s'élève à 2753 m (mont Korab). Le reste est constitué de plaines alluviales, dont le terrain est plutôt de piètre qualité pour l'agriculture, alternativement inondé ou desséché. Les terres les plus fertiles sont situées dans le district des lacs (lac d'Ohrid, Grand Prespa et Petit Prespa) et sur certains plateaux intermédiaires entre la plaine et la montagne. La seule île notable est celle de Sazan qui fut tour à tour occupée par diverses grandes puissances européennes.
Le plus grand fleuve albanais est la Drini. Long de 282 km, elle est un des seuls à connaître un débit relativement stable tout au long de l'année. Les autres cours d'eau sont généralement presque secs durant l'été, même les rivières Semani et Vjosa qui ont pourtant une longueur de plus de 160 km.
Le climat y est méditerranéen dans les régions littorales (moyenne hivernale : 7°), et devient plus continental dans le relief . Les précipitations sont assez élevées (1 000 à 1 500 mm annuels), le flux d'air humide rencontrant la masse d'air continentale plus froide, surtout pendant l'hiver, qui est la saison pluvieuse. (Wikipédia, article

« Albanie »)

Concernant le filtrage de la ressource, on peut remarquer que les liens rejetés suite à la phase de filtrage (reportés dans le tableau 5.11) paraissent globalement peu pertinents dans le contexte des extraits où ils apparaissaient. Cela permet de corroborer par l'exemple une relative fiabilité du filtrage mis en place. Des liens peu interprétables subsistent bien sûr ; on peut par exemple relever la paire `sol_N`/`tête_N` dans la figure 5.6, et les paires `reste_N`/`terrain_N` et `terre_N`/`tour_N` dans la figure 5.8. L'élagage opéré est d'importance différente dans les deux exemples : pour l'exemple (1), 37.0% des liens projetés sont rejetés ; pour l'exemple (2), la proportion tombe à 18,7%. Il s'agit dans un cas comme dans l'autre d'une proportion non négligeable.

Ex. (1) (Girafe)	Exemple (2) (Albanie)
animal_N/effet_N (2 liens)	hiver_N/relief_N
animal_N/réseau_N (4 liens)	île_N/terrain_N
chasse_N/circulation_N (2 liens)	lac_N/terrain_N
effet_N/réseau_N (2 liens)	mont_N/tour_N (2 liens)
	montagne_N/tour_N (2 liens)
	plateau_N/tour_N (2 liens)
	terrain_N/tour_N (2 liens)

TABLEAU 5.11 – Résumé des liens supprimés par le filtrage

Bien sûr, des liens relevant de la cohésion lexicale restent ignorés. On peut par exemple regretter que des mots comme « sang » et « cœur » ne soient pas du tout connectés : on aurait pu espérer repérer des liens tels que `sang`/`sanguin`, `sang`/`veine`, `coeur`/`muscle`, etc. Ces lacunes ne sont pas dues au filtrage des voisins, puisque ces liens n'apparaissent pas non plus lorsque la totalité de la ressource est projetée – ces rapprochements n'ont simplement pas été effectués par l'analyse distributionnelle du corpus, et étaient donc « hors de portée ».

Finalement, au vu de ces exemples, la phase de filtrage semble bien apporter une amélioration à la détection de la cohésion lexicale réalisée par la projection des *Voisins de Wikipédia*. Nous montrons ci-dessous si cette amélioration est reflétée par les résultats du système de segmentation thématique.

5.3.2.2 Filtrage des voisins et segmentation thématique : résultats

Le tableau 5.12 présente les résultats de notre système de segmentation thématique, selon la stratégie de filtrage des voisins. Après les scores obtenus par les deux *baselines* (*cf.* 5.2.3), nous reportons :

- les scores déjà présentés dans le tableau 5.8, obtenus en plaçant un seuil sur le score de Lin pour filtrer les voisins ;
- les scores obtenus en projetant tous les voisins, sans les filtrer ni pondérer les liens ;

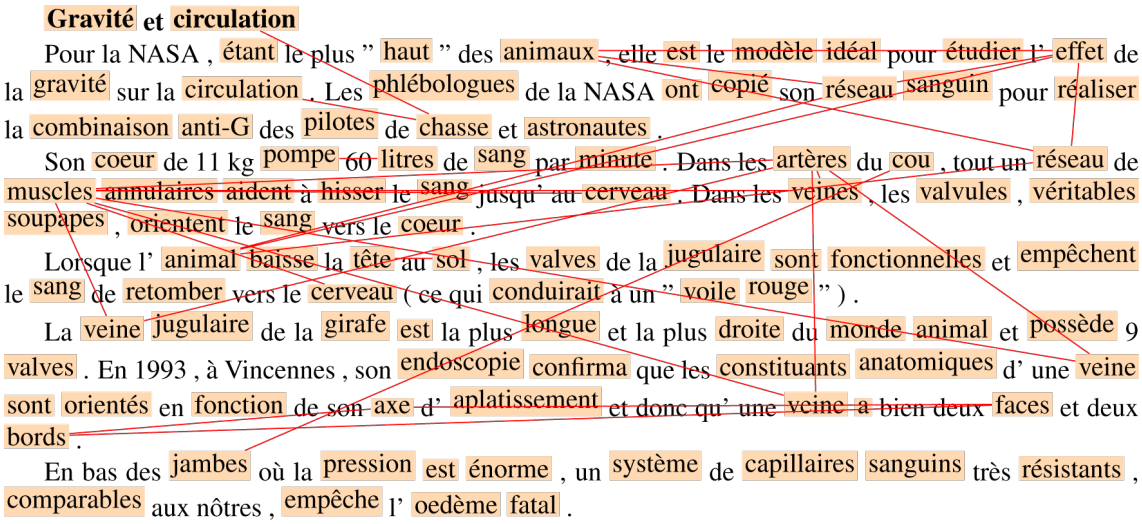


FIGURE 5.5 – Reprise de l'exemple (1) avec projection de tous les liens de voisinage

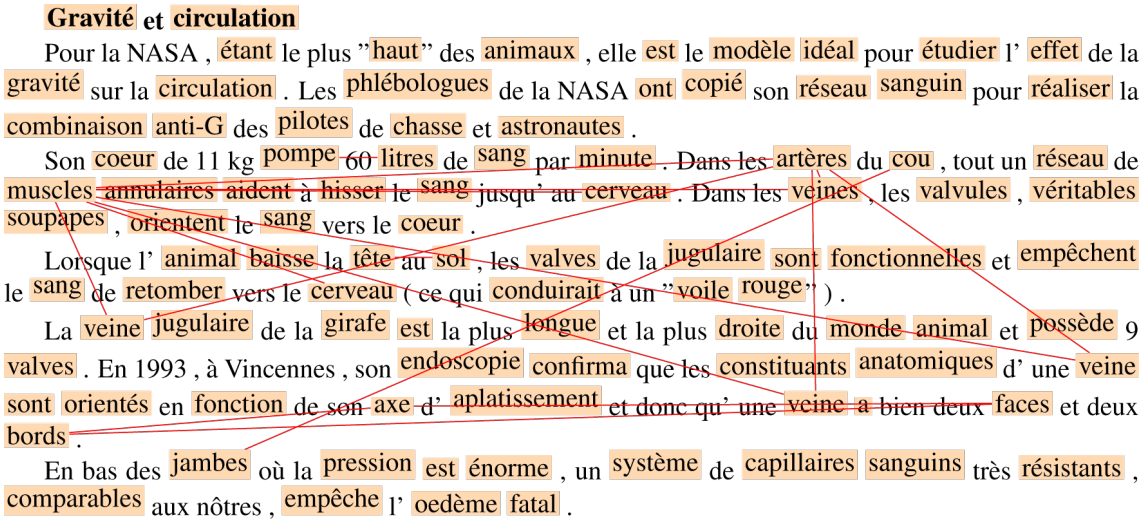
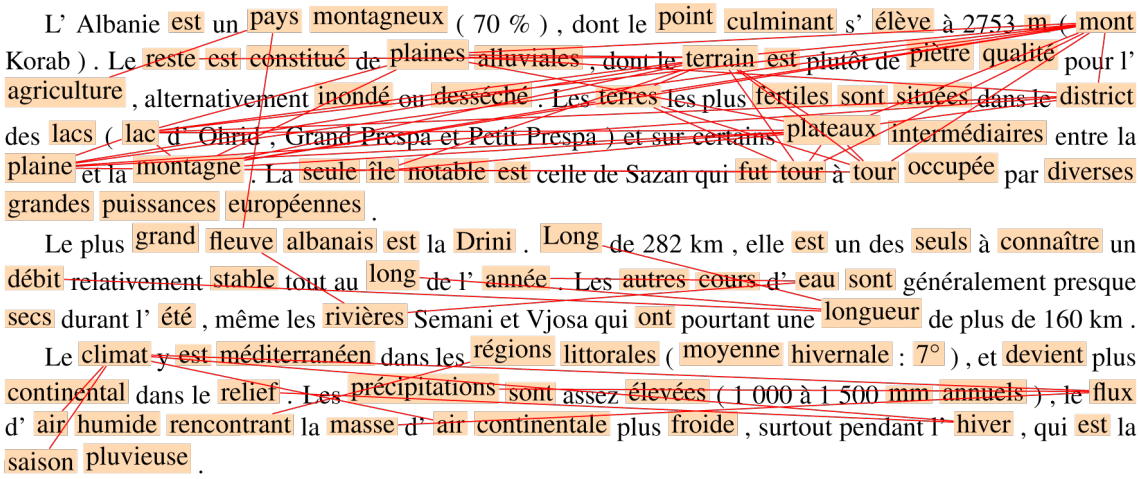


FIGURE 5.6 – Reprise de l'exemple (1) avec filtrage des liens de voisinage



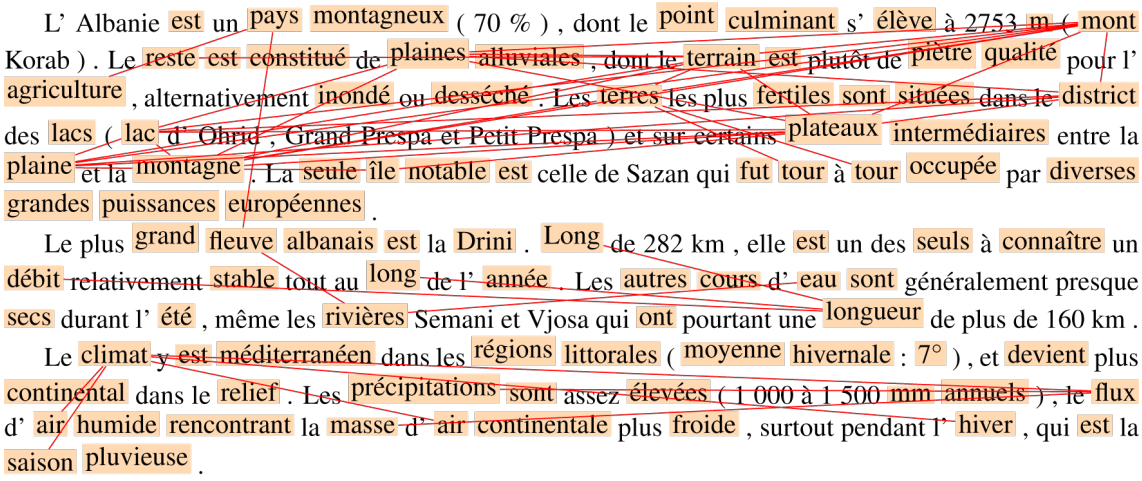
L' Albanie est un pays montagneux (70 %) , dont le point culminant s' élève à 2753 m (mont Korab) . Le reste est constitué de plaines alluviales , dont le terrain est plutôt de piètre qualité pour l' agriculture , alternativement inondé ou desséché . Les terres les plus fertiles sont situées dans le district des lacs (lac d' Ohrid , Grand Prespa et Petit Prespa) et sur certains plateaux intermédiaires entre la plaine et la montagne . La seule île notable est celle de Sazan qui fut tour à tour occupée par diverses grandes puissances européennes .

Le plus grand fleuve albanais est la Drini . Long de 282 km , elle est un des seuls à connaître un débit relativement stable tout au long de l' année . Les autres cours d' eau sont généralement presque secs durant l' été , même les rivières Semani et Vjosa qui ont pourtant une longueur de plus de 160 km .

Le climat y est méditerranéen dans les régions littorales (moyenne hivernale : 7°) , et devient plus continental dans le relief . Les précipitations sont assez élevées (1 000 à 1 500 mm annuels) , le flux d' air humide rencontrant la masse d' air continentale plus froide , surtout pendant l' hiver , qui est la saison pluvieuse .

Résumé des liens : agriculture_N/pays_N, air_N/climat_N, année_N/eau_N, climat_N/flux_N, climat_N/hiver_N, climat_N/saison_N, débit_N/longueur_N, district_N/lac_N, district_N/mont_N, district_N/plaine_N, district_N/plateau_N, eau_N/rivière_N, fleuve_N/pays_N, fleuve_N/rivière_N, flux_N/masse_N, flux_N/précipitation_N, hiver_N/relief_N, île_N/montagne_N, île_N/plateau_N, île_N/terrain_N, lac_N/mont_N, lac_N/montagne_N, lac_N/plaine_N, lac_N/plateau_N, lac_N/terrain_N, long_N/longueur_N, mont_N/montagne_N, mont_N/plaine_N, mont_N/plateau_N, mont_N/terrain_N, mont_N/terre_N, mont_N/tour_N, montagne_N/plaine_N, montagne_N/plateau_N, montagne_N/terrain_N, montagne_N/terre_N, montagne_N/tour_N, plaine_N/plateau_N, plaine_N/terrain_N, plateau_N/terrain_N, plateau_N/tour_N, région_N/rencontrer_V, reste_N/terrain_N, terrain_N/terre_N, terrain_N/tour_N, terre_N/tour_N

FIGURE 5.7 – Reprise de l'exemple (2) avec projection de tous les liens de voisinage



L' Albanie est un pays montagneux (70 %) , dont le point culminant s' élève à 2753 m (mont Korab) . Le reste est constitué de plaines alluviales , dont le terrain est plutôt de piètre qualité pour l' agriculture , alternativement inondé ou desséché . Les terres les plus fertiles sont situées dans le district des lacs (lac d' Ohrid , Grand Prespa et Petit Prespa) et sur certains plateaux intermédiaires entre la plaine et la montagne . La seule île notable est celle de Sazan qui fut tour à tour occupée par diverses grandes puissances européennes .

Le plus grand fleuve albanais est la Drini . Long de 282 km , elle est un des seuls à connaître un débit relativement stable tout au long de l' année . Les autres cours d' eau sont généralement presque secs durant l' été , même les rivières Semani et Vjosa qui ont pourtant une longueur de plus de 160 km .

Le climat y est méditerranéen dans les régions littorales (moyenne hivernale : 7°) , et devient plus continental dans le relief . Les précipitations sont assez élevées (1 000 à 1 500 mm annuels) , le flux d' air humide rencontrant la masse d' air continentale plus froide , surtout pendant l' hiver , qui est la saison pluvieuse .

Résumé des liens : agriculture_N/pays_N, air_N/climat_N, climat_N/flux_N, climat_N/hiver_N, climat_N/saison_N, débit_N/longueur_N, district_N/lac_N, district_N/mont_N, district_N/plaine_N, district_N/plateau_N, eau_N/rivière_N, fleuve_N/pays_N, fleuve_N/rivière_N, flux_N/masse_N, flux_N/précipitation_N, île_N/montagne_N, île_N/plateau_N, lac_N/mont_N, lac_N/montagne_N, lac_N/plaine_N, lac_N/plateau_N, long_N/longueur_N, mont_N/montagne_N, mont_N/plaine_N, mont_N/plateau_N, mont_N/terrain_N, mont_N/terre_N, montagne_N/plaine_N, montagne_N/plateau_N, montagne_N/terrain_N, montagne_N/terre_N, plaine_N/plateau_N, plaine_N/terrain_N, plateau_N/terrain_N, reste_N/terrain_N, terrain_N/terre_N, terre_N/tour_N

FIGURE 5.8 – Reprise de l'exemple (2) avec filtrage des liens de voisinage

– les scores obtenus en filtrant les voisins selon la stratégie présentée en 4.5.2.1 (page 135) : seuls les couples dont la probabilité qu'ils soient pertinents (selon l'algorithme d'apprentissage automatique) est supérieure à 0.3 sont projetés. Ils sont pondérés par cette probabilité (indice qui est donc compris entre 0.3 et 1).

Méthode	P_k*100	WD*100	nb seg/txt
référence	0	0	7.89
baseline "hasard bruité"	36.59	37.38	9.46
baseline "hasard exact"	34.17	34.52	7.89
VD filtrés par *lin*	30.91	31.29	5.09
VD non filtrés	30.86	31.04	4.34
VD filtrés	**30.15**	**30.69**	4.13

TABLEAU 5.12 – Évaluation des différentes stratégies de filtrage des voisins distributionnels (VD)

Comme attendu, les résultats du système apparaissent légèrement meilleurs lorsque les voisins subissent une phase de filtrage. Le filtrage par score de Lin n'apporte quant à lui rien par rapport à une exploitation des voisins sans aucun filtrage. Toutefois, il faut constater le fort tassement de ces résultats, qui ne permet pas de tirer une conclusion claire. Nous discutons de ce phénomène dans les sections 5.3.3.2 et 5.3.4.

5.3.3 Comparaison entre les voisins distributionnels et d'autres types d'informations lexicales

Dans cette section, nous comparons les liens mis au jour par l'utilisation de différentes sources d'informations :

– un simple repérage de répétitions de lemmes, ne faisant donc pas appel à des connaissances extérieures ;
– la projection d'un dictionnaire de synonymes, le Dicosyn [2] ;
– la projection des liens de voisinage telle que précédemment décrite.

5.3.3.1 Informations lexicales prises en compte et détection de la cohésion lexicale : exemple

L'exemple (3) est extrait de la section intitulée « Description », dans l'article « Alouette des champs ». À partir de ce même exemple, nous identifions les liens de répétition (figure 5.9), de synonymie (figure 5.10), puis de voisinage (figure 5.11).

2. 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 à l'adresse suivante : http ://www.crisco.unicaen.fr/des/

(3) Le plumage de l'alouette des champs est peu voyant, brun strié de brun-noirâtre dans la partie supérieure avec une calotte un peu plus foncée et une gorge jaune, finement striée de brun foncé. La crête sur le sommet de la calotte se hérisse à certains moments. Les yeux brun foncé sont rehaussés d'un sourcil blanc-jaune, le bec est plutôt court et couleur corne. La partie inférieure du corps est crème sauf la poitrine chamois clair striée de brun-noir, la queue allongée et presque noire a les rectrices externes tachetées de blanc. Les ailes ont le liséré plus clair, pattes et orteils sont marron clair, le doigt arrière est plus long que les autres.
L'alouette court à ras le sol et s'y aplatit en cas de danger, le "trrlit" qui peut durer des minutes et le vol montant en spirale suivi d'une descente en piqué sont caractéristiques. L'alouette des champs chante – on dit aussi grisolle, tirelire ou turlutte – également au sol de façon très mélodieuse, parfois pendant plus d'une heure, et comme celui du rossignol, ce chant a fasciné les humains. (Wikipédia, article « Alouette des champs »)

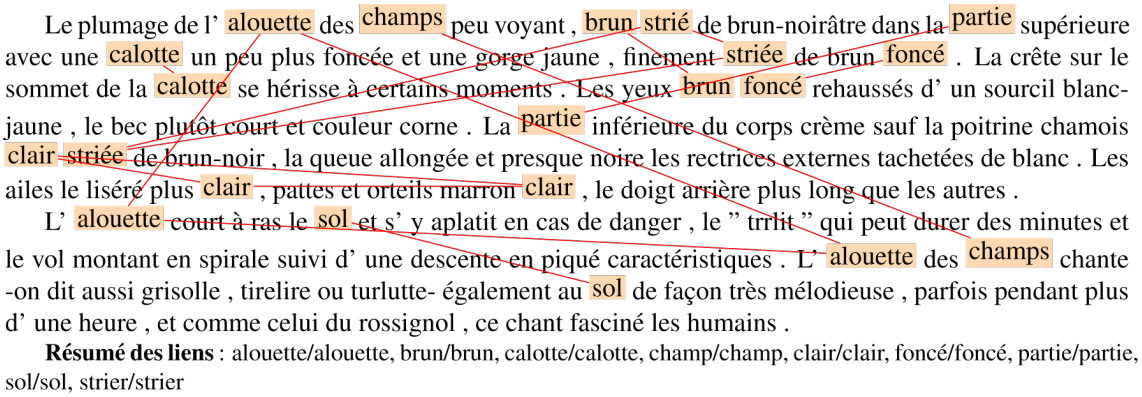
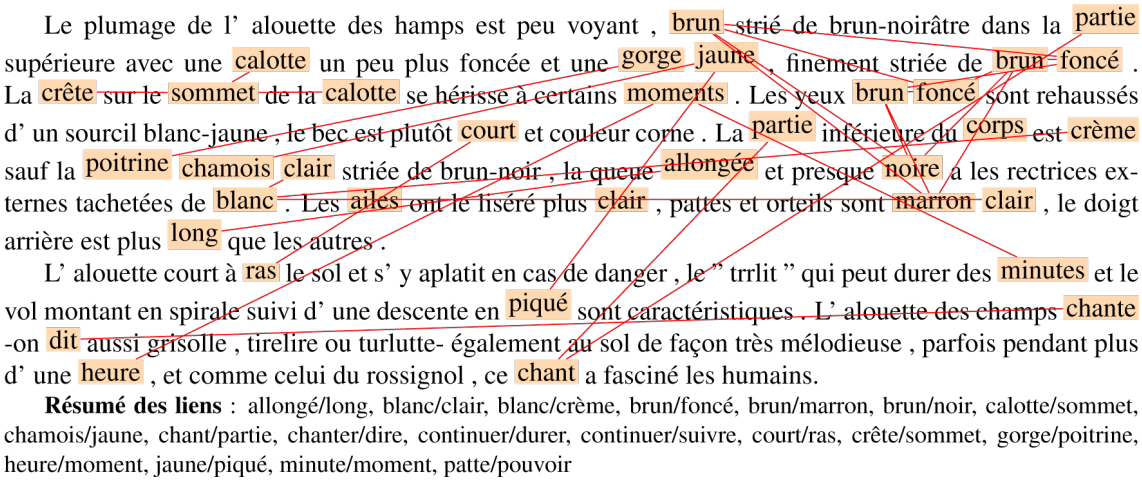


FIGURE 5.9 – Reprise de l'exemple (3) avec identification des liens de répétition

La figure 5.9 nous permet de voir que 9 mots sont répétés dans l'exemple considéré, générant 15 liens lexicaux. Les reprises des mots « alouette » et « champ », qui composent le nom de l'animal dont traite cet article (« l'alouette des champs ») représentent à elles seules 4 liens. On peut supposer que ces mots sont répétés tout au long de l'article. Ils sont toutefois les seuls à connecter les deux paragraphes.

La projection du dictionnaire de synonymes permet quant à elle de faire émerger 27 liens, correspondant à 19 paires différentes. On peut remarquer que l'utilisation d'une ressource classique ne permet pas d'éviter complètement le bruit. Notamment, la synonymie entre les paires `chant`/`partie` et `jaune`/`piqué` repose sur des sens de ces mots qui ne sont pas ceux réalisés dans ce contexte. Les paires `court`/`ras` et `chanter`/`dire` ne nous semblent pas non plus faire émerger de liens de cohésion lexicale pertinents. Les deux paragraphes de l'extrait ne sont que peu connectés, voire pas connectés si l'on exclut les relations non-pertinentes sus-citées.

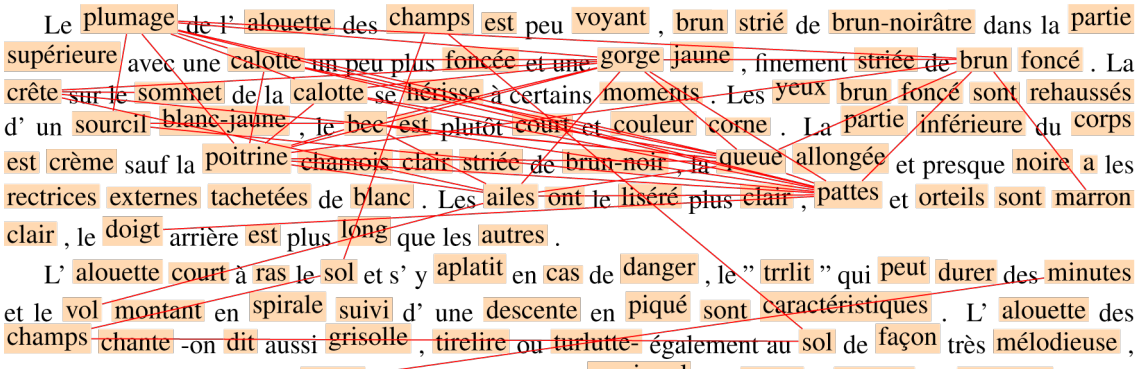


Le plumage de l' alouette des hamps est peu voyant , brun strié de brun-noirâtre dans la partie supérieure avec une calotte un peu plus foncée et une gorge jaune , finement striée de brun foncé . La crête sur le sommet de la calotte se hérisse à certains moments . Les yeux brun foncé sont rehaussés d' un sourcil blanc-jaune , le bec est plutôt court et couleur corne . La partie inférieure du corps est crème sauf la poitrine chamois clair striée de brun-noir , la queue allongée et presque noire a les rectrices externes tachetées de blanc . Les ailes ont le liséré plus clair , pattes et orteils sont marron clair , le doigt arrière est plus long que les autres .

L' alouette court à ras le sol et s' y aplatit en cas de danger , le " trrlit " qui peut durer des minutes et le vol montant en spirale suivi d' une descente en piqué sont caractéristiques . L' alouette des champs chante -on dit aussi grisolle , tirelire ou turlutte- également au sol de façon très mélodieuse , parfois pendant plus d' une heure , et comme celui du rossignol , ce chant a fasciné les humains.

Résumé des liens : allongé/long, blanc/clair, blanc/crème, brun/foncé, brun/marron, brun/noir, calotte/sommet, chamois/jaune, chant/partie, chanter/dire, continuer/durer, continuer/suivre, court/ras, crête/sommet, gorge/poitrine, heure/moment, jaune/piqué, minute/moment, patte/pouvoir

FIGURE 5.10 – Reprise de l'exemple (3) avec identification des liens de synonymie



Le plumage de l' alouette des champs est peu voyant , brun strié de brun-noirâtre dans la partie supérieure avec une calotte un peu plus foncée et une gorge jaune , finement striée de brun foncé . La crête sur le sommet de la calotte se hérisse à certains moments . Les yeux brun foncé sont rehaussés d' un sourcil blanc-jaune , le bec est plutôt court et couleur corne . La partie inférieure du corps est crème sauf la poitrine chamois clair striée de brun-noir , la queue allongée et presque noire a les rectrices externes tachetées de blanc . Les ailes ont le liséré plus clair , pattes et orteils sont marron clair , le doigt arrière est plus long que les autres .

L' alouette court à ras le sol et s' y aplatit en cas de danger , le " trrlit " qui peut durer des minutes et le vol montant en spirale suivi d' une descente en piqué sont caractéristiques . L' alouette des champs chante -on dit aussi grisolle , tirelire ou turlutte- également au sol de façon très mélodieuse , parfois pendant plus d' une heure , et comme celui du rossignol , ce chant a fasciné les humains .

Résumé des liens : aile_N/bec_N, aile_N/crête_N, aile_N/gorge_N, aile_N/patte_N, aile_N/poitrine_N, aile_N/queue_N, aile_N/vol_N, bec_N/gorge_N, bec_N/patte_N, bec_N/poitrine_N, bec_N/queue_N, bec_N/sourcil_N, brun_N/gorge_N, brun_N/marron_N, brun_N/patte_N, brun_N/plumage_N, brun_N/poitrine_N, brun_N/queue_N, calotte_N/gorge_N, calotte_N/patte_N, calotte_N/plumage_N, calotte_N/poitrine_N, calotte_N/queue_N, champ_N/sol_N, crête_N/gorge_N, crête_N/queue_N, crête_N/sommet_N, doigt_N/patte_N, gorge_N/patte_N, gorge_N/plumage_N, gorge_N/poitrine_N, gorge_N/queue_N, heure_N/minute_N, patte_N/plumage_N, patte_N/poitrine_N, patte_N/queue_N, patte_N/sourcil_N, plumage_N/poitrine_N, plumage_N/queue_N, plumage_N/sourcil_N, poitrine_N/queue_N

FIGURE 5.11 – Reprise de l'exemple (3) avec identification des liens de voisinage

La projection des voisins distributionnels fait émerger un nombre de liens bien plus important que les méthodes précédentes (49 liens correspondant à 40 couples de voisins). Ces liens relèvent de relations lexicales variées ; on y retrouve des relations précédemment étiquetées comme synonymiques par la projection de Dicosyn (`brun`/`marron`, `gorge`/`poitrine`). Différents mots désignant des parties du corps de l'alouette apparaissent tous interconnectés, ce qui participe grandement à la cohésion lexicale détectée – `aile`, `crête`, `bec`, `queue`, etc. – alors que `gorge`/`poitrine` était la seule relation de ce type mise au jour grâce aux synonymes. Malgré la richesse des liens lexicaux mis au jour, les deux paragraphes apparaissent encore une fois faiblement connectés, ce qui confirme le peu de cohésion lexicale qu'ils entretiennent, bien qu'ils fassent partie de la même section. Le couple `aile`/`vol` permet toutefois de les relier (la relation entre les deux unités polylexicales « alouette des champs » et « ras le sol » *via* le couple `champ`/`sol` paraît quant à elle peu pertinente).

5.3.3.2 Informations lexicales prises en compte et segmentation thématique : résultats

Le tableau 5.13 présente les résultats de notre système de segmentation thématique, selon la source de connaissances lexicales utilisée. Après la reproduction des scores obtenus par les deux *baselines* (*cf.* 5.2.3), nous reportons :

- les scores obtenus en ne prenant en compte que des répétitions de lemmes, déjà présentés dans le tableau 5.8 ;
- les scores obtenus en prenant en compte la synonymie ;
- les scores obtenus en utilisant les voisins filtrés, déjà présentés dans le tableau 5.12.

Méthode	P_k∗100	WD∗100	nb seg/txt
référence	0	0	7.89
baseline "hasard bruité"	36.59	37.38	9.46
baseline "hasard exact"	34.17	34.52	7.89
répétitions	31.14	31.44	4.93
synonymes	30.85	30.99	4.29
VD filtrés	**30.15**	**30.69**	4.13

TABLEAU 5.13 – Évaluation de l'apport des voisins distributionnels par rapport à d'autres types d'informations lexicales

On peut remarquer que ce sont les voisins distributionnels qui permettent d'obtenir les meilleurs résultats sur notre corpus d'évaluation, suivis des synonymes puis des répétitions. Mais ici encore, on ne peut que constater une grande proximité entre les résultats.

Il serait intéressant de comparer les voisins avec d'autres types d'informations lexicales, notamment avec une autre ressource construite en corpus comme par

exemple une base de collocations. Mais les faibles différences déjà observées n'incitent pas à ce stade à poursuivre des expériences qui se révèlent incapables de faire émerger de résultats marqués, mais à revenir tout d'abord sur les résultats déjà produits.

À partir des tableaux 5.12 et 5.13 on peut par ailleurs remarquer un phénomène étonnant : plus le nombre de ruptures postulées par une méthode est faible, et s'éloigne donc de celui de la référence, meilleurs sont les résultats. Pourtant, comme on l'a vu, le score *WindowDiff* est conçu de manière à ne pas favoriser les faux négatifs sur les faux positifs. On peut donc supposer que les différentes méthodes présentées placent de moins en moins de ruptures, mais que celles-ci sont de plus en plus précises. Le faible nombre de ruptures placées explique partiellement le tassement des résultats : il est plus difficile de faire émerger des différences entre ces méthodes en se basant sur un nombre de ruptures par texte très faible. Une solution pourrait être de choisir des textes plus longs pour l'évaluation. Par ailleurs, le nombre de ruptures placées est fortement dépendant du choix de paramètres (taille de la fenêtre, importance du lissage...) ; or, il apparaît que l'optimisation des paramètres (*cf.* section 5.2.3.1) a mis en avant des configurations aboutissant à des nombres faibles de ruptures ; un paramétrage différent permettrait peut-être de mieux observer d'éventuelles différences entre méthodes. Ainsi, lors d'une précédente expérience (Adam et Morlane-Hondère, 2009) utilisant pour le système de segmentation thématique les paramètres par défaut fixés par Hearst (1997), des différences plus importantes avaient été relevées selon les différentes informations lexicales prises en compte ; le classement entre les méthodes était quant à lui identique : *baseline* - répétitions - synonymes - voisins.

Au delà de ces considérations techniques touchant au système de segmentation mis en œuvre, et qui invitent à mener de nouveaux tests, un retour peut également être fait sur les textes constituant le matériel d'évaluation. La section 5.3.4 y est consacrée.

5.3.4 Quelques éléments de mise en perspective des résultats

L'hypothèse régissant l'évaluation des voisins par la tâche de segmentation thématique était que, puisque cette tâche repose principalement sur la cohésion lexicale, une meilleure détection de la cohésion lexicale devrait aboutir à de meilleures performances. Au terme de l'évaluation menée, qui ne fait émerger que de très faibles différences selon les informations lexicales prises en compte (répétitions, voisins, voisins filtrés, synonymes), nous avons souhaité revenir sur cette hypothèse.

Nos résultats se caractérisent surtout par un grand nombre de faux-négatifs, et ce quelle que soit la méthode utilisée. On voit en effet dans les tableaux 5.12 et 5.13 que le nombre de segments postulés par les différentes méthodes est toujours nettement inférieur au nombre de segments de la référence. Nous avons repéré dans

les textes des exemples de faux-négatifs, afin de réfléchir aux causes de ces faux négatifs, du point de vue de la cohésion lexicale détectable par nos méthodes.

Dans certains cas, c'est notre choix de considérer tous les titres de section comme des ruptures thématiques qui est mis à mal. Par exemple, dans l'article « Danemark », la section intitulée « Culture » est suivie par le titre (de même niveau) « Subcultures ». Dans un autre article, « Tigre du Bengale », la section « Alimentation » est suivie par une section intitulée « Le saviez-vous ? », contenant essentiellement des informations sur l'alimentation du tigre : « Poussé par la faim, un tigre peut tuer l'équivalent de 30 buffles par an. », « Un tigre adulte peut manger 31kg de viande en une nuit. », « Contrairement aux autres félins, les tigres mangent souvent de la viande en début de décomposition. ». Dans de tels cas, il ne paraît pas justifié d'attendre d'un système de segmentation thématique qu'il identifie une rupture thématique. Toutefois, dans l'ensemble des données observées, ces cas semblent très minoritaires.

Dans les cas où le changement de thème paraît effectif, la non détection d'une rupture peut être rapportée à une trop faible variation de cohésion lexicale entre les deux segments. Nous donnons deux exemples pour lesquels un changement de thème ne s'accompagne pas d'un creux identifiable dans la cohésion lexicale détectée par projection des voisins distributionnels.

Dans l'exemple 5.12, extrait de l'article « Bouvreuil pivoine », la section « Alimentation » apparaît fortement liée à la section précédente, dédiée à l'habitat du bouvreuil pivoine (6 liens connectent les deux sections, c'est autant que le nombre de liens internes à la section « Alimentation »). Il est difficile d'attribuer cette forte connectivité à du bruit, dans la mesure où les liens projetés paraissent globalement pertinents, à l'exception peut-être du lien `arbre/source`[3]. On peut plutôt remarquer que les deux thèmes apparaissent en fait très liés, puisque l'habitat du bouvreuil est déterminé par son type d'alimentation. De manière générale, les changements de thème s'accompagnant par un changement radical de vocabulaire sont rares.

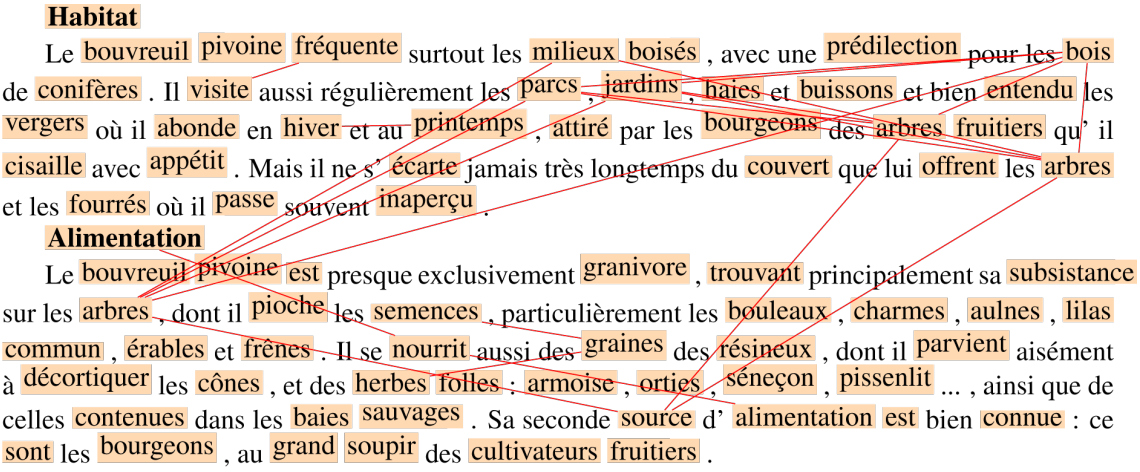
La figure 5.13 présente un exemple un peu plus long, extrait du texte « Albanie ». Les sections consécutives « Géographie » et « Économie » comptent respectivement 4 et 2 paragraphes. Pour plus de lisibilité, nous ne montrons que les liens inter-paragraphiques. L'observation des liens projetés montre que les différents paragraphes de la section « Géographie » sont bien plus connectés à la section « Économie » qu'aux autres paragraphes de la même section, ce qui explique la non détection de cette frontière.

Cette forte connectivité peut être attribuée à différentes causes.

Tout d'abord, on peut observer un certain chevauchement des thèmes :

– dans les deux sections, il est question d'agriculture ;
– l'introduction des ressources naturelles, à la fin de la section consacrée à la

3. Ces mots ont été rapprochés sur la base de contextes tels que `exploitation_de`, `protéger_obj`, `parc_de`.



Résumé des liens : alimentation_N/nourrir_V, arbre_N/bois_N, arbre_N/jardin_N, arbre_N/milieu_N, arbre_N/parc_N, arbre_N/source_N, bois_N/jardin_N, bois_N/parc_N, fréquenter_V/visiter_V, graine_N/herbe_N, graine_N/semence_N, hiver_N/printemps_N

FIGURE 5.12 – Sections « Habitats » et « Alimentation » du texte « Bouvreuil pivoine »

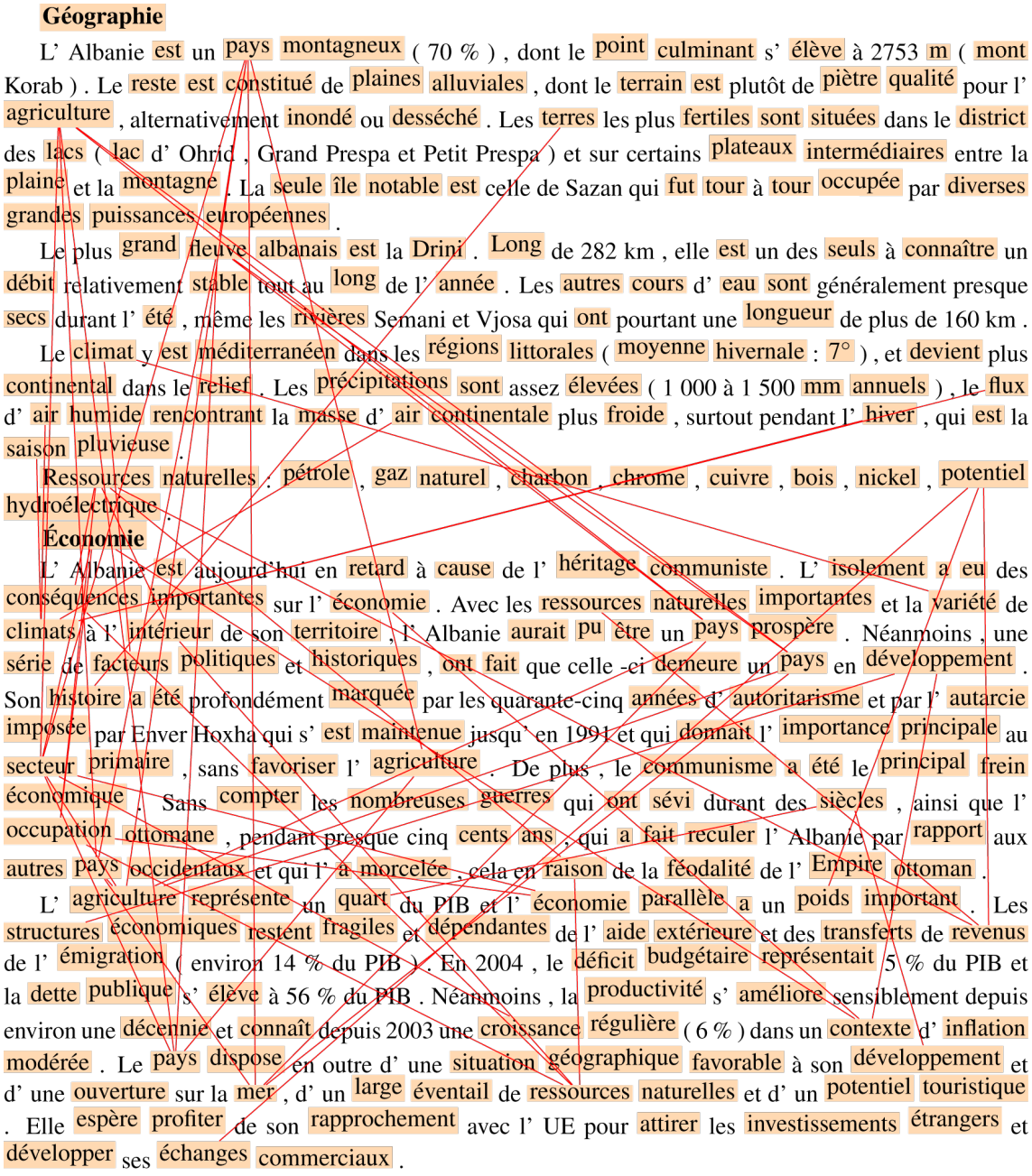
géographie de l'Albanie, amorce la transition vers le thème de l'économie ; on peut d'ailleurs constater que le mot « ressource » est uniquement connecté vers la suite du texte, et pas du tout à la section où il apparaît pour la première fois ;

– à l'inverse, la section « Économie » débute avec un rappel de ce qui a été dit dans la section précédente, notamment *via* la phrase « Avec les ressources naturelles importantes et la variété de climats à l'intérieur de son territoire, l'Albanie aurait pu être un pays prospère. »

Par ailleurs, un brouillement des pistes est introduit par des mots ne relevant pas d'un sous-thème particulier mais étant fortement connectés. Par exemple, la forte productivité du mot `pays`, qui, dans ce seul extrait, a pour voisins `agriculture`, `fleuve`, `mer` et `secteur`, peut également être mise en cause. En effet, ce mot entretient des liens avec des représentants des différents thèmes, et il est fréquemment répété (5 fois dans l'extrait présenté), ce qui contribue largement à la cohésion lexicale observée entre les deux sections. On peut supposer que ce phénomène concerne l'ensemble du texte, puisque celui-ci est consacré à un pays.

Si, dans cet exemple, du bruit est encore une fois bien sûr observable (par exemple, il est difficile d'interpréter dans ce contexte la paire `fleuve_N`/`secteur_N`[4]), son influence sur le phénomène considéré (la plus forte cohésion lexicale détectée entre sections qu'à l'intérieur d'une section) semble mineure.

4. rapprochée uniquement par des adjectifs tels que `côtier`, `large`, `maritime`, `français`, etc.



Géographie

L' Albanie est un pays montagneux (70 %) , dont le point culminant s' élève à 2753 m (mont Korab) . Le reste est constitué de plaines alluviales , dont le terrain est plutôt de piètre qualité pour l' agriculture , alternativement inondé ou desséché . Les terres les plus fertiles sont situées dans le district des lacs (lac d' Ohrid , Grand Prespa et Petit Prespa) et sur certains plateaux intermédiaires entre la plaine et la montagne . La seule île notable est celle de Sazan qui fut tour à tour occupée par diverses grandes puissances européennes .

Le plus grand fleuve albanais est la Drini . Long de 282 km , elle est un des seuls à connaître un débit relativement stable tout au long de l' année . Les autres cours d' eau sont généralement presque secs durant l' été , même les rivières Semani et Vjosa qui ont pourtant une longueur de plus de 160 km .

Le climat y est méditerranéen dans les régions littorales (moyenne hivernale : 7°) , et devient plus continental dans le relief . Les précipitations sont assez élevées (1 000 à 1 500 mm annuels) , le flux d' air humide rencontrant la masse d' air continentale plus froide , surtout pendant l' hiver , qui est la saison pluvieuse .

Ressources naturelles : pétrole , gaz naturel , charbon , chrome , cuivre , bois , nickel , potentiel hydroélectrique .

Économie

L' Albanie est aujourd'hui en retard à cause de l' héritage communiste . L' isolement a eu des conséquences importantes sur l' économie . Avec les ressources naturelles importantes et la variété de climats à l' intérieur de son territoire , l' Albanie aurait pu être un pays prospère . Néanmoins , une série de facteurs politiques et historiques , ont fait que celle -ci demeure un pays en développement . Son histoire a été profondément marquée par les quarante-cinq années d' autoritarisme et par l' autarcie imposée par Enver Hoxha qui s' est maintenue jusqu' en 1991 et qui donnait l' importance principale au secteur primaire , sans favoriser l' agriculture . De plus , le communisme a été le principal frein économique . Sans compter les nombreuses guerres qui ont sévi durant des siècles , ainsi que l' occupation ottomane , pendant presque cinq cents ans , qui a fait reculer l' Albanie par rapport aux autres pays occidentaux et qui l' a morcelée , cela en raison de la féodalité de l' Empire ottoman .

L' agriculture représente un quart du PIB et l' économie parallèle a un poids important . Les structures économiques restent fragiles et dépendantes de l' aide extérieure et des transferts de revenus de l' émigration (environ 14 % du PIB) . En 2004 , le déficit budgétaire représentait 5 % du PIB et la dette publique s' élève à 56 % du PIB . Néanmoins , la productivité s' améliore sensiblement depuis environ une décennie et connaît depuis 2003 une croissance régulière (6 %) dans un contexte d' inflation modérée . Le pays dispose en outre d' une situation géographique favorable à son développement et d' une ouverture sur la mer , d' un large éventail de ressources naturelles et d' un potentiel touristique . Elle espère profiter de son rapprochement avec l' UE pour attirer les investissements étrangers et développer ses échanges commerciaux .

Résumé des liens : agriculture_N/pays_N, agriculture_N/secteur_N, air_N/climat_N, climat_N/facteur_N, climat_N/flux_N, climat_N/hiver_N, climat_N/saison_N, climat_N/variété_N, conséquence_N/ressource_N, contexte_N/développement_N, contexte_N/économie_N, croissance_N/secteur_N, développement_N/siècle_N, développement_N/structure_N, échange_N/potentiel_N, économie_N/occupation_N, économie_N/secteur_N, fleuve_N/pays_N, histoire_N/terre_N, intérieur_N/ressource_N, mer_N/pays_N, occupation_N/ressource_N, pays_N/secteur_N, poids_N/potentiel_N, potentiel_N/revenu_N, quart_N/siècle_N, raison_N/ressource_N, ressource_N/revenu_N

FIGURE 5.13 – Sections « Géographie » et « Économie » du texte « Albanie »

Ces deux exemples aident à comprendre le grand nombre de faux-négatifs observés. On sait que la cohésion lexicale couvre l'ensemble des textes ; la segmentation thématique repose sur l'hypothèse que des variations de cohésion lexicale accompagnent les changements de thèmes d'un texte, variations suffisantes pour être repérées automatiquement. Dans les deux exemples présentés, cette hypothèse n'est pas vérifiée. Le caractère bruité de la ressource peut être partiellement mis en cause, il ne semble pas seul responsable.

Les données annotées décrites et utilisées dans le chapitre 4 peuvent fournir des éléments plus quantitatifs pour confirmer ces observations. Rappelons qu'un ensemble de 10000 couples environ, sélectionnés aléatoirement dans un corpus de 42 textes, ont été annotés comme participant ou non à la cohésion lexicale de ces textes. Ces textes étant également des articles de l'encyclopédie Wikipédia, ils sont similaires à ceux utilisés dans l'expérience de segmentation thématique décrite dans ce chapitre.

Dans ces 42 textes, à chaque intersection entre paragraphes (le paragraphe étant notre unité minimale de segmentation), nous avons compté le nombre de liens de voisinage dans une fenêtre allant de 50 mots pleins avant l'intersection à 50 mots pleins après (il s'agit de la même fenêtre que celle appliqué par notre système de segmentation thématique). Le tableau 5.14 résume les nombres moyens de liens selon que les deux paragraphes à l'intersection desquels on se trouve appartiennent à la même section ou non. Nous reportons les moyennes de liens projetés, de liens annotés, et parmi ceux-ci, de liens annotés comme étant pertinents.

	section identique	sections différentes	rapport
liens projetés	689.13 (242.88)	601.89 (248.98)	53.3% / 46.7%
liens annotés	43.72 (38.40)	35.58 (32.28)	55.1% / 44.9%
liens pertinents	10.33 (16.22)	6.19 (7.50)	62.5% / 37.5%

TABLEAU 5.14 – Moyennes (et écarts-types) de liens dans une fenêtre de 100 mots pleins

Si l'on se réfère aux moyennes de liens projetés, on constate que les paragraphes appartenant à la même section sont un peu plus connectés que les paragraphes appartenant à deux sections différentes, mais que cette différence est très faible. Les moyennes de liens annotés permettent de vérifier leur représentativité par rapport aux liens projetés (la proportion de liens entre paragraphes d'une même section est légèrement supérieure à celle observée parmi les liens projetés, donc on voit qu'on a légèrement favorisé cette catégorie). Enfin, si l'on s'intéresse uniquement aux liens annotés comme pertinents, on se rend compte qu'ils sont effectivement un peu plus nombreux lorsque les paragraphes considérés font partie de la même section.

Il ressort donc que :

– lorsque les deux paragraphes considérés appartiennent à la même section, un nombre un peu plus important de liens sont projetés en moyenne ;

– ces liens ont un peu plus tendance à être annotés comme pertinents par les annotateurs (la corrélation de Pearson (*cf.* section 4.4) est de 0.07***).

Toutefois, la subtilité des différences observées ainsi que les forts écarts-types reportés font douter de la possibilité de les exploiter pour la segmentation des textes en sections. La figure 5.14 montre la distribution des nombres de liens pour chaque intersections selon qu'elle soit au milieu d'une section ou entre deux sections ; elle permet d'apprécier la proximité entre les deux catégories.

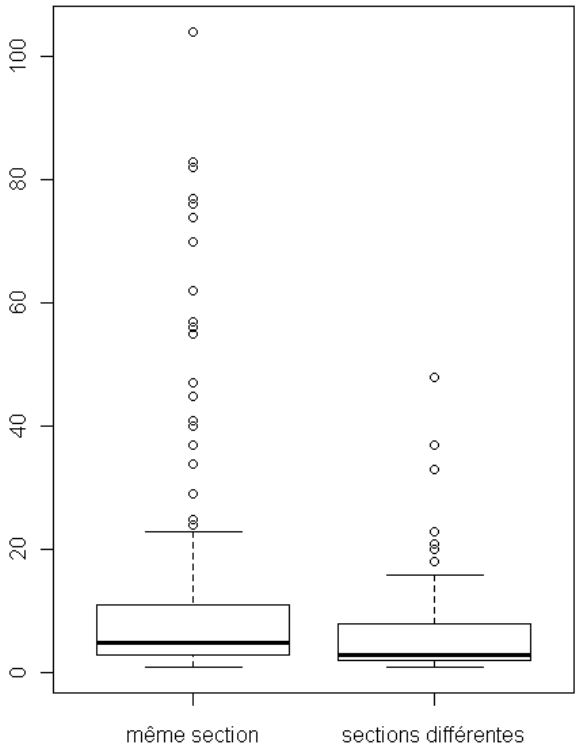


FIGURE 5.14 – Nombres de liens pertinents selon le type d'intersection

Ces données nous montrent que d'une part la cohésion lexicale telle qu'on la mesure est très variable au fil du texte, mais que d'autre part, l'influence imputable aux changements de sections, si elle semble avérée, reste faible.

Ces informations amènent à relativiser fortement le poids de la qualité de la détection de la cohésion lexicale dans l'approche de la segmentation thématique que nous avons mise en œuvre, puisque, même si l'on ne considère que des liens validés par des juges, la distinction entre paragraphes appartenant à une même section (c'est-à-dire ce qu'on a considéré comme un même segment thématique) ou à deux

sections différentes paraît difficile à établir.

Ces observations ne disqualifient pas la tâche de segmentation thématique, mais elles montrent les limites du matériel d'évaluation que nous avons utilisé. On peut supposer que ce problème de faible impact des changements de thèmes sur la quantité de liens de cohésion lexicale se présente à moindre mesure lorsque l'on évalue sur des pseudo-textes, puisque dans ce cas-là on n'attend aucune forme de cohésion lexicale entre segments.

Ces observations plaident également, selon nous, en faveur d'une meilleure intégration de la notion de « thème » aux systèmes de segmentation. Il apparaît en effet que tous les liens de cohésion lexicale ne sont pas porteurs de cohérence thématique. Ainsi, le lien `alimentation_N`/`habitat_N` dans l'exemple 5.12, dans la mesure où il connecte deux mots représentants de deux thèmes différents du texte, ne gagne pas à être pris en compte lorsque le but est de distinguer différents thèmes. De la même manière, les liens de voisinage impliquant un mot comme « pays » dans un texte portant justement sur un pays semblent peu pertinents pour cette tâche, même lorsqu'ils constituent de « bons » liens de cohésion lexicale (on a vu dans la section 4.4.3 p. 124 que les liens impliquant un mot fréquent dans le texte ont davantage tendance à être annotés comme participant à la cohésion lexicale du texte).

Ainsi, l'hypothèse selon laquelle les performances d'un système de segmentation thématique seraient forcément améliorées par une meilleure détection de la cohésion lexicale semble être à mitiger. Détecter une plus large gamme de liens lexicaux semble être un atout (comme on l'a vu dans la section (3)), mais demande également plus de contrôle de ces liens. Par exemple, le bruit induit par des mots apparaissant régulièrement dans l'ensemble du texte (et donc peu pertinents pour la segmentation thématique) se trouve fortement amplifié du fait que l'on considère une plus grande palette de liens.

Finalement, lorsque le but de la segmentation est d'identifier plusieurs thèmes au sein du même texte, les approches basées sur une détection des thèmes (Ferret, 2007; Chen *et al.*, 2009), ou pondérant les liens de cohésion lexicale non pas en fonction d'un indice de leur qualité (comme le score de Lin) mais plutôt de leur représentativité vis-à-vis du thème supposé (*via* des mesures de *tf·idf* local par exemple) (Malioutov et Barzilay, 2006; Dias *et al.*, 2007) nous paraissent plus adaptées que celles basées uniquement sur la notion de cohésion lexicale.

5.4 Bilan et perspectives

Dans ce chapitre, nous avons abordé la tâche de segmentation thématique, qui présentait un intérêt particulier du point de vue du projet VOILADIS, car elle repose principalement sur la notion de cohésion lexicale. Nous avons développé un système de segmentation thématique fonctionnant selon le principe du *TexTiling* et avons présenté plusieurs expériences utilisant ce système.

Une première expérience nous a permis de valider l'intuition selon laquelle la

segmentation thématique n'est pas appropriée pour traiter tous les types de textes. Nous avons extrait de manière automatique un sous-ensemble de textes de Wikipédia que nous avons regroupés sous l'appellation « corpus THEM ». Nous avons ensuite montré que les résultats obtenus sur ce corpus étaient significativement meilleurs que ceux obtenus par l'utilisation de deux *baselines*, à l'inverse des résultats résultant du traitement d'un autre corpus (nommé NON-THEM).

Dans un second temps, nous avons exploité le corpus THEM pour évaluer l'apport des voisins distributionnels à la tâche de segmentation thématique, en deux temps :

- tout d'abord en comparant plusieurs stratégies de filtrage des voisins, afin de valider les propositions effectuées dans le chapitre précédent ;
- puis en comparant l'utilisation des voisins à celle de répétitions ou de synonymes.

Nos résultats pointent vers une bonne efficacité du filtrage mis en place, et vers un apport des voisins distributionnels par rapport aux autres types d'informations lexicales. Toutefois, alors que les analyses qualitatives semblaient montrer des différences assez nettes, ces différences ne se sont que faiblement reportées sur les variations de performances du système de segmentation thématique. Nous avons discuté certaines causes possibles de ce phénomène, mais de manière non exhaustive. Nous fournissons ainsi une nouvelle illustration, détaillée et exemplifiée, de la difficulté (déjà soulignée, notamment par Bestgen et Piérard, 2006) qu'ont les systèmes de segmentation thématique à reproduire l'organisation thématique de textes réels.

Nos perspectives sont diverses.

Concernant la segmentation thématique, la poursuite des expériences menées doit passer selon nous par une phase de retour en arrière, c'est-à-dire d'examen poussé des résultats obtenus et de remise en cause de nos prémisses. Notre objectif était double en abordant la tâche de segmentation thématique, comme nous l'avons souligné dans l'en-tête de ce chapitre ; il s'agissait pour nous d'explorer l'apport des voisins distributionnels :

(a) à la détection de la cohésion lexicale (la segmentation thématique étant alors considérée comme une tâche *via* laquelle on évalue) ;

(b) à la détection de l'organisation thématique des textes postulée par la segmentation thématique (on s'intéresse alors à la segmentation thématique en elle-même).

Ces deux versants sont bien sûr très liés, et nous les avons donc abordés conjointement. Nous pensons toutefois qu'il serait bon de les traiter de manière séparée dans une prochaine phase.

(a) Si l'on considère l'objectif consistant à éprouver la capacité de notre méthodologie de projection des *Voisins de Wikipédia* à détecter la cohésion lexicale, nous pensons que nous avons opté pour un matériel d'évaluation démesurément « difficile », d'une part à cause de la difficulté connue des systèmes à traiter ce type de matériel, d'autre part parce qu'il était finalement assez peu contrôlé (et très

hétérogène, avec des textes très longs ou très courts, une grande variabilité de la taille des sections, etc.). Il serait intéressant de réitérer certaines expériences avec un matériel d'évaluation plus classique, constitué d'extraits d'articles concaténés. Une telle évaluation ne serait selon nous pas portable à la problématique de l'organisation thématique de textes réels, mais elle permettrait sans doute des comparaisons plus nettes entre différentes ressources, différentes stratégies de filtrage des voisins, etc.

(b) La problématique de l'organisation thématique des textes devrait quant à elle être appréhendée avec une plus grande réflexion sur la notion de thème, sur les conditions qui font qu'un lien de cohésion lexicale est porteur de cohérence thématique, et en travaillant à l'élaboration d'un corpus d'évaluation mieux contrôlé. Dans ce cadre, nous souhaiterions notamment étudier le statut des phrases de transitions entre segments, et le brouillage qu'elles font subir à la segmentation thématique des textes.

Mais notre principal horizon, dans le cadre du projet VOILADIS, est surtout d'aller au delà de la segmentation thématique et de la vision limitée qu'elle offre de l'organisation textuelle, en travaillant sur des structures discursives mieux décrites linguistiquement. C'est pourquoi nous nous détachons dans la suite de cette thèse de la problématique de la segmentation thématique, pour étudier d'autres types d'objets, identifiés lors du projet ANNODIS. Dans le chapitre consacré aux structures énumératives (chapitre 6), nous montrons que des méthodes issues de la segmentation thématique peuvent être mobilisées pour aborder certains aspects de ces objets discursifs.

Chapitre 6

Structures énumératives

Sommaire

« Nous pourrons tenter, si cela convient à Votre Seigneurie, de déblayer un sphinx enfoui, de désobstruer un naos, d'ouvrir un hypogée... »
Voyant que le lord restait impassible à cette alléchante énumération, et qu'un sourire sceptique errait sur les lèvres du savant, Argyropoulos comprit qu'il n'avait pas affaire à des dupes faciles.

Théophile Gautier, Le Roman de la momie

Dans ce chapitre, nous proposons d'appliquer le critère de cohésion lexicale à l'analyse des structures énumératives. Nous nous appuyons pour cela sur les annotations effectuées lors du projet ANNODIS « descendant ». Les structures énumératives sont des structures discursives bien identifiées et dont la description linguistique ne fait pas directement appel à la cohésion lexicale. Nous insistons sur les méthodes d'approche que nous avons mises en place pour aborder ces objets discursifs et proposons des pistes de recherche.

6.1 Contexte et motivations

6.1.1 Le projet ANNODIS « descendant » et les structures énumératives

un soleil d'Austerlitz
un siphon d'eau de Seltz
un vin blanc citron
un Petit Poucet un grand pardon un calvaire
de pierre une échelle de corde
(...)
plusieurs ratons laveurs.

Jacques Prévert, Inventaire

6.1.1.1 Le choix des structures énumératives

Le sous-projet ANNODIS « descendant » (Ho-Dac *et al.*, 2009), abordant les textes avec un point de vue global, s'est intéressé au repérage de structures discursives de haut niveau. L'annotation descendante s'est focalisée essentiellement sur une méta-structure : la structure énumérative (SE). Énumérer est un acte textuel, une stratégie de base de mise en texte qui consiste à rassembler plusieurs éléments dans un même objet textuel en fonction d'une identité de statut : on parle de coénumérabilité [1]

1. contrainte que ne respecte pas l'extrait poétique placé en épigraphe

(Luc, 2000). Une SE est composée d'une énumération (c'est-à-dire d'une succession d'items coénumérés), qui peut être précédée d'une amorce et/ou suivie d'une clôture. Le critère d'égalité des différentes unités coénumérées est nommé énuméraThème ; l'énuméraThème est indispensable à l'interprétation de toute SE, mais il n'est pas toujours réalisé explicitement, pouvant être inféré à partir du contenu des items de l'énumération (Bras *et al.*, 2008).

L'exemple (1) montre une SE extraite de l'article « Jules César ». Cette SE est dotée d'une amorce, de deux items et d'une clôture. Son énuméraThème est explicité par les mots en gras dans le texte.

(1) [Des divers **écrits** qu'il avait composés, il ne nous reste que ses **Commentaires** (Commentarii rerum gestarum) :]$_{AMORCE}$
[- De Bello Gallico, « Commentaires sur la Guerre des Gaules », relatant la campagne de César en Gaule.]$_{ITEM1}$
[- De Bello ciuile, « Commentaires sur la Guerre civile », relatant la guerre civile contre Pompée.]$_{ITEM2}$
[Ces **œuvres** constituent le modèle du genre des mémoires historiques, même si leur objectivité est discutée par les historiens.]$_{CLOTURE}$

Ho-Dac *et al.* (2009) citent quatre propriétés des SE qui ont incité à étudier cette méta-structure dans le cadre du projet ANNODIS « descendant » :

> 1. les textes, surtout de type expositif, font fortement appel à une organisation en SE ;
> 2. les SE sont présentes à différents niveaux de grain, du très global (tout un chapitre ou section) au local (quelques propositions) ;
> 3. ce mode de structuration est associé à une grande variété de patrons textuels : du découpage en sections aux patrons d'amorce et séquences de marqueurs d'items en passant par les listes formatées ;
> 4. ce mode de structuration est aisément identifiable par le lecteur (Turco et Coltier, 1988, p. 57).
>
> (Ho-Dac *et al.*, 2010)

Ainsi, la multiplicité de leurs réalisations fait des SE un bon point d'entrée dans la complexité de l'organisation textuelle abordée d'un point de vue descendant.

6.1.1.2 Campagne d'annotation

Le corpus utilisé pour la campagne d'annotation ANNODIS « descendant » est constitué de textes longs de types expositif, de différentes provenances : articles de l'encyclopédie Wikipédia (sous-corpus WIKI), articles traitant de géopolitique publiés par l'Institut Français des Relations Internationales (sous-corpus GEOPO) et articles scientifiques issus des actes du premier Congrès Mondial de Linguistique Française (sous-corpus CMLF). Ce corpus est caractérisé dans le tableau 6.1. Le

sous-corpus WIKI, qui nous intéresse plus particulièrement, est le plus important en nombre de mots bien qu'il ne soit constitué que de 25 textes.

	Nombre de textes	Nombre de mots
WIKI	25	199 858
GEOPO	31	181 414
CMLF	30	133 515
Tot.	86	514 787

TABLEAU 6.1 – Caractérisation du corpus ANNODIS « descendant »

L'annotation manuelle des SE s'est appuyée sur un pré-marquage automatique de traits linguistiques : marqueurs d'intégration linéaire (Turco et Coltier, 1988), cadratifs, patrons typo-dispositionnels, etc. Le but de ce pré-marquage est d'assister l'annotation en facilitant le repérage de SE à partir de zones denses en traits associés à la structuration du discours. Elle a été réalisée grâce à l'interface GLOZZ (*cf.* section 3.3.3.2 page 84) par trois annotateurs étudiants en sciences du langage (niveau master), ayant lu un manuel d'annotation rédigé lors d'une phase exploratoire. L'identification de chaque SE peut s'accompagner de l'annotation de différents éléments : amorce, items, clôture, énuméraThème, indices. Les indices peuvent être des traits pré-marqués qui ont été identifiés comme indices de repérage des SE par l'annotateur, ou des traits supplémentaires non pré-marqués qu'il a lui-même repérés. Neuf textes ont reçu une annotation multiple, ce qui a permis de calculer un taux d'accord, qui est de 0.7 (Ho-Dac *et al.*, 2010).

6.1.1.3 Bilan de l'annotation

Nous donnons ici les résultats de l'annotation sur le sous-corpus WIKI, qui est le seul que nous utilisons pour les expériences et analyses relatées dans ce chapitre. Dans ce sous-corpus, 332 SE ont été annotées ; cela représente en moyenne entre 13 et 14 SE par texte. En moyenne, 60% d'un article est couvert par au moins une SE (les SE étant des structures récursives, une portion de texte peut être couverte par plusieurs SE).

Les SE sont principalement caractérisées par :

- leur taille (en nombre de mots) ;
- leur cardinalité (en nombre d'items) ;
- leur composition (présence ou non des éléments optionnels que sont l'amorce et la clôture) ;
- leur niveau de grain (interaction avec la structure du document.

Une typologie basée sur le niveau de grain a été proposée, car cette caractéristique permet une répartition assez équilibrée des SE et est la plus fortement liée aux autres caractéristiques. Cette typologie distingue 4 types de SE :

- Type 1 : dont les items correspondent à des sections titrées ;
- Type 2 : dont les items correspondent à des listes formatées ;

– Type 3 : couvrant plus d'un paragraphe sans marques visuelles spécifiques ;
– Type 4 : intra-paragraphiques.
(Ho-Dac *et al.*, 2010)

Le sous-corpus WIKI compte 64 SE de type 1 (19,3%), 130 SE de type 2 (39.1%), 69 SE de type 3 (20.8%) et 69 SE de type 4 (20.8%). Ces SE sont de taille très variable (de 8 à 8000 mots) avec une moyenne de 412 mots. Elles ont entre 2 et 23 items, mais 75% des SE ont seulement 2 ou 3 items. 77% des SE débutent par une amorce, moins de 8% se terminent par une clôture.

6.1.2 Motivations pour explorer les aspects lexicaux des structures énumératives ?

Comme nous l'avons montré dans la section précédente, les SE sont des objets linguistiques complexes, multi-échelles (qui se manifestent à tous les niveaux de grain) et très variées du point de vue de leur structure. Ces structures bien déterminées sont également très fortement présentes dans les textes. Elles sont caractérisées principalement par des propriétés structurelles (taille, nombre d'items, etc.).

Nous avons souligné dans le chapitre précédent les difficultés rencontrées lors de l'utilisation des voisins distributionnels dans le cadre de la segmentation thématique, difficultés essentiellement dues au caractère artificiel de la tâche et de son évaluation. Certaines SE (les SE de type 1) s'apparentent partiellement au mode d'organisation abordé lors de nos travaux sur la segmentation thématique, puisque leurs différents items correspondent à des sections du texte. Mais les SE sont des objets linguistiques bien définis, qui ont donné lieu à une campagne d'annotation, ce qui nous permet de nous appuyer sur des données fiables et maîtrisées. Les SE se caractérisent certes par une discontinuité entre items, qui induit une forme de segmentation du texte, mais elles supposent également une continuité, portée par l'énuméraThème.

Alors que la tâche de segmentation thématique repose essentiellement sur la prise en compte d'indices lexicaux, les interactions entre SE et lexique restent à déterminer. C'est pourquoi nous renouons ici avec une démarche plus qualitative : nous ne nous fixons pas pour but de détecter automatiquement les SE ou leur structure interne, mais nous situons plutôt dans une perspective de caractérisation linguistique.

Les travaux présentés dans ce chapitre ont été initiés en collaboration avec L. Tanguy, membre du projet ANNODIS « descendant », dont l'expertise dans l'exploration et la visualisation de données linguistiques complexes (*cf.* Tanguy, 2012) a été mise à profit.

Dans la suite de ce chapitre nous présentons nos premiers résultats sur la cohésion lexicale des SE (section 6.2), avant de proposer des pistes de recherche s'appuyant sur ces résultats (section 6.3).

6.2 Aborder les structures énumératives : premières visualisations, intuitions et expérimentations

Dans cette section, nous décrivons les premières expériences menées pour caractériser les SE du point de vue de leur cohésion lexicale ; nous avons notamment observé la place des SE par rapport à l'évolution de la cohésion lexicale des textes (section 6.2.1), évalué leur cohésion lexicale globale (section 6.2.2) et caractérisé la répartition des liens de cohésion lexicale en relation avec leur structure interne (section 6.2.3).

6.2.1 Aborder les structures énumératives avec des méthodes issues de la segmentation thématique

Dans un système de segmentation thématique de type TextTiling, tel que présenté dans le chapitre précédent (5), des scores de cohésion lexicale locale (à l'intérieur d'une fenêtre glissante) sont calculés pour tout le texte. Ces scores sont ensuite uniquement exploités pour repérer les vallées (c'est-à-dire les minimums locaux) les plus profondes, qui sont interprétées comme les ruptures thématiques du texte. Il nous a paru intéressant de mettre en correspondance la courbe des scores calculés par notre système de segmentation thématique avec les positions des SE au sein des textes, afin de donner un premier aperçu de leur place au sein des textes, du point de vue de la répartition de la cohésion lexicale dans ces derniers.

La figure 6.2 montre les positions des SE sur la courbe de cohésion lexicale locale de deux textes, telle que calculée par notre système de segmentation thématique. L'algorithme utilisé est celui décrit en 5.2.1 (page 154 : une fenêtre glissante parcourt le texte d'unité en unité ; à chaque étape, un score basé sur le nombre de liens lexicaux contenus par la fenêtre[2] est calculé. Les paramètres (taille de la fenêtre, type de lissage, etc.) sont ceux retenus après optimisation sur un corpus de test (*cf.* tableau 5.7 page 161).

Les figures présentées combinent ces scores avec des informations sur la position et la structure des SE. Elles ont été générées avec l'environnement R (R Development Core Team, 2012). Les SE sont représentées sur ces figures par des segments de couleur rouge. Pour permettre de visualiser leur structure interne, nous avons choisi de tracer des segments verticaux correspondant aux frontières d'items. Ainsi, si l'on se réfère à la figure 6.1, on peut voir que la SE 23 est composée d'une longue amorce suivie de deux items (non suivis d'une clôture), alors que la SE 25 n'est composée que de 4 items (sans amorce ni clôture). Ces segments ont été placés sur le graphique d'une part en tenant compte de leur position dans le texte (en tokens), et d'autre part en faisant la moyenne des scores de cohésion calculés sur

2. Plus précisément, il s'agit des liens reliant une moitié de la fenêtre à l'autre, comme montré par la figure 5.2 page 155.

l'étendue de chaque SE (pour déterminer leur position sur l'axe vertical). Les fines lignes verticales (par exemple en position 2000) indiquent les titres de section, pris comme ruptures de référence dans les expériences que nous avons menées sur la segmentation thématique.

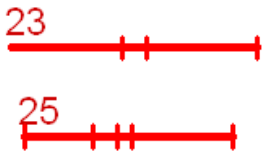


FIGURE 6.1 – Exemples de segments représentant les SE

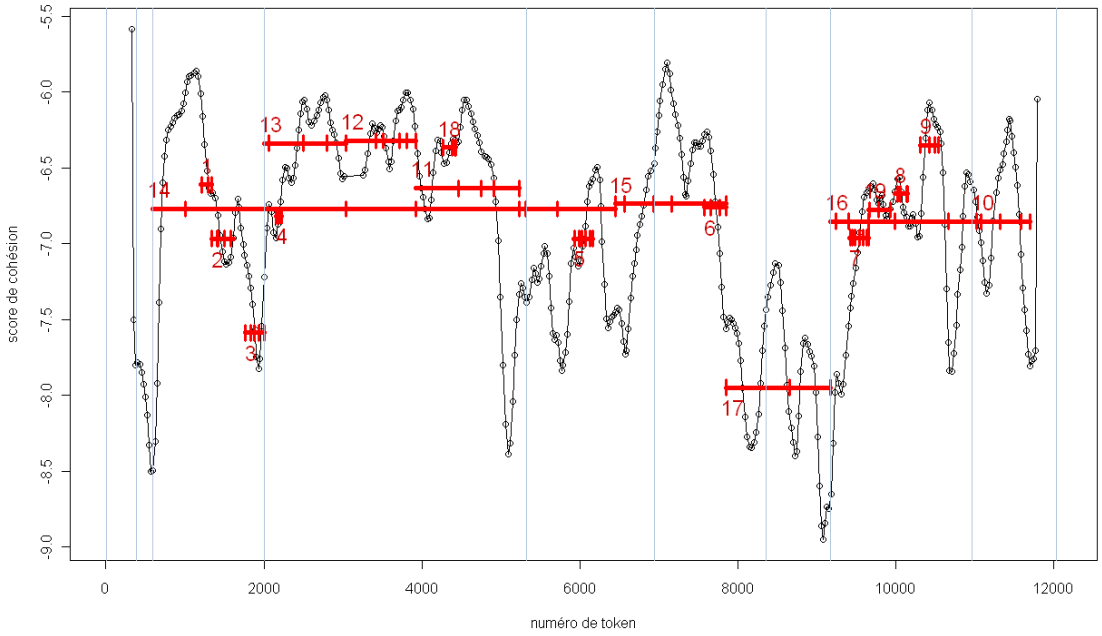


FIGURE 6.2 – Courbe de cohésion lexicale et position des SE pour le texte « Attentats du 11 septembre »

Cette méthode permet ainsi une visualisation croisant des données de natures très différentes. Au vu de ces graphiques, nous notons tout d'abord qu'aucun phénomène corrélant les positions ou la structure interne des SE à l'évolution de la cohésion lexicale du texte ne se manifeste de manière claire et systématique. L'observation de ces graphiques suscite toutefois plusieurs interrogations qui peuvent représenter autant de pistes de recherches, et que nous illustrons sur la base de la figure 6.2.

(a) Certaines frontières de SE semblent coïncider avec des creux importants dans la courbe de cohésion lexicale. Par exemple, des vallées profondes accompagnent

le début et la fin des SE 14 et 15 ; on peut par ailleurs constater que ces vallées ne concordent pas avec la présence d'un titre de section.

(b) De la même manière, certaines frontières d'items semblent être reflétées par des évolutions dans la courbe de cohésion lexicale. Ainsi, la SE 17 apparaît fortement « segmentée », avec deux pics de cohésion lexicale correspondant à ses deux items.

(c) À part quelques exceptions notables (SE 3 et 17), les SE apparaissent plutôt dans les hauteurs de la courbe de cohésion (rappelons que la position de chaque SE sur l'axe vertical est déterminée en faisant la moyenne des scores de cohésion calculés sur toute son étendue). Ainsi, on peut se demander si les SE constituent des objets présentant une cohésion interne particulièrement forte par rapport au reste du texte.

Nous soulignons à nouveau que toutes ces observations ne reflètent absolument pas des fonctionnements systématiques au sein des textes étudiés. Il s'agit plutôt de pistes à explorer pour aborder un nouvel objet – les SE – à partir de méthodes que nous maîtrisons – issues de nos expériences en segmentation thématique. C'est seulement l'exploration plus systématique de ces pistes qui permettrait de déterminer s'il s'agit d'une tendance générale des SE, d'éléments qui pourraient aider leur caractérisation (en distinguant plusieurs types de SE selon le comportement observé), ou d'un artefact de notre interprétation.

Les observations (a) et (b) incitent à confronter la segmentation thématique à des structures textuelles telles que les SE. On a en effet vu (section 5.1.1 page 146) qu'il est loin d'être évident que le mode de structuration des textes présupposé par la segmentation thématique corresponde à un fonctionnement réel des textes. Confronter la segmentation thématique à des structures linguistiques bien identifiées, mises au jour par une campagne d'annotation discursive de corpus, pourrait constituer une perspective intéressante pour cette tâche dont le mode d'évaluation est généralement très artificiel et peu satisfaisant, comme nous avons eu l'occasion de le souligner. Toutefois, les modalités d'une telle confrontation restent à définir et peuvent s'avérer complexes, du fait du caractère multi-échelle des SE et de leur fort degré de chevauchement (par exemple, la SE 18 de la figure 6.2 est incluse dans l'amorce de la SE 11, qui constitue elle-même un item de la SE 14).

Si les deux premières pistes (concernant les frontières des SE ou des items à l'intérieur des SE) reviennent à poser le problème de la position et de la structure des SE avec un point de vue très empreint des objectifs de la segmentation thématique (là encore, il s'agit de trouver des ruptures), la dernière piste mentionnée (c) revient par contre à interpréter les scores calculés d'une toute autre manière, en s'intéressant cette fois à leurs aspects jusque là ignorés : recherche de zones de cohésion et non seulement de rupture, prise en compte de la hauteur « absolue » des scores – et non profondeur relative des creux. Dans la section suivante, nous explorons l'hypothèse selon laquelle les SE présenteraient une forte cohésion lexicale interne.

6.2.2 Calculer la cohésion lexicale globale des SE

Évaluer la cohésion lexicale interne des SE est un problème relativement complexe.

La première stratégie que nous avons envisagée, et qui a été utilisée pour représenter les SE sur la courbe d'évolution de la cohésion lexicale du texte, consiste à faire pour chaque SE la moyenne de l'ensemble des scores calculés par notre système de segmentation thématique lorsque la fenêtre glissante définie se trouve à une position comprise à l'intérieur des bornes de la SE considérée. Toutefois, cette méthode ne nous a pas paru adaptée pour aborder la question de la cohésion lexicale interne des SE, pour différentes raisons illustrées par la figure 6.3. D'une part, les scores calculés au début et à la fin de chaque SE font intervenir des liens lexicaux connectant la SE à son environnement direct. Ainsi, une SE en relation de continuité thématique avec le texte la précédant apparaîtra plus cohésive qu'une SE introduisant un nouveau thème. Cette limite concerne toutes les SE, mais est plus critique lorsque la SE considérée est de plus petite taille que la fenêtre glissante déployée : tous les scores calculés sur son étendue impliquent alors des liens avec des items lexicaux se situant à l'extérieur de la SE. D'autre part, si la SE est de taille supérieure à celle de la fenêtre glissante, les liens connectant des mots apparaissant au début et à la fin de la SE ne sont alors jamais pris en compte car ils ne sont pas compris dans la portée de la fenêtre ; or, il nous paraît dommageable d'évaluer la cohésion lexicale interne des SE sans prendre en compte la possibilité que, par exemple, l'amorce d'une SE puisse être connectée avec son dernier item ou sa clôture.

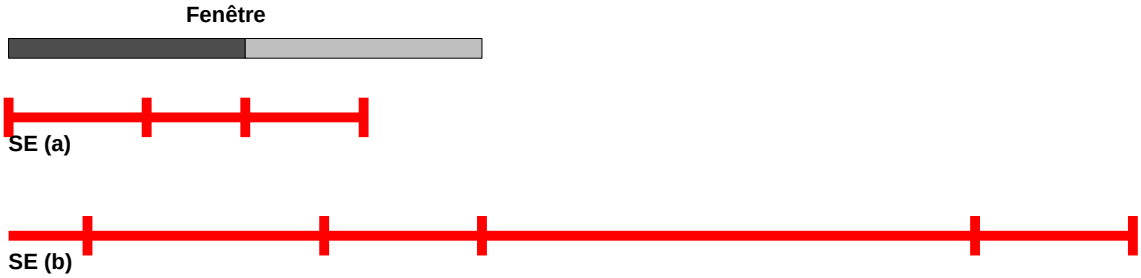


FIGURE 6.3 – Représentations de deux SE de tailles inférieure et supérieure à la taille de la fenêtre glissante

Nous avons donc mis en œuvre une autre méthode, permettant d'évaluer la cohésion lexicale globale d'une SE indépendamment de sa taille. Pour chaque SE, nous comptons tous les liens de voisinage qu'elle contient (et non uniquement ceux connectant les moitiés gauche et droite d'une fenêtre glissante de taille fixe). Nous normalisons ce score en fonction de la taille de la SE, mais comme nous l'avons vu (section 3.4.2 page 91), les effets d'échelle sont difficiles à compenser pleinement par la normalisation. Nous parcourons alors le texte avec une fenêtre glissante de même taille (en nombre de tokens) que la SE considérée, afin de calculer le même score à différents points du texte. La moyenne (et l'écart type) des scores calculés

sur l'ensemble du texte offre une référence de la cohésion lexicale attendue pour une taille de segment donnée. Nous calculons alors pour chaque SE son écart réduit à la moyenne, c'est-à-dire sa différence signée à la moyenne normalisée par l'écart type. Le tableau 6.2 montre un exemple de calcul pour une SE de taille 1461 tokens. Son score de cohésion normalisé est de 2.955. La moyenne des scores de cohésion calculés sur l'ensemble du texte pour des segments de 1461 tokens est de 2.496, avec un écart type de 0.281. Le score calculé pour la SE considéré est ainsi supérieur à la moyenne de 1.634 écart type.

Nombre de tokens	1461
Nombre de liens	3152
Score de cohésion×1000	2.955
Moyenne des scores de cohésion×1000	2.496
Écart type×1000	0.281
Écart réduit	1.634

TABLEAU 6.2 – Cohésion lexicale d'une SE exprimée par l'écart réduit à la cohésion moyenne

L'écart réduit est plus facilement interprétable que le score brut : un écart réduit positif indique que la SE est plus cohésive que la moyenne, tandis qu'un écart réduit négatif indique qu'elle est moins cohésive. La cohésion lexicale d'une SE telle qu'exprimée par l'écart-réduit calculé permet de comparer entre elles des SE de tailles très différentes.

La moyenne des écarts réduits calculés pour l'ensemble des SE permet d'apprécier si celles-ci présentent une tendance à être plus fortement cohésives. Cette moyenne est de 0.37, ce qui semble confirmer l'hypothèse. Toutefois, cette conclusion est à mitiger : l'observation des données montre que des cas extrêmes (SE présentant des écarts-réduits allant jusqu'à 13) perturbent ce résultat. Ce phénomène est visible dans la figure 6.4, qui permet de visualiser l'ensemble des écarts réduits calculés pour les 331 SE de nos données. Si l'on adopte une démarche plus robuste en écartant ces cas extrêmes [3], on constate alors que les SE ne sont pas significativement plus cohésives que la moyenne du texte.

Plusieurs raisons peuvent expliquer ce résultat. Concernant la méthode déployée, on peut remarquer que la cohésion lexicale de chaque SE est évaluée à partir d'une moyenne calculée en faisant circuler une fenêtre de taille identique sur l'ensemble du texte. Or, on a vu que les textes sont couverts par des SE à 60% environ. La fenêtre glissante traverse donc de nombreuses autres SE, ainsi que la SE cible. Une comparaison de la cohésion lexicale des SE avec la cohésion de pans de textes non-couverts par des SE donnerait peut-être un résultat différent. Mais, au delà de ces modifications pouvant être faites à notre protocole, le résultat obtenu ne nous paraît

3. ceux dont l'écart à la moyenne des écarts réduits dépasse trois écarts types, *cf.* figure 6.4

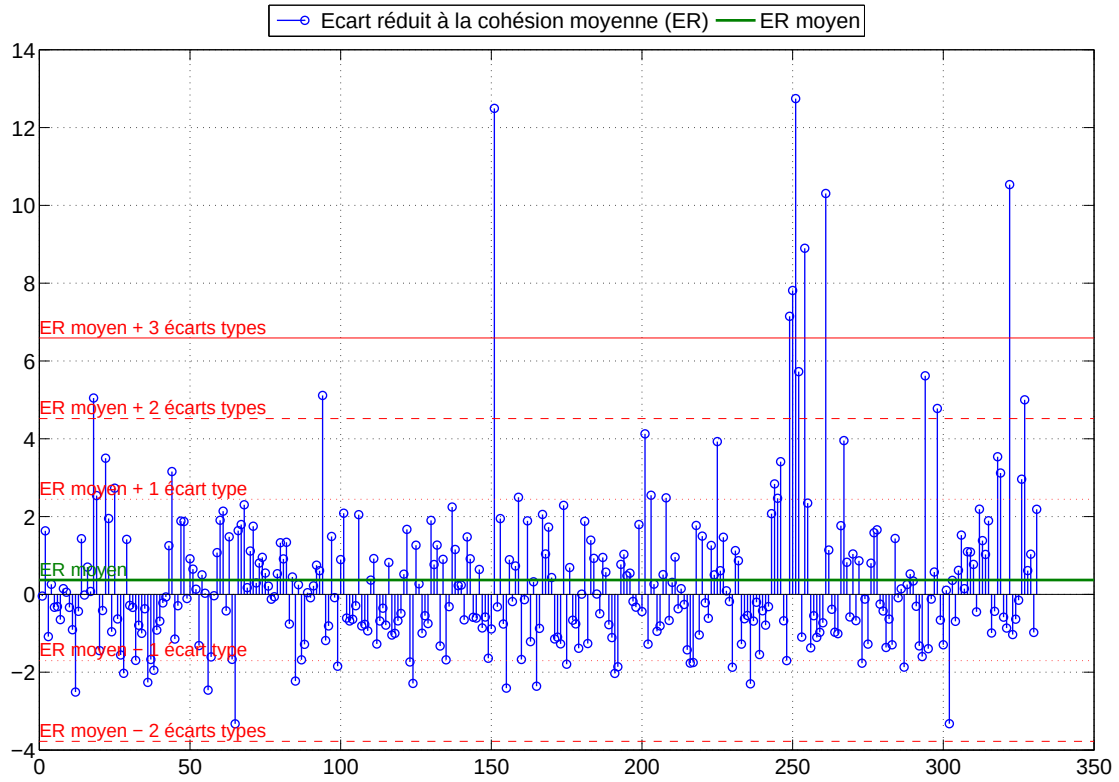


FIGURE 6.4 – Représentation des écarts réduits calculés pour l'ensemble des SE

pas contre-intuitif. Nous avons présenté les SE comme des lieux de continuité (sur un plan) mais aussi de discontinuité (sur un autre plan) ; une SE peut notamment présenter de la discontinuité sur le plan thématique (nous avons vu dans la section 5.2.3 page 159 qu'une structuration thématique s'exprime plus facilement en termes de variations de cohésion lexicale qu'une structuration temporelle ou rhétorique). Ainsi, nous avons noté, à partir de la visualisation proposée avec la figure 6.2, que la structuration en items pouvait correspondre à de fortes variations de cohésion lexicale.

Si une forte cohésion lexicale ne semble pas constituer une tendance générale des SE, notre approche a par contre permis de pointer des fonctionnements marqués et divergents : il y a des SE très cohésives (beaucoup plus que la moyenne du texte) et d'autres qui le sont très peu. La cohésion lexicale des SE pourrait donc participer à leur caractérisation. Nous avons calculé les coefficients de corrélation entre la cohésion lexicale des SE et les autres variables qui étaient à notre disposition : leur type et l'ensemble de leurs caractéristiques : présence ou non d'une amorce, d'une clôture, d'un énumérathème explicite, nombre de mots, nombre d'items, signalement par les différents types d'indices, etc. Aucune corrélation significative n'a pu être mise au jour. Ainsi, la typologie actuelle des SE, basée sur des critères essentiellement structurels, n'est pas appuyée par des différences significatives du point de vue du

lexique. Le fait qu'il n'y ait aucune corrélation significative entre la cohésion lexicale d'une SE et sa taille (en nombre de mots ou d'items) montre que notre méthode a bien pris en compte les différences d'échelle. On peut se demander à se stade si la cohésion lexicale des SE pourrait être corrélée avec d'autres propriétés (sémantiques ou rhétoriques) des SE, mais aucune typologie des SE selon de tels critères n'a été effectuée à ce jour. Nous développerons partiellement cette piste de recherche dans la section 6.3.2.

6.2.3 Les SE, un lieu d'observation de liens « à distance » ?

Lors de la procédure qui nous a permis d'évaluer la cohésion lexicale globale des SE, nous avons souhaité tester une hypothèse corolaire : si les SE pouvaient selon nous constituer des zones de forte cohésion lexicale, c'était parce que le critère fondamental de la coénumérabilité semblait garantir une certaine homogénéité thématique minimale entre tous les items. Nous nous attendions donc à trouver à l'intérieur des SE plus de liens « distants », par exemple connectant l'amorce ou les premiers items aux derniers items ou à la clôture.

Nous illustrons cette hypothèse à partir des exemples (2) et (3).

L'exemple (2) est une liste formatée (SE de type 2) de deux items, précédés d'une amorce et suivis d'une clôture. On peut remarquer dans cette SE la grande proximité lexicale entre l'amorce et la clôture, avec notamment une forte proportion de vocabulaire partagé : « individus », « reproduire », « avantage sélectif », etc. En plus de ces répétitions, de nombreux liens de cohésion lexicale sont détectés par la projection des voisins : `individu`/`milieu`, `environnement`/`milieu`, `reproduire`/`reproduction`, `avantage`/`capacité`, `variation`/`trait`, etc.

(2) **[**Certains **individus** portent des **variations** qui leur permettent de se **reproduire** davantage que les autres, dans un **environnement** précis. On dit qu'ils disposent d'un **avantage sélectif** sur leurs congénères :**]**$_{AMORCE}$
[- La première possibilité est, par exemple, qu'en échappant mieux aux prédateurs, en étant moins malades, en accédant plus facilement à la nourriture, ces individus atteignent plus facilement l'âge adulte, pour être apte à la reproduction. Ceux qui ont une meilleure capacité de survie pourront donc se reproduire davantage.**]**$_{ITEM\ 1}$
[- Dans le cas particulier de la reproduction sexuée, les individus ayant survécu peuvent être porteurs d'un caractère particulièrement attirant pour les partenaires de sexe opposé. Ceux-là seront capables d'engendrer une plus grande descendance en copulant davantage.**]**$_{ITEM\ 2}$
[Dans les deux cas, l'augmentation de la **capacité** à survivre et à se **reproduire** se traduit par une augmentation du taux de **reproduction** et donc par une descendance plus nombreuse, pour les **individus** porteurs de ces caractéristiques. On dit alors que ce **trait** de caractère donné offre un **avantage sélectif**, par rapport à d'autres. C'est dans ce principe d'adaptation

uniquement, qu'intervient le **milieu** de vie.]$_{CLOTURE}$
(Wikipédia, article « Sélection naturelle »)

L'exemple (3) est une SE intrapragraphique (type 4) constituée de 6 items, sans amorce ni clôture. Du point de vue de la cohésion lexicale, Les différents items de cette SE apparaissent tous fortement inter-connectés, notamment *via* deux principales composantes :

- `est`, `ouest`, `nord`, `sud`, `côte`, `frontière` ;
- `altitude`, `colline`, `massif`, `plateau`, `sommet`, `bois`, `forêt`, `plaine`, `vallée`, `cordillère`, `pâturage`, `prairie`, `sierra`, etc.

Le tableau 6.3 résume les nombres de liens (de voisinage et de répétition) connectant chaque item aux autres items. La quantité de liens connectant deux items semble ici indépendante de la distance entre ces deux items.

(3) [À l'**est**, les Pyrénées sont hautes et parsemées de forêt et de pâturages d'altitude avec des vallées assez profondes.]$_{ITEM\ 1}$ [À l'**ouest**, la chaine Pyrénéenne est plus calme et forme des plateaux herbeux et des sommets arrondis jusqu'à l'Océan.]$_{ITEM\ 2}$ [Au **nord** du massif, en Iparralde, les collines vertes dominent jusqu'à l'Adour. On y trouve des prairies, des bois et des champs cultivés de maïs.]$_{ITEM\ 3}$ [Au-dessus de l'Adour se forme une plaine alluviale marécageuse appelée les barthes.]$_{ITEM\ 4}$ [Au **sud** de la chaine axiale, en Navarre, les Pyrénées sont présentes tout le long de la **frontière** et se prolongent jusqu'à la **côte** basque espagnole avec des vallées plus vertes et moins étroites.]$_{ITEM\ 5}$ [À l'**ouest**, dans la communauté autonome basque, on trouve la cordillère Cantabrique qui se prolonge vers Bilbao. Elle est formée d'une succession de sierras : la sierra d'Aralar, la sierra d'Urbasa et la sierra d'Andia.]$_{ITEM\ 6}$
(Wikipédia, article « Pays basque »)

	It1	It2	It3	It4	It5	It6
Item 1		**8**	14	8	4	5
Item 2			**4**	7	2	5
Item 3				**8**	16	8
Item 4					**4**	5
Item 5						**1**

TABLEAU 6.3 – Nombres de liens entre les différents items de la SE (3)

Lors de la procédure décrite dans la section 6.2.2, nous avons modélisé la longueur moyenne attendue pour les liens de cohésion lexicale, en calculant la moyenne et l'écart type sur une fenêtre de taille identique à celle de la SE et parcourant l'ensemble du texte. Le tableau 6.4 montre un exemple de calcul : les liens lexicaux repérés dans cette SE de 1461 tokens « mesurent » en moyenne 487 tokens ; ce

nombre est supérieur de 1.21 écart type au nombre moyen observé sur des zones de même taille à différents points du texte.

Nombre de tokens	1461
Longueur moyenne des liens	486.95
Moyenne des longueurs moyennes des liens	470.47
Écart type	13.62
Écart réduit	1.21

TABLEAU 6.4 – Longueur moyenne des liens d'une SE exprimée par l'écart réduit à la moyenne

Cette procédure, répétée pour l'ensemble des SE, a permis de montrer que celles-ci présentent des liens significativement plus longs que ce qui serait attendu étant donnée leur taille. Ainsi, bien que les SE n'apparaissent pas particulièrement denses du point de vue de leur cohésion lexicale, l'hypothèse selon laquelle elles se caractériseraient par plus de liens « à distance » se vérifie.

Afin d'affiner ce résultat, nous avons souhaité comparer les nombres de liens connectant deux éléments d'une SE, selon que ces éléments soient adjacents ou non. Nous avons vu (notamment dans la section 4.4.3 page 124) que le nombre de liens de cohésion lexicale pertinents connectant deux zones de texte est fortement corrélé à la distance entre ces zones de texte. Si cette observation se reportait sur les SE, on s'attendrait donc à ce que les éléments adjacents soient plus fortement connectés. Mais une répartition différente expliquerait la longueur supérieure des liens à l'intérieur des SE.

Nous avons pour chaque SE extrait une matrice similaire à celle présentée dans le tableau 6.3. Dans cette matrice, les liens entre items contigus sont représentés par les nombres apparaissant le long de la diagonale (en gras). Les liens entre items non contigus sont représentés par l'ensemble des autres nombres. On peut donc constater que les items contigus sont reliés par 5 liens en moyenne [4], tandis que les items non-contigus sont reliés par 7.4 liens en moyenne [5]. Le nombre de liens ne peut toutefois pas être considéré tel quel, car les items n'ont pas tous une taille égale, comme illustré dans le tableau 6.5. La normalisation apparaissant la plus adaptée

It1	It2	It3	It4	It5	It6
07	07	16	13	15	05

TABLEAU 6.5 – Taille des items de la SE (3), en nombre de tokens

est de diviser le nombre de liens observé entre deux items par le nombre maximum

4. $(8+4+8+4+1)/5 = 5$
5. $(14+8+4+5+7+2+5+16+8+5)/10 = 7.40$

de liens possibles entre ces items. Ce nombre maximum est simplement le produit de la taille des deux items en question. Le tableau 6.6 illustre le résultat de cette normalisation pour l'exemple de la SE (3). Pour cet exemple, le taux de lien entre

	It1	It2	It3	It4	It5	It6
Item 1		**0.1633**	0.1250	0.0879	0.0408	0.1429
Item 2			**0.0357**	0.0769	0.0204	0.1429
Item 3				**0.0385**	0.0714	0.1000
Item 4					**0.0220**	0.0769
Item 5						**0.0143**

TABLEAU 6.6 – Taux de liens entre les différents items de la SE (3), normalisés par la taille des items

items contigus est de 5.47%, tandis que le taux de liens entre items non-contigus est en moyenne de 8.85%.

Nous avons répété ces calculs sur l'ensemble des SE comprenant au moins 3 items (les SE à deux items ne présentant pas d'items non adjacents), c'est-à-dire 232 SE. Le tableau 6.7 présente les taux de liens obtenus sur l'ensemble de ces données. Il s'agit de moyennes établies sur 95% des données, en excluant les 5% de cas les plus extrêmes, c'est-à-dire présentant les taux de liens s'écartant le plus du taux moyen pour l'ensemble des SE. Nous avons également appliqué la même procédure pour calculer les nombres de liens moyen entre une amorce et le premier item d'une SE *vs* entre une amorce et les autres items.

Nombre de SE analysées	Objets comparés	Taux de liens moyen (±écart type) ×100
219	Paire d'items contigus	5.04 (±5.03)
	Paire d'items non contigus	4.49 (±4.51)
207	Amorce + item 1	3.24 (±2.75)
	Amorce + autre item	3.30 (±2.18)
023	Clôture + dernier item	3.36 (±2.75)
	Clôture + autre item	3.23 (±1.86)

TABLEAU 6.7 – Taux de liens moyens selon la proximité entre éléments d'une SE

Les données du tableau 6.7 permettent de conclure que la différence entre le taux de liens moyen entre items contigus et non-contigus est faible (de l'ordre de 10%) par rapport à ce taux moyen. Si l'on examine ces différences individuellement pour chacune des SE (figure 6.5), on peut constater que la différence est minime pour la grande majorité des SE.

Pour un nombre restreint de cas (14 dans la figure 6.5), on observe plutôt un comportement plus linéaire, avec plus de liens adjacents que éloignés, ce qui est plus similaire à ce qui est observé en général dans les textes, comme nous l'avions illustré

par la figure 3.7 (page 82). La présence de ce type de SE contribue partiellement à expliquer le fort écart-type observé pour les taux de liens moyens.

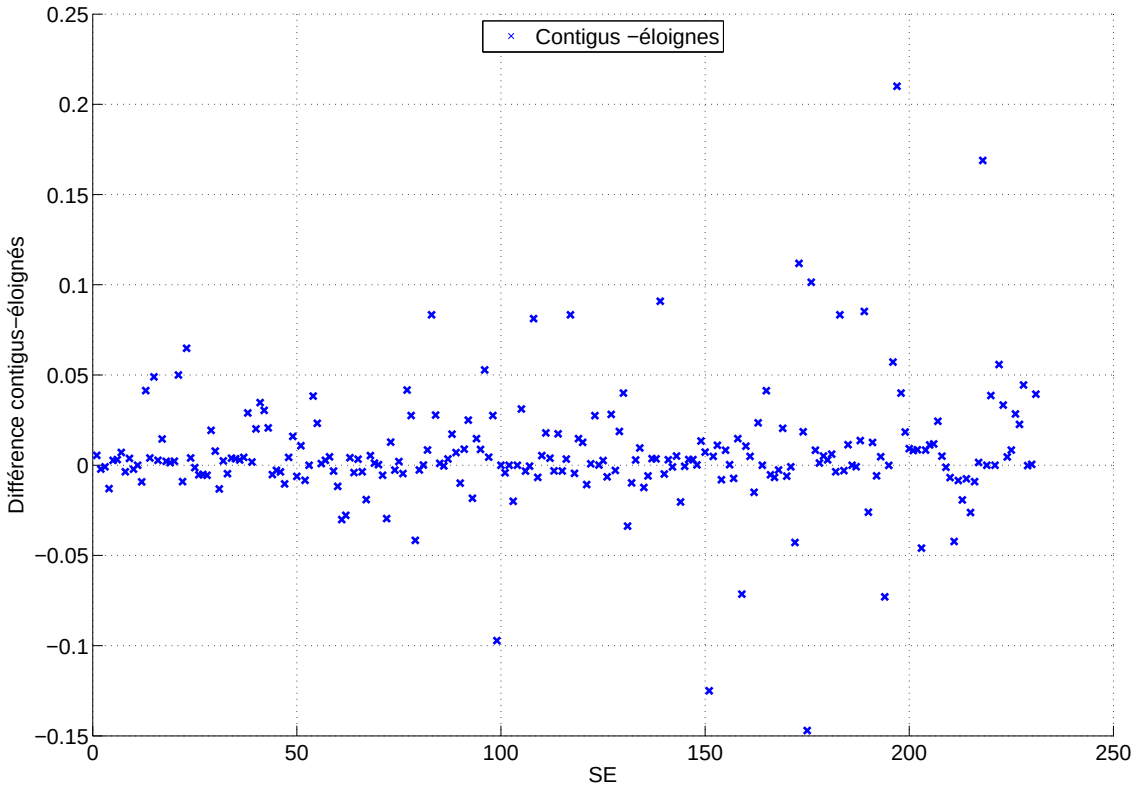


FIGURE 6.5 – Différence des taux de liens entre items contigus et éloignés pour 232 SE

Finalement, on peut aussi observer que les amorces et clôtures sont moins liées aux items que les items entre eux. Et pour ces deux éléments, on ne constate aucune augmentation du taux de liens pour l'élément contigu.

6.3 Vers une observation plus fine : quelques pistes de recherche

Nous avons présenté dans la section 6.2 quelques premiers résultats concernant la cohésion lexicale des SE. Partant de ces résultats préliminaires, nous discutons ici des pistes de recherche qui pourraient être poursuivies par la suite.

6.3.1 Observer le contenu lexical des SE en interaction avec leur structure interne

Nous avons vu dans la section 6.2.3 que les SE se caractérisent par une répartition particulière des liens de cohésion lexicale. Nous avons appréhendé cette répartition particulière à partir de deux indicateurs :

- la longueur moyenne des liens lexicaux : elle est significativement plus grande dans les SE que dans des segments de texte de même taille ;
- le nombre de liens connectant l'amorce à un item ou un item à un autre : sur l'ensemble des SE, ce nombre n'est pas significativement plus important lorsque les deux objets considérés sont contigus que lorsqu'ils sont distants.

Ces résultats quantitatifs incitent à observer plus finement le contenu lexical des SE en interaction avec leur structure interne, afin de proposer des pistes d'interprétation linguistique des tendances mesurées. Pour cette phase exploratoire d'analyse, une visualisation appropriée, croisant les différents types de données (structures énumératives et liens lexicaux), est nécessaire. Notre travail s'est donc concentré sur la définition d'une stratégie de sélection et de visualisation des informations lexicales associées aux SE.

Nous avons déjà noté les difficultés soulevées par la visualisation de l'information liée à la cohésion lexicale (section 3.3.3 page 82). À ces difficultés s'ajoutent celles découlant du caractère multi-échelle des SE, rendant peu pratique une visualisation des liens lexicaux en texte, et incitant à adopter une approche plus synthétique. Nous avons finalement opté pour la stratégie suivante : nous avons extrait pour chaque SE les composantes connexes (*cf.* section 3.4.1.2 page 89) couvrant tous les éléments de cette SE, excluant donc les composantes internes à un item ou ne couvrant qu'une partie des items. Nous affichons ensuite ces composantes en rendant compte de la structure des SE par des paliers verticaux respectant l'ordre de ses différents éléments (amorce, items, clôture).

Les figures 6.6, 6.7 et 6.8, réalisées par L. Tanguy, présentent des exemples de graphes obtenus par cette stratégie de sélection / visualisation de nos données. Ces graphes ont été calculés à partir de SE de formats variés.

La figure 6.6 est construite à partir d'une SE de type 4, c'est-à-dire intra-paragraphique. Les relations lexicales présentées laissent entrevoir une continuité thématique sur l'ensemble de la SE autour du thème de la politique.

La figure 6.7 est construite à partir d'une longue SE de type 3, dont chaque item est un paragraphe entier (la SE couvre donc 5 paragraphes). Ici, l'extraction de composantes couvrant l'ensemble des items a fait émerger une clique regroupant plusieurs mois de l'année, rendant compte de l'organisation temporelle de cette SE qui décrit différentes étapes du scandale du Watergate, auquel est consacré l'article.

Enfin, la figure 6.8 est construite à partir d'une SE de type 2 (c'est-à-dire une liste formatée) provenant de l'article « Vin de Champagne ». Bien que cette SE soit plus courte que celle représentée en 6.7, l'information lexicale extraite est bien plus riche. On devine une SE thématiquement très homogène, présentant différents

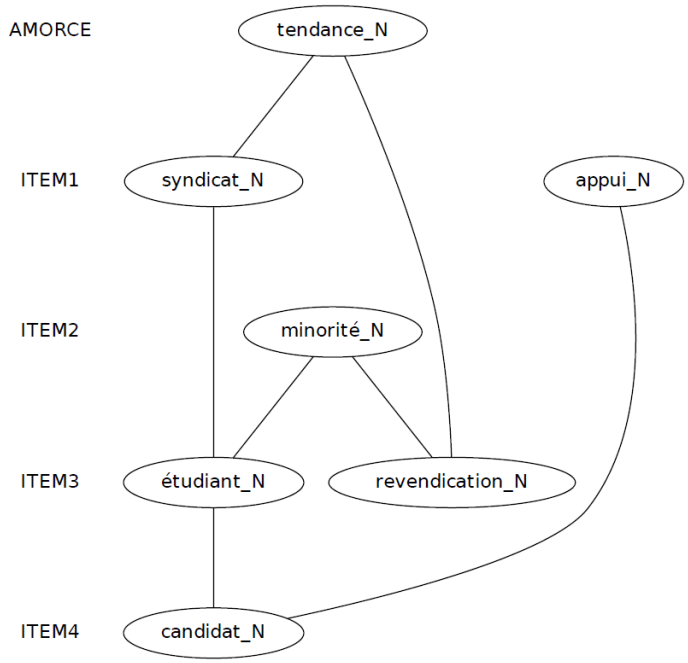


FIGURE 6.6 – Informations lexicales associées à une SE de type 4 appartenant à l'article « Scandale du Watergate »

cépages du vin de Champagne.

Finalement, ces graphes offrent une vision synthétique, qui met l'accent sur une particularité des SE : une répartition particulière des liens de cohésion lexicale, qui tend à mettre sur le même plan les différents items successifs. Les résultats que nous présentons montrent un aspect moins abordé de la cohésion lexicale : ici, les voisins distributionnels sont envisagés non pas comme un indice de repérage des structures, mais comme un moyen de faciliter l'accès à une représentation de ce qui fait l'unité lexicale d'une structure discursive particulière. Il serait intéressant d'étudier le lien entre les relations lexicales mise au jour par cette méthode et l'énuméraThème, qui est le critère thématique qui permet de regrouper les différents items.

6.3.2 Exploiter la cohésion lexicale pour aider au typage des SE

L'autre piste recherche que nous avons commencé à explorer consiste à partir des fonctionnements marqués que nous avons soulignés dans la section 6.2.2 (certaines SE ont un score cohésion lexicale très élevé ; d'autres sont au contraire très peu cohésives ; les SE ont également des fonctionnements différents du point de vue de la répartition des liens de cohésion lexicale qu'elles contiennent) afin de déterminer s'ils peuvent être associés à certaines propriétés des SE. Nous avons déjà établi que les scores de cohésion lexicale n'étaient corrélés avec aucune des caractéristiques

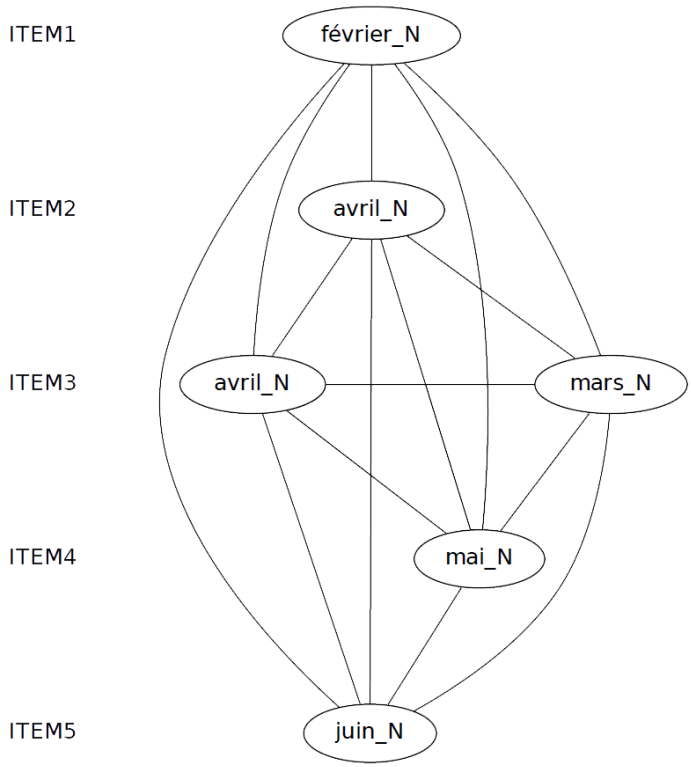


FIGURE 6.7 – Informations lexicales associées à une SE de type 3 appartenant à l'article « Scandale du Watergate »

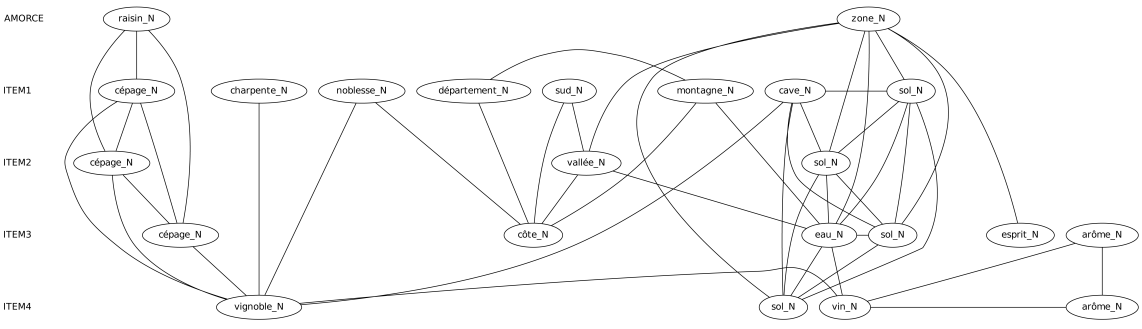


FIGURE 6.8 – Informations lexicales associées à une SE de type 2 appartenant à l'article « Vin de Champagne »

structurelles des SE (nombre de mots, d'items, présence d'une amorce ou d'une clôture, etc.). Il s'agirait ici d'investiguer la relation entre cohésion lexicale et d'autres propriétés des SE. Par exemple :

- les SE peuvent s'articuler selon différents plans : thématique, temporel, rhétorique... ; il serait intéressant d'observer si ces modes d'organisations sont appuyés par des différences du point de vue de la cohésion lexicale des SE ;
- si l'on se place du point de vue d'une analyse en relations de discours, la structure interne des SE peut s'analyser avec différentes relations (Bras *et al.*, 2008) : plus particulièrement, les différents items peuvent entrer dans des relations de narration, d'élaboration et d'explication (par rapport à l'amorce) ; or, nous verrons dans le chapitre qui suit que ces relations sont différentes du point de vue de leur cohésion lexicale (l'explication par exemple est moins cohésive que l'élaboration) ; dans l'optique d'une typologie des SE selon ce critère, des indices lexicaux pourrait donc être utilisés ;
- les SE peuvent être qualifiées de syntagmatiques ou de paradigmatiques (Luc, 2000, 2001) :
  - *énumération syntagmatique* : tous les items entretiennent une relation de dépendance (syntaxique ou rhétorique) les uns par rapports aux autres.
  - *énumération paradigmatique* : les items sont fonctionnellement équivalents (syntaxiquement et rhétoriquement) au sein de l'énumération.

  (Luc, 2001)

  La répartition des liens lexicaux au sein des SE pourrait peut-être être rapprochée de leurs fonctionnements paradigmatiques ou syntagmatiques.

Les annotations n'ayant pas porté sur ces aspects, il est difficile de répondre à ces questions. La stratégie que nous avons adoptée consiste à extraire des sous-ensembles de SE se caractérisant par une cohésion lexicale très faible ou très forte afin de les analyser. Plus particulièrement, nous avons observé :

- les SE dont le score de cohésion est supérieur à 4 écarts types à la moyenne (14 SE) ;
- les SE dont le score de cohésion est inférieur à -2 écarts types à la moyenne (12 SE).

Parmi les SE faiblement cohésives :

- 3 SE sont des énumérations de noms propres ou comprenant beaucoup de noms propres : noms de logiciels (*cf.* exemple (4)), villes du pays Basque, ou auteurs de romans ;
- les 8 SE restantes ont toutes une organisation temporelle, parfois étalée sur plusieurs sections ; l'exemple (5) montre une SE faiblement cohésive et présentant une organisation temporelle. Une corrélation semble ainsi exister entre faiblesse des scores de cohésion lexicale et organisation des SE selon un axe temporel.

(4) **[**Des exemples de logiciels donnés à titre indicatif : **]**$_{AMORCE}$
[- la bureautique avec OpenOffice.org,**]**$_{ITEM\ 1}$

[- Internet avec Mozilla Firefox, Konqueror, IceWeasel, Gnuzilla, Mozilla Thunderbird, Pidgin ou BitTorrent,]$_{ITEM\ 2}$
[- le multimédia avec Xine, MPlayer, VLC media player, XMMS ou Amarok,]$_{ITEM\ 3}$
[- le graphisme, avec GIMP, Inkscape ou Scribus, la 3D avec Blender .]$_{ITEM\ 4}$
(Wikipédia, article « Linux »)

(5) [Les manifestations de la faim se multiplient.]$_{AMORCE}$ [En mars 1930, 35 000 personnes défilent dans les rues de New York.]$_{ITEM\ 1}$ [En juin 1932, les Anciens Combattants réclament le paiement des pensions à Washington DC : ils sont violemment délogés par les soldats.]$_{ITEM\ 2}$ [Une grande grève dans le secteur du textile éclate en 1934.]$_{ITEM\ 3}$ [Dans les campagnes, la situation économique se dégrade, notamment à cause de la sécheresse et du Dust Bowl (1933–1935). En 1933, la diminution de 60 % des prix agricoles affecte durement les agriculteurs (effet ciseaux). La ruine des fermiers des Grandes Plaines poussent des milliers de personnes à s'installer dans les États de l'Ouest.]$_{ITEM\ 4}$
(Wikipédia, article « Grande Dépression »)

Les SE fortement cohésives sont plus difficiles à catégoriser. Sur 14 SE, une seule présente une organisation temporelle, la SE (6) ; les liens de répétition et de voisinage reliant les différentes occurrences d'`embarquer`, `débarquer`, `équipage` et `passager` (qui forment une clique) font de cette courte SE une zone très dense en cohésion lexicale.

(6) [Le 11 avril 1912 à 11h30, le Titanic arrive à Queenstown où débarquent sept passagers inter-ports et 120 passagers embarquent. Les passagers qui embarquent à ce moment là sont en grande majorité des passagers de 3e classes immigrant vers les États-Unis.]$_{ITEM\ 1}$ [À 13h30, le RMS Titanic quitte Queenstown pour New York avec à son bord 1 316 passagers et 885 membres d'équipage.]$_{ITEM\ 2}$
(Wikipédia, article « Titanic »)

On trouve également des SE dont les items présentent un fort parallélisme, comme dans l'exemple (7). Les deux items de cette SE partagent une grande partie de leur structure et de leur vocabulaire. Les parties variables mettent souvent en parallèle des antonymes, tous identifiés par la projection des voisins : (*ce qui est de l'ordre du*) `sujet` *vs* `objet` ; `reconnaissance` *vs* `déni` (*du droit au libre arbitre*) ; `adulte` ou `parent` *vs* `enfant`.

(7) [Dans l'étendue des choses qu'il traite en matière de droit des personnes, et pour éclairer le sujet, le droit peut s'analyser en deux notions complémentaires :]$_{AMORCE}$
[- ce qui est de l'ordre du sujet : reconnaissance du droit au libre arbitre, responsabilité de ses actes. Ceci s'applique aux adultes (parents en particulier),

qui sont par ailleurs rédacteurs des règles et des lois qu'ils s'appliquent, par délégation à travers l'élection des députés qui font la loi.]$_{ITEM\ 1}$
[- ce qui est de l'ordre de l'objet : déni ou forte restriction du droit au libre arbitre, responsabilité absente ou limitée, existence d'une protection. Ceci s'applique aux choses, aux évènements, aux animaux, et en matière de personnes aux enfants mineurs et aux incapables. Les objets du droit ne participent pas à sa rédaction, ils le subissent.]$_{ITEM\ 2}$
(Wikipédia, article « Pédophilie »)

Poussé à l'extrême, ce parallélisme est réalisé par des SE constituées de listes de syntagmes nominaux, comme dans l'exemple (8). Dans cet exemple, `membre`, `sénateur`, `conseiller` et `président` forment une clique ; le lien `démocrate` / `républicain` est également repéré.

(8) [Les membres de la commission d'enquête sénatoriale :]$_{AMORCE}$
[- Howard H. Baker, sénateur républicain du Tennessee]$_{ITEM\ 1}$
[- Samuel Dash, conseiller démocrate]$_{ITEM\ 2}$
[- Sam J. Ervin, président de la commission, sénateur démocrate de la Caroline du Nord]$_{ITEM\ 3}$
[- Edward J. Gurney, sénateur républicain de la Floride]$_{ITEM\ 4}$
[- Daniel K. Inouye, sénateur démocrate de Hawaii]$_{ITEM\ 5}$
[- Joseph M. Montoya, sénateur démocrate du nouveau Mexique]$_{ITEM\ 6}$
[- Herman E. Talmadge, sénateur démocrate de Géorgie]$_{ITEM\ 7}$
[- Fred D. Thompson, conseiller républicain]$_{ITEM\ 8}$
[- Lowell P. Weicker, Jr, sénateur républicain du Connecticut]$_{ITEM\ 9}$
(Wikipédia, article « Scandale de Watergate »)

D'autres listes similaires apparaissent parmi les SE fortement cohésives que nous avons observées : listes de personnes associées à leurs fonctions : 5 SE ; autres listes de co-hyponymes : 2 SE.

Finalement, cette démarche permet de dégager des caractéristiques associées au caractère faiblement ou fortement cohésif des SE. Notamment, une typologie des SE selon le plan suivant lequel elles s'organisent bénéficierait certainement d'une prise en compte de la cohésion lexicale.

6.4 Bilan et perspectives

Dans ce chapitre, nous avons présenté les travaux que nous avons menés en exploitant les annotations effectuées lors du projet ANNODIS « descendant ».

Quelques premiers résultats ont été présentés. Nous avons notamment vu que les SE, bien qu'étant des objets présentant une certaine homogénéité (garantie par la contrainte de coénumérabilité) ne se caractérisent pas forcément par une forte

cohésion lexicale interne. Bien que les SE énumératives ne puissent pas être significativement associées à une forte densité de liens lexicaux, nous avons montré qu'elles se caractérisent par une répartition particulière de ces liens : la distance moyenne entre items lexicaux connectés au sein d'une SE est significativement supérieure à celle observée dans des zones de texte de même taille. L'explication à ce phénomène réside dans la tendance des différents éléments de la SE a présenter entre eux une connectivité qui n'est pas fonction de leur distance : deux items contigus ne présentent pas une cohésion lexicale significativement plus forte que deux items non-contigus. Les différents indices mesurés (force de la cohésion lexicale et longueur des liens) n'ont pas pu être corrélés avec des caractéristiques structurelles des SE.

Partant de ces résultats préliminaires, nous avons proposé deux pistes de recherche :

- la première consiste à observer les interactions entre contenu lexical et structure interne des SE ;
- la seconde consiste à partir des divergences que font émerger les indices liés à la cohésion lexicale pour déterminer s'ils peuvent être corrélés à certaines caractéristiques des SE ; il pourrait notamment s'agir de leur nature paradigmatique ou syntagmatique, ou du type d'organisation qu'elles mettent en œuvre (plans sur lesquels elles présentent de la continuité ou de la discontinuité).

Nos perspectives résident donc essentiellement dans la continuation de ces pistes de recherches, ce chapitre s'étant surtout attaché à montrer la démarche pouvant être adoptée pour aborder un nouvel objet du point de vue de sa cohésion lexicale. L'exploration de données a en effet ses propres problématiques, avec notamment le problème crucial de la visualisation des informations linguistiques, auquel nous avons accordé une place importante dans ce chapitre (avec les modes de visualisation proposés en 6.2.1 et 6.3.1).

Chapitre 7

Structure rhétorique du discours

Sommaire

Ce chapitre s'intéresse au discours tel que modélisé par des approches hiérarchiques et ascendantes. Il s'inscrit à nouveau dans une collaboration avec le projet ANNODIS, cette fois dans son versant dit « ascendant ».

Après avoir présenté le contexte théorique et la constitution du corpus ANNODIS « ascendant » (section 7.1), nous montrons qu'il est possible de faire appel à la cohésion lexicale pour aborder les deux principales problématiques de l'analyse en relations de discours : l'attachement des segments de discours et la détermination de la relation liant deux segments (section 7.2). Un cas particulier, correspondant à un problème réel rencontré lors de la campagne d'annotation ANNODIS, est ensuite présenté ; il s'agit de la confusion par les annotateurs de deux relations de discours : élaboration et élaboration d'entité. Nous proposons une approche par apprentissage automatique pour répondre à ce problème, et montrons l'impact de la prise en compte de la cohésion lexicale dans un tel système (section 7.3).

7.1 Contexte

Dans cette section, nous introduisons brièvement les approches relationnelles du discours et posons le problème du repérage automatique de relations de discours (section 7.1.1) ; nous présentons ensuite le projet ANNODIS dans son versant « micro » ou « ascendant » (section 7.1.2).

7.1.1 Structure discursive et relations de discours

Un certain nombre de travaux abordent le discours en s'intéressant plus particulièrement aux relations pouvant intervenir entre différentes parties du discours. Ces approches *bottom-up* du discours formalisent la structure du discours de façon hiérarchique et ascendante, de manière analogue à la structure syntaxique : le discours est décomposé en unités minimales, qui sont assemblées *via* des relations de discours en unités complexes, qui elles-mêmes entrent en relation avec d'autres unités de manière récursive jusqu'à obtenir une structuration de la globalité du texte.

L'idée fondamentale selon laquelle le discours serait structuré hiérarchiquement *via* des relations de discours qui lient entre eux des segments de discours a donné lieu à différents modèles, qui se distinguent par les choix théoriques faits concernant notamment la nature des unités minimales et le nombre et la nature des relations de discours.

La nature des unités minimales du discours Les approches relationnelles du discours nécessitent de segmenter entièrement les textes analysés en unités minimales. La nature de ces unités minimales varie d'une approche à l'autre. Dans le modèle de Grosz et Sidner (1986) et dans la RST (Mann et Thompson, 1987, 1988), les unités sont dites intentionnelles, c'est-à-dire qu'on peut leur associer un but communicatif de la part du locuteur/scripteur. D'autres approches telles que la

SDRT (Asher, 1993; Lascarides et Asher, 1993) définissent des unités sémantiques. D'un point de vue syntaxique, les segments minimaux retenus correspondent le plus souvent à des phrases ou des propositions.

Lorsque plusieurs segments minimaux sont regroupés *via* des relations de discours, ils forment de nouveaux segments dits segments complexes.

Le nombre et la nature des relations de discours Établir une liste finie de relations de discours est difficile. Grosz et Sidner (1986) choisissent de mettre de côté l'aspect sémantique des relations et de ne s'intéresser qu'à leurs caractéristiques structurelles. Ils ne définissent ainsi que deux relations ayant des propriétés structurelles différentes correspondant à des intentions différentes de la part du locuteur (ils parlent de théorie intentionnelle du discours). Les autres approches soit proposent une liste de relations « ouverte » (Mann et Thompson, 1988), soit considèrent qu'il est possible de définir un ensemble fermé de relations principales suffisantes pour l'interprétation des discours Hobbs (1985); Asher (1993).

Les types de structuration imposés par les différentes relations de discours Chaque nouveau segment de discours apporte une contribution qui peut être de deux ordres : continuer ce qui est en cours ou le préciser. On oppose ainsi généralement deux grands types de relations de discours qui se distinguent par leurs propriétés structurelles :

- les relations coordonnantes (ou multi-nucléaires en RST) relient deux énoncés de même importance ;
- et les relations subordonnantes (ou noyau-satellite) relient deux énoncés dont le second apporte des précisions sur le premier.

Approches relationnelles du discours et TAL La détection automatique de la structure du discours telle que modélisée par les approches relationnelles est au coeur de nombreux travaux de recherche (Moore et Wiemer-Hastings, 2003; Péry-Woodley et Scott, 2006). En effet, certains problèmes empiriques comme la résolution d'anaphores bénéficient de la prise en compte de la structure discursive. Des tentatives d'analyse relationnelle automatique de la structure du discours ont ainsi été menées (Baldridge et Lascarides, 2005; Subba et Di Eugenio, 2009). Toutefois, Moore et Wiemer-Hastings (2003) constatent que les applications telles que le résumé automatique ou les systèmes de question-réponse, qui prennent en compte des informations liées à la structure discursive, ont rompu avec les théories de la structure du discours, laissant place aux performances statistiques des systèmes informatiques. Réconcilier approches théoriques et TAL en faisant progresser l'identification automatique de relations de discours constitue donc un défi actuellement.

7.1.2 Le projet ANNODIS « ascendant »

Le projet ANNODIS dans son versant ascendant a visé la création d'un corpus annoté en relations de discours. Ce corpus est principalement constitué d'articles de presse (provenant de l'Est Républicain) et d'extraits d'articles Wikipédia. Le tableau 7.1 résume ses caractéristiques essentielles.

	Nb de textes	Nb de mots
Wikipédia	42	17330
Autres (Est Républicain)	45	27098
Corpus total	87	28146

TABLEAU 7.1 – Caractérisation du corpus ANNODIS « ascendant »

L'annotation du corpus a été effectuée en plusieurs étapes :

(i) Lors d'une première annotation dite « exploratoire », 50 textes ont été doublement annotés par deux étudiants en master de sciences du langage. Le but de cette première phase était d'assister la création et l'amélioration d'un manuel d'annotation.

(ii) Lors d'une deuxième phase d'annotation dite « naïve », 3 annotateurs (également étudiants en master de sciences du langage) ont annoté 87 textes (*cf.* tab. 7.1) en se servant du manuel d'annotation. Chacun des 87 textes a été annoté par deux annotateurs naïfs différents. Le Kappa moyen obtenu pour cette tâche est de 0.4.

(iii) Enfin, la dernière phase d'annotation est dite « experte » : 42 textes issus de l'annotation naïve ont été revus et corrigés par des annotateurs experts. Ce travail est encore en cours, puisque 45 textes n'ont pas encore été revus par les experts[1]. Lors de cette phase d'annotation, les textes n'ont pas été doublement annotés (par deux experts différents) ; il n'a donc pas été calculé d'accord inter-annotateur pour l'annotation experte.

Les deux dernières annotations (effectuées par les naïfs et les experts) font partie des livrables du projet. Ainsi, l'annotation naïve peut être exploitée, par exemple pour investiguer certains principes cognitifs de l'organisation du discours (Afantenos et Asher, 2010; Afantenos *et al.*, 2012).

À chaque niveau d'annotation (exploratoire, naïve, experte), la procédure suivie a été la suivante :

(a) **Tâche de segmentation :**
- dans un premier temps, les annotateurs effectuent chacun une segmentation du texte en unités de discours élémentaires (UDE) ;

1. Le travail d'annotation experte d'ANNODIS a été complété depuis, tous les textes sont maintenant annotés.

- puis ils se concertent afin de proposer une segmentation commune à partir de laquelle sera effectuée l'annotation en relations de discours.

(b) **Tâche d'annotation :**
- pour chaque UDE considérée, la première étape consiste à décider de son point d'attachement. L'attachement d'une nouvelle unité (le « segment attaché ») par une relation de discours se fait toujours vers un segment antérieur du texte (appelé « segment cible ») ;
- la seconde étape consiste à établir quelle relation de discours lie les deux segments.

La nature des UDE est discutée dans le manuel d'annotation, qui indique qu'« en général, une UDE correspond à la description d'un événement ou d'un état unique ». S'il est noté que « l'UDE prototypique est une proposition indépendante », d'autres cas sont également envisagés et définis « en extension ». Sont notamment considérés comme des UDE à part entière :
- les appositions, qu'elles soient adjectivales, verbales ou nominales ; cela implique qu'une UDE puisse être enchâssée dans une autre ;
- certains syntagmes adverbiaux, comme les syntagmes adverbiaux comportant un nom d'événement ou d'état ou ceux apparaissant détachés en tête de proposition.

Une « fausse » relation de discours, la relation Fusion, a été créée en vue de corriger des erreurs de segmentation lors de la phase d'attribution de relations.

Concernant les relations de discours, une liste fermée a été définie ; il s'agit d'une adaptation du modèle de la SDRT (Asher et Lascarides, 2003a), mais qui a également été inspirée d'autres modèles du discours, notamment la RST (Mann et Thompson, 1987) et le *Linguistic Discourse Model* (Polanyi, 1988) ; finalement, cette liste couvre les principales relations communes à la plupart des théories du discours. Nous reprenons ci-dessous les définitions et exemples des différentes relations, tels que fournis par le manuel d'annotation d'ANNODIS.

Explication (étiquette : `explanation`) La relation d'Explication lie deux segments dont le second (celui qui est attaché) explique le premier (la cible) de façon explicite ou non :

(1) [L'équipe a perdu lamentablement hier.]_1 [Elle avait trop de blessés.]_2
`explanation(1,2)`

But (étiquette : `goal`) La relation relie deux segments, dont l'un (le segment attaché) présente de façon explicite le but, l'objectif, pour lequel est réalisée l'action décrite dans l'autre segment (le segment cible) :

(2) [Les chercheurs ont fait grève]_1 [pour montrer leur mécontentement.]_2
`goal(1,2)`

Résultat (étiquette : `result`) La relation Résultat caractérise des liens entre deux segments portant deux éventualités (événements ou états) dont la 2e résulte de la première.

(3) [Nicholas avait bu trop de vin]_1 [et a donc dû rentrer chez lui en métro]_2.
`result(1,2)`

Parallèle (étiquette : `parallel`) Cette relation porte sur deux segments ayant une construction similaire (la plupart du temps syntaxique).

(4) [Jean aime Marie ;]_1 [il aime Claire aussi.]_2
`parallel(1,2)`

Contraste (étiquette : `contrast`) Contraste est une relation entre deux segments A et B qui est soit marquée par un marqueur explicite (comme *mais, par contre, cependant*), soit par un contraste « formel ».

(5) [Jean aime Marie,]_1 [mais il déteste Jeanne.]_2
`contrast(1,2)`

Continuation Continuation est une relation qui a une sémantique « faible ». On met Continuation entre deux segments quand le deuxième continue le rôle rhétorique du premier, par exemple une explication ou une élaboration faite dans la première.

(6) [Jean était fatigué]_1 [parce qu'il avait beaucoup travaillé]_2 [et qu'il avait peu dormi.]_3
`explanation(1,2), continuation(2,3)`

Conditionnel (étiquette : `conditional`) Cette relation relie deux constituants ou le premier est une hypothèse et le deuxième est la conséquence de l'hypothèse.

(7) [S'il pleut,]_1 [je resterai a la maison.]_2
`conditional(1,2)`

Alternation (étiquette : `alternation`) Cette relation marque une disjonction entre deux propositions.

(8) [Soit Philippe est au bureau à l'IRIT]_1 [soit il est au Mirail.]_2
`alternation(1,2)`

Attribution (étiquette : `attribution`) Attribution(a,b) est une relation qui lie un acte de parole **b** à l'agent de cet acte, qui doit être explicite dans le segment **a**.

(9) [La direction générale de Citroën a informé ses employés]_1 [que les nouveaux contrats de travail prendront effet lundi prochain.]_2
`attribution(1,2)`

Arrière-Plan (étiquette : `background`) Arrière-Plan est une relation où un constituant décrit la scène servant d'arrière-plan à l'événement décrit par l'autre constituant. Le constituant qui donne l'arrière plan a un état comme éventualité principale, ce qui, dans les textes narratifs, correspond souvent à l'usage de l'imparfait.

(10) [Marie entra dans la cuisine.]_1 [Pierre faisait la vaisselle.]_2
`background(1,2)`

Narration (étiquette : `narration`) La relation de Narration est fondée sur la maxime d'ordre de Grice : deux segments reliés par narration décrivent, dans l'ordre d'occurrence, deux éventualités (événements) d'une même histoire.

(11) [Pierre prit son parapluie.]_1 [Il sortit.]_2
`narration(1,2)`

Flashback (étiquette : `flashback`) Intuitivement, la relation de Flashback correspond à une narration dans l'autre sens : au lieu de raconter les événements dans l'ordre dans lequel ils se sont produits, on fait un retour en arrière (sur un événment qui s'est produit avant).

(12) [Paul a déménagé ce été.]_1 [Il avait trouvé un nouvel appartement au printemps.]_2
`flashback(1,2)`

Encadrement (étiquette : `frame`) La relation d'Encadrement a pour arguments un segment introducteur de cadre, *i.e.*, un adverbial détaché en tête de phrase ou en début de constituant, et le segment sur lequel porte ce cadre. Ce second argument peut être un segment complexe. En d'autres termes, un introducteur de cadre en tête de phrase peut avoir une portée au delà de cette phrase.

(13) [A la fin des années 1980,]_4 [un consensus semble se dégager en France autour de la fin de ce que l'on appelait l'« exception francaise ».]_5 [Les clivages idéologiques irréconciliables et l'atmosphère de guerre civile larvée ont disparu,]_6 [tandis que les grands conflits sociaux, [mais aussi l'engagement et la participation politique]_7 , sont en net déclin.]_8
`frame(4,[5-8])`
`continuation(5,6), continuation(6,8)`
`parralel(7,8)`
`contrast(7,8)`

Temp (relation temporelle sous-spécifiée) (étiquette : `temploc`) Il y a une relation Temp entre deux segments (1,2) quant le segment 2 permet de localiser temporellement l'événement ou l'état décrit dans le segment 1. Le segment 2 est une subordonnée temporelle ou un adverbial temporel comportant un nom d'éventualité. Temp n'est pas une relation de discours comme les autres, c'est une relation sous-spécifiée dont le rôle se limite à indiquer une localisation temporelle.

(14) [Dimanche en milieu d'après-midi.]_35 [quelques enfants s'adonnaient à des glissades]_36 [alors que l'eau, [cachée,]_37 recouvrait encore tout le secteur]_38
`temp(36,38)`
`frame(35, [36-38])`
`elaboration(38,37)`

Élaboration (étiquette : `elaboration`) La relation d'Élaboration relie deux propositions si la seconde proposition décrit un sous-état ou sous-événement de l'état ou l'événement décrit dans la premiere proposition. La relation d'Élaboration inclut également les cas d'exemplification, de reformulation et paraphrase.

(15) [Cette année-la vit de nombreux changements dans la vie de nos héros.]_1 [Jean epousa Adèle,]_2 [Marie s'acheta une maison à la campagne,]_3 [et Paul partit pour le Brésil.]_4
`elaboration(1,[2-4])`

Élaboration d'entité (étiquette : `e-elab`) La relation d'Élaboration d'entité lie deux segments dont le second (celui qui est attaché) précise une propriété d'une des entités impliquées dans le premier segment (la cible). Cette précision peut être importante (*e.g.*, identificatoire) ou marginale.

(16) [Mikhaïl Saakachvili, [le jeune et bouillant président géorgien]_1 avait besoin d'action pour sauver son régime]_2
`e-elab(2,1)`

Commentaire (étiquette : `comment`) Un constituant relié par Commentaire à un constituant A donne un point de vue d'un agent ou de l'auteur sur ce qui est décrit dans A.

(17) [La session s'est bien passée.]_1 [Jean a dit]_2 [qu'elle était très prometteuse.]_3
`comment(1,2)`
`attribution(2,3)`

Le tableau 7.2 montre la représentation de chaque relation dans le corpus expert.

Relation	total (nb)	(%)
alternation	18	0.5
attribution	75	2.2
background	155	4.6
comment	78	2.3
continuation	681	20.3
contrast	144	4.3
e-elab	527	15.7
elaboration	625	18.6
explanation	130	3.9
flashback	27	0.8
frame	211	6.3
goal	95	2.8
narration	349	10.4
parralel	59	1.8
result	163	4.9
temploc	18	0.5
total	3355	100

TABLEAU 7.2 – Représentation des différentes relations de discours dans le corpus « expert » ANNODIS « ascendant »

7.2 Explorer le rôle de la cohésion lexicale dans la structure rhétorique du discours

7.2.1 Méthodologie

Les approches rhétoriques de l'organisation discursive se sont focalisées sur des marqueurs explicites ou une modélisation sémantique pour décrire et prédire la structure du discours. La cohésion lexicale peut aussi être considérée comme un indice pour détecter la structuration hiérarchique des textes. Par exemple, des relations typiquement lexicales sont considérées comme des déclencheurs de relations rhétoriques dans Asher et Lascarides (2003b). Ceci a été la source de plusieurs investigations linguistiques qualitatives (Tanskanen, 2006; Berzlánovich *et al.*, 2008), et il y a eu des tentatives d'utilisation de mesures de proximité lexicale d'unités de discours dans des approches automatiques de l'analyse rhétorique du discours (Subba et Di Eugenio, 2009; Wellner et Pustejovsky, 2007). Notre but est d'étudier de façon plus systématique la corrélation entre la similarité lexicale et les relations de discours dans la perspective de la modélisation rhétorique d'un texte. Plus précisément, cette étude exploratoire a été construite dans le but de tester les hypothèses suivantes :

- la cohésion lexicale est un indicateur de la cohérence du discours et suit la structure rhétorique : elle est plus forte entre segments reliés rhétoriquement
- la présence de cohésion lexicale dépend du type de relation rhétorique inter-

venant entre segments de discours.

Étant donnée la faible taille des unités que nous comparons (les UDE sont typiquement des propositions), le calcul d'un score de cohésion est complexe et ne peut reposer sur le nombre de liens entre unités. Les mesures auxquelles nous faisons appel s'appuient sur les similarités entre mots (données par le score de Lin) et les agrègent pour calculer des similarités entre phrases.

Une première approche classique consiste à ne prendre en compte que la paire avec la similarité maximale entre les deux phrases. Mais, pour évaluer la similarité entre deux ensembles, il est plus judicieux de prendre en compte tous les éléments de chaque ensemble, par exemple comme cela est fait dans le calcul de la distance de Hausdorff. Hausdorff utilise la distance entre les éléments les plus différents d'un set. Cette approche n'est pas directement applicable ici, car notre mesure de similarité n'est définie que pour un sous-ensemble d'éléments lexicaux (certains mots ne sont pas connectés). Il est dans notre cas plus commode de procéder à une opération de moyennage sur les liens existants.

Une façon élégante de moyenner les liens entre deux sets de mots est proposée par Mihalcea *et al.* (2006). Soit $s(w, S)$ le maximum de similarité entre un mot w et chaque mot plein de la phrase S :

$$s(w, S) = \max_{w' \in S} sim(w, w') \tag{7.1}$$

Nous pouvons moyenner cette quantité sur tous les mots pleins d'une phrase, et obtenir ainsi une similarité non-symétrique entre deux phrases :

$$(S_1 \to S_2) = \sum_{w_1 \in S_1} \frac{sim(w_1, S_2)}{\|S_1\|} \tag{7.2}$$

La similarité entre phrases définie par Mihalcea *et al.* (2006) utilise une définition symétrique :

$$sim(S_1, S_2) = \frac{((S_1 \to S_2) + (S_2 \to S_1))}{2} \tag{7.3}$$

Des variantes ont été proposées, par exemple par Malik *et al.* (2007), ou le moyennage des similarités maximales n'est pas réalisé séparément pour chaque $(S_i \to S_j)$, mais en fin de calcul par une division par $\|S_1\| + \|S_2\|$.

Nous nous appuyons sur ces mesures pour tester nos hypothèses dans les sections qui suivent.

7.2.2 Attachement des relations et cohésion lexicale

Afin de tester les corrélations entre la cohésion lexicale entre unités du discours et leurs liens rhétoriques, nous avons constitué trois groupes de paires d'UDE :

– un ensemble de paires d'UDE venant de textes différents dans le corpus, permettant d'évaluer le « bruit de fond » de similarité dans le corpus ;

- un ensemble de paires d'UDE venant du même texte dans le corpus, mais qui ne sont pas reliées rhétoriquement d'après l'annotation ;
- un ensemble de paires d'UDE reliées par l'annotation.

Dans la mesure où nous disposons de 1927 paires d'UDE annotées dans le corpus, nous nous sommes adaptés à cette taille d'échantillon en prenant 2000 paires non-reliées venant de documents différents, et le même nombre de paires non-reliées venant du même texte.

Le tableau 7.3 présente les principaux résultats obtenus sur notre corpus : une comparaison de la cohésion lexicale moyenne pour les paire d'UDE, accompagnée de l'intervalle à 95% de confiance, de la taille de l'échantillon, et de la significativité statistique entre les moyennes (un t-test a été appliqué sur les échantillons non-reliés). Comme le montrent les résultats du t-test, les différences entre UDE non-reliées provenant de documents différents et UDE non-reliées provenant d'un même texte, ainsi qu'entre UDE reliées et non-reliées provenant des mêmes textes, sont significatives.

Position	Relation	Cohérence moyenne	±	taille échantillon	t-test par rapport à la ligne précédente
même corpus	docs diff.	40.0	2.00	2000	
même document	non-reliées	61.5	4.26	2000	$p < 10^{-17}$
même document	reliées	73.7	4.86	1927	$p < 10^{-3}$

TABLEAU 7.3 – Comparaison de la cohésion lexicale (×1000) entre UDE reliées et non-reliées, dans le même document ou dans des documents différents

Nous avons vu que la force de la cohésion lexicale entre deux segments dépend fortement de leur distance dans le texte – deux segments proches dans le texte se caractérisent par plus de liens (*cf.* section 3.3.2 page 79) et par une plus forte tendance de ces liens à être pertinents (*cf.* section 4.4.3 page 124). Ce facteur est susceptible d'interférer avec les conditions de l'expérimentation ; en effet, la grande majorité des UDE reliées dans le corpus sont contiguës (c'est-à-dire appartiennent à la même phrase ou à deux phrases adjacentes). Notre hypothèse sur l'implication de la cohésion lexicale dans les relations rhétoriques devrait conduire à une différence entre unités contiguës, selon qu'elles soient reliées ou pas. Dans le cas contraire, la cohésion lexicale apparaîtrait uniquement dépendante de l'agencement textuel. Les aspects suivants ont donc été testés :

- la cohésion entre UDE est très différente suivant qu'elles sont présentes ou pas dans la même phrase : 59 ± 5 contre 109 ± 12, ce qui pourrait indiquer que les relations intra-phrastiques n'ont pas autant recours à la cohésion lexicale et utilisent moins de réitérations lexicales.
- la différence entre UDE dans des phrases adjacentes, suivant qu'elles sont reliées ou pas, est significative ($p = 0.00023$), avec les UDE non-reliées à 63.5 ± 12 contre 96 ± 12 pour les UDE reliées.

Nous avons répété les expériences en calculant la cohésion uniquement à partir de la répétition de mots entre UDE, en utilisant des vecteurs de mots et une similarité cosinus, et n'avons pas observé d'effet statistiquement significatif. Une analyse des paires d'UDE montre que seule une petite proportion d'entre elles contiennent des répétitions lexicales, et que cet indice n'est pas pertinent sur de si petits segments. Nous avons également testé avec une pondération IDF, mais dans ce cas également les répétitions exactes semblent trop rares pour avoir une influence (sauf pour la distinction entre UDE appartenant au même document / à des documents différents).

Nous résumons les mesures de similarité lexicale que nous avons testées dans le tableau 7.4. Nous ne comparons pas les valeurs de cohésion car elles ne sont pas toutes produites de façon comparable. Nous notons cependant que les méthodes employant la similarité de Lin ont une variance bien inférieure aux autres méthodes, qui sont basées sur des indices à très faible couverture (répétitions ou synonymes).

Méthode	Statistiquement significatif à 1%
répétition seule	
répétition + IDF	
synonymes	
Lin / Mihalcea et al.	✓
Lin / Mihalcea et al. + IDF	✓
Lin / Malik et al.	✓
Lin / Malik et al. + IDF	✓

TABLEAU 7.4 – Comparaison des mesures de similarités+agrégation pour tester le caractère relié ou non-relié d'une paire d'EDU.

7.2.3 Nature des relations et cohésion lexicale

Dans un second temps, nous avons testé la corrélation du score de cohésion lexicale avec les différentes relations de discours. Le tableau 7.5 présente les relations classées par ordre décroissant de cohésion lexicale moyenne, avec intervalle à 95% de confiance. Ces données sont illustrées par la figure 7.1.

Étant donnée la taille du corpus, le nombre d'échantillons de chaque relation est souvent trop bas pour obtenir une séparation marquée entre les champs de valeurs possibles, mais il est néanmoins possible de remarquer que la cohésion est plus élevée pour les relations qui sont supposées être moins susceptibles d'être marquées. En effet des relation telles que l'élaboration et la continuation sont typiquement implicites, et des relations telles que le parallèle ou le contraste, qui sont parfois marquées explicitement, impliquent souvent des relations lexicales (des items lexicalement similaires, respectivement opposés). D'autres relations ont un comportement mixte, et une étude détaillée nécessiterait d'analyser pleinement le rôle joué par le

Relation	Cohérence moyenne	intervalle de confiance	taille échantillon
parallel	124.0	±49.6	037
elaboration	97.5	±18.4	231
continuation	83.6	±11.5	431
contrast	83.3	±23.2	78
result	80.9	±23.5	91
goal	76.5	±19.3	75
narration	71.4	±14.5	197
temploc	70.5	±34.3	18
background	67.4	±18.9	87
e-elab	60.2	±09.3	399
explanation	59.5	±16.9	70
comment	58.6	±32.3	34
conditional	56.3	±28.5	16
alternation	48.6	±42.3	15
attribution	44.2	±19.8	31
frame	43.1	±16.3	104
flashback	28.6	±19.4	13

TABLEAU 7.5 – Cohésion moyenne (×1000) entre UDE reliées, par type de relation et par valeur décroissante

facteur que nous avons relevé (voir par exemple notre analyse de parallèle dans la section d'exemple ci-dessous). Les relations souvent marquées incluent but, résultat, explication, temp, conditionnel, alternation et flashback. Pour les autres cas, nous nous serions attendu à ce que Arrière-plan présente un niveau plus élevé de cohésion lexicale. Commentaire est une relation qui peut prendre des formes diverses et est difficile à généraliser. Attribution relie un discours rapporté et son auteur, et n'est donc pas propice à la présence de nombreux liens lexicaux. Ceci nous laisse donc deux relations fréquentes : encadrement et e-élaboration. Encadrement a typiquement un premier argument très court (par exemple « Au XXème siècle »), alors que e-élaboration a généralement un second argument court, car un grand nombre d'occurrences sont des appositions ou des propositions relatives.

Ces observations incitent à tester la corrélation entre cohésion lexicale et utilisation d'un marqueur de discours. Il est peu commode de vérifier ce point, car cela nécessite une liste exhaustive des marqueurs potentiels et des relations qu'ils indiquent. Nous avons utilisé un lexique de marqueurs du discours (Roze *et al.*, 2010) [2]. Comme une grande partie de ces marqueurs sont ambigus, nous devons être

2. Ce lexique, fourni par l'équipe Alpage, est disponible sous http ://www.linguist.univ-paris-diderot.fr/~croze/D/Lexconn.xml.

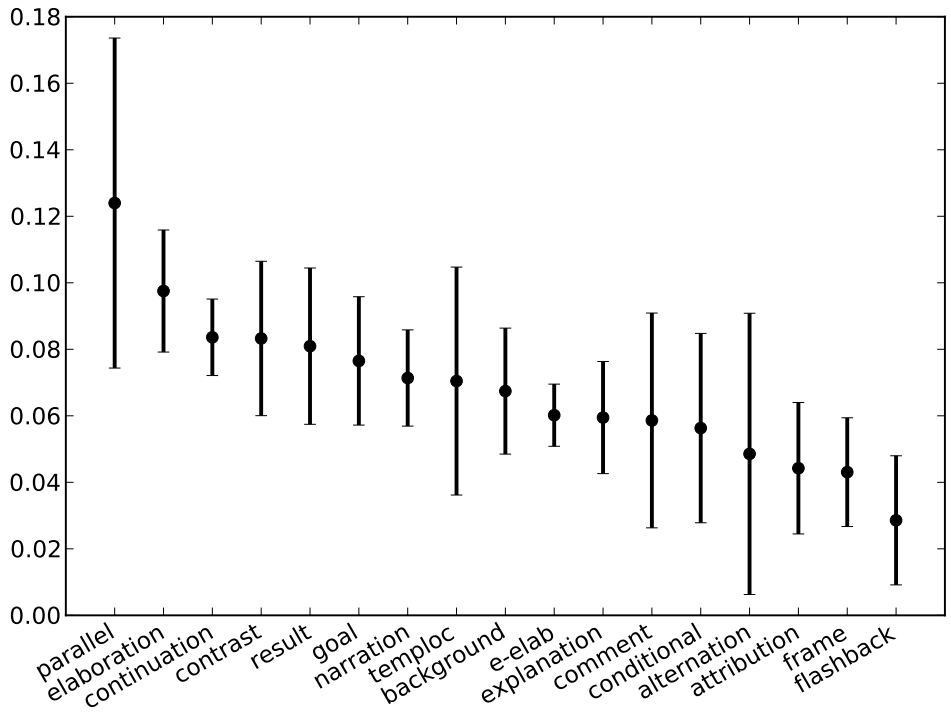


FIGURE 7.1 – Cohésion moyenne (avec intervalle de confiance) entre UDE reliées, par type de relation.

en mesure de vérifier qu'ils sont employés pour marquer la relation visée entre les UDE considérées. Nous avons choisi d'ignorer les marqueurs trop fréquemment employés dans un sens non-discursif. Nous avons ainsi été amenés à éliminer « ou », qui peut marquer une alternation, mais indique très fréquemment seulement une disjonction entre phrases nominales, ou « pour » qui peut indiquer un but mais est trop commun comme préposition dans d'autres usages. Finalement, seules 55 des 1927 paires d'UDE incluent un marqueur qui indique potentiellement la relation qui les relie. L'analyse de ces données n'a pas fait émerger de différence significative de cohésion lexicale entre UDE marquées et non-marquées.

7.2.4 Exemples et analyse d'erreurs

La figure 7.2 présente un histogramme des valeurs de similarité entre UDE reliées. Nous n'avons pas inclus la valeur pour 0, qui représente près de la moitié des cas (environ 800 instances sur 1927). On peut observer que les paires d'UDE ont des valeurs majoritairement comprises entre $0 - 0.2$. Nous pouvons donc considérer une

valeur supérieure à 0.2 comme l'indication d'une forte proximité lexicale.

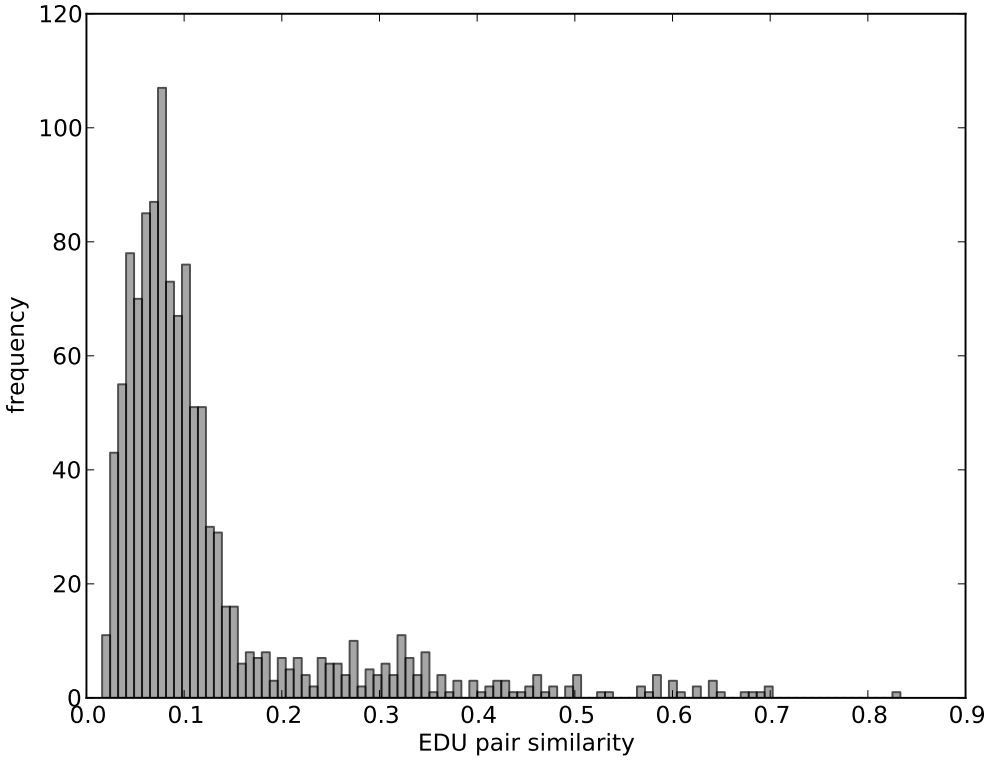


FIGURE 7.2 – Distribution des valeurs de similarité (supérieures à 0) entre paires d'UDE reliées.

Comme on pouvait s'y attendre, dans la mesure où les répétitions reçoivent la valeur maximale (1) dans notre score de similarité, les plus hauts niveaux de similarité lexicale entre propositions sont observés lorsque des mots sont répétés dans les deux propositions.

L'exemple (18) montre une relation de continuation, où la répétition de `touche` et la proximité entre `noir` et `blanc` se combinent et conduisent à une valeur de cohésion lexicale de 0.44 (on peut noter que nous manquons une autre relation lexicale pertinente dans ce contexte, la relation `ébène` / `ivoire` n'étant pas listée dans la ressource distributionnelle).

(18)
UDE 1 : les touches noires étaient recouvertes d'ébène
UDE 2 : et les touches blanches d'ivoire.

L'exemple (19) illustre une relation d'élaboration et présente de la cohésion lexicale sans répétition. La relation entre *art* et ses deux hyponymes *peinture* et *sculpture* a été détectée, parmi d'autres relations sémantiques plus lâches : *peinture-inspirer*,

peinture-règle, *peinture-établir*. La valeur de similarité est assez faible : 0.12. Cet exemple montre que des relations distributionnelles pertinentes mais très ténues peuvent tirer vers le bas le score de cohésion calculé.

(19)
UDE 1 : Par la suite, il s'inspire considérablement des règles établies par l'art égyptien,
UDE 2 : notamment en peinture et en sculpture

Dans un second temps nous avons procédé à une analyse qualitative d'un échantillon des instances relevées lors de nos tests, dans le but spécifique de rechercher de potentielles « erreurs » des types suivants :

- UDE non-reliées avec de hautes valeurs de cohésion ;
- UDE reliées avec de faibles valeurs de cohésion, alors que la relation lexicale est supposée avoir une forte cohésion lexicale.

Nous n'avons été en mesure d'examiner que des exemples extrêmes du premier cas, et avons trouvé beaucoup de très courtes UDE, souvent des titres de paragraphe qui sont nécessairement reliés à beaucoup d'autres segments dans le même texte : cela suggère la présence de relations de plus haut niveau, puisque la cohésion lexicale d'un titre avec l'ensemble de la section chapeautée est prévisible. Cela montre une certaine concurrence entre différents niveaux de structuration, mais met également en évidence un biais dans la mesure que nous avons employée dans le cas de segments de texte courts présentant une réitération lexicale.

Dans le second cas, nous nous sommes focalisés sur la relation parallèle, puisqu'elle semble être la mieux corrélée avec une cohésion lexicale relativement forte. Nous avons analysé les 37 cas trouvés dans le corpus ; 8 d'entre eux avaient une similarité de zéro, 21 avaient une similarité faible ($0.03 - 0.18$), et 8 avaient une haute similarité (> 0.25). Les cas à haute similarité impliquent toujours au moins une répétition parmi d'autres liens lexicaux, alors que les cas zéro étaient des segments courts sans liens lexicaux (mais avec un marqueur explicite d'une construction parallèle, marqueur de discours ou construction syntaxique parallèle). Parmi les 21 cas restant, nous avons observé que 10 d'entre eux étaient des cas discutables de construction parallèle, c'est-à-dire des erreurs d'annotation ou des cas difficiles, et les autres cas n'avaient pas un très haut score de similarité à cause d'une combinaison des facteurs suivants : vrais liens lexicaux qui sont absents de notre ressource (6 cas), pas de relation lexicale claire hormis les marqueurs de discours explicites (7 cas), ellipses ou anaphores (3 cas), et finalement, dans 6 cas il existait plusieurs liens lexicaux pertinents entre des UDE longues, mais le moyennage selon la longueur des UDE masquait ce fait.

Ceci est illustré par l'exemple (20) :

(20)
UDE 1 : *La principale utilité des plantes et animaux domestiques est la production alimentaire*
UDE 2 : *ainsi que celle d'autres produits utiles*

On observe ici trois paires lexicales : `produit`/`production` ($lin = 0.27$), `produit`/`animal` (0.12), `plante`/`produit` (0.15), et la moyenne prenant en compte la longueur des UDE donne 0.18. Ce cas est un cas limite, avec de bonnes et de moins bonnes relations, mais le score calculé donne dans ce cas une information ambiguë.

La leçon à tirer ici est que la similarité lexicale distributionnelle entre segments courts gagnerait probablement à mieux incorporer le (faible) nombre de liens, mais il n'existe pas de méthode évidente pour arriver à ce résultat.

Finalement, nous avons montré dans cette section la relation qui existe entre structure rhétorique du discours et cohésion lexicale : la cohésion lexicale est plus haute entre segments reliés rhétoriquement, et dépend de la nature de la relation rhétorique intervenant entre deux segments de discours. Ces résultats rendent envisageable d'exploiter la cohésion lexicale dans des systèmes visant à détecter automatiquement la structure rhétorique du discours, tant pour l'identification des relations que pour l'attachement des segments. Dans la section suivante, nous présentons un sous-système qui vise uniquement à distinguer entre deux relations fréquentes : l'élaboration et l'e-élaboration.

7.3 Un cas pratique : la distinction entre élaboration et e-élaboration

Dans cette section, nous décrivons une expérience exploitant les deux niveaux d'annotation du corpus ANNODIS (annotations naïve et experte) afin de prédire une erreur courante d'annotation : la confusion entre les relations d'élaboration [3] et d'élaboration d'entité [4]. Sur ce problème très précis, nous mettons en œuvre un système d'apprentissage automatique faisant notamment appel à la cohésion lexicale. Nous exploitons ainsi, sur une problématique bien circonscrite, les résultats présentés dans la section 7.2. Nous posons tout d'abord le problème de la confusion entre élaboration et e-élaboration dans le corpus ANNODIS, et décrivons plus précisément les deux relations (section 7.3.1). Puis nous présentons notre proposition pour traiter ce problème : l'intégration dans un classifieur automatique d'indices issus de la description linguistique des deux relations, parmi lesquels un score de cohésion lexicale. Enfin, nous discutons de la manière dont ce classifieur pourrait être exploité dans le cadre d'ANNODIS (section 7.3.3).

3. La relation d'e-élaboration fera également l'objet du chapitre 8

4. Cette expérience a été présentée lors du 6e *Linguistic Annotation Workshop* en juillet 2012 (Adam et Vergez-Couret, 2012)

7.3.1 Un problème pratique : la confusion élaboration et e-élaboration dans le corpus ANNODIS

7.3.1.1 élaboration et e-élaboration dans le projet ANNODIS

Dans le corpus ANNODIS « ascendant », élaboration et e-élaboration (pour élaboration d'entité) sont les deux relations les plus fréquentes après continuation (voire avant continuation pour l'annotation naïve). À elles seules, elles représentent 50% des relations annotées par les naïfs et 35% des relations annotées par les experts. Il s'agit également des relations les plus souvent corrigées par les experts. Enfin, comme le montre le tableau 7.6, l'élaboration est la relation la souvent confondue avec l'e-élaboration, et réciproquement :

- sur 793 élaborations selon les annotateurs naïfs, 302 cas sont réellement des élaborations, 158 cas sont des e-élaborations, et 333 cas sont concernés par une autre relation de discours ;
- sur 452 e-élaborations selon les annotateurs naïfs, 216 cas sont réellement des e-élaborations, 70 cas sont des élaborations, et 166 cas sont concernés par une autre relation de discours.

Nous parlerons désormais de confusion élaboration→e-élaboration pour désigner les cas où une élaboration (selon la référence) est annotée par un naïf comme étant une e-élaboration (70 cas dans les données présentées), et de confusion e-élaboration→élaboration pour désigner les cas où une e-élaboration est annotée comme étant une élaboration (158 cas).

		Annot. naïve	
		Elab.	E-Elab.
Annot. experte	Elab.	302	70
	E-Elab	158	216
	Fusion	81	57
	Continuation	70	32
	Arrière-Plan	32	18
	Autre	150	59

TABLEAU 7.6 – Principales confusions pour les relations d'élaboration et d'e-élaboration

7.3.1.2 Description des deux relations : parenté et altérité

La distinction entre l'élaboration d'un événement ou d'un état et l'élaboration d'une entité est couramment admise dans les théories du discours. Toutefois, cette distinction n'a pas toujours amené à définir deux relations de discours différentes (élaboration et e-élaboration). Ainsi, dans le cadre théorique de la RST (Mann et Thompson, 1987), l'élaboration d'une entité, nommée « *Object-attribute Elaboration* » est considérée comme un cas particulier de la relation d'élaboration (Knott

et al., 2001) ; cette décision est notamment motivée par l'absence de marqueur prototypique pour *Object-attribute Elaboration*. Fabricius-Hansen et Behrens (2001) introduisent deux relations distinctes, nommées « *E[ventuality] Elaboration* » et « *I[ndividual] Elaboration* ». Dans le cadre de la SDRT, Prévot *et al.* (2009) montrent la nécessité d'introduire la relation d'élaboration d'entité pour rendre compte de certains cas qui, s'ils ne sont pas isolés, brouillent la distinction entre les relations d'élaboration et d'Arrière-plan. En conséquence, dans la campagne d'annotation du projet ANNODIS, le choix a été fait de considérer séparément ces deux relations.

L'exemple (21) illustre les deux relations.

(21) [La Lausitz, [une région pauvre de l'est de l'Allemagne,]$_1$ [réputée pour ses mines de charbon à ciel ouvert,]$_2$ a été le théâtre d'une première mondiale, mardi 9 septembre.]$_3$ [Le groupe suédois Vattenfall a inauguré, dans la petite ville de Spremberg, une centrale électrique à charbon expérimentale]$_4$ [qui met en œuvre toute la chaîne des techniques de captage et de stockage du carbone]$_5$

L'annotation experte pour cet exemple est la suivante :

Continuation (1,2)
E-Elaboration (3,[1-2])
Elaboration (3,4)
E-Elaboration (4,5)

Le segment complexe [1-2] est enchâssé dans le segment 3, et donne des informations sur une entité introduite dans ce segment 3, « la Lausitz » ; il est donc connecté à ce segment par la relation d'e-élaboration. Le segment 4 décrit un événement (« inaugurer une centrale électrique à charbon expérimentale ») qui précise l'événement « être le théâtre d'une première mondiale » ; il est donc attaché au segment 3 par la relation d'élaboration. Enfin, le segment 5 indique une propriété de l'entité « centrale électrique à charbon expérimentale » ; il est ainsi relié au segment 4 par la relation d'e-élaboration.

Le manuel d'annotation rédigé suite à la phase exploratoire de la campagne d'annotation ANNODIS discute de la possible confusion entre élaboration et e-élaboration. Les solutions qu'il propose font principalement appel aux notions sémantiques d'événement, d'état et d'entité : il rappelle à l'annotateur que la distinction majeure entre élaboration et e-élaboration est que élaboration donne des détails sur l'événement ou l'état décrit par le segment élaboré, tandis que e-élaboration donne des détails sur une entité particulière de ce segment. On a vu en 8.2.1 que malgré ces précautions, les résultats de l'annotation naïve révèlent que la distinction entre les deux relations demeure problématique. Dans la section suivante, nous discutons des éléments linguistiques de surface qui peuvent être utilisés pour différencier les deux relations.

7.3.1.3 Éléments linguistiques pour une différenciation

La complexité de l'identification (humaine ou automatique) d'élaboration et e-élaboration est renforcée par le fait qu'il n'existe pas de marqueur prototypique pour ces relations (Knott, 1996). Ce constat est repris dans le manuel d'annotation d'ANNODIS, qui précise que « le plus souvent, élaboration apparaît sans marqueur », et que l'e-élaboration n'a « pas de marqueurs explicites ». Le manuel d'annotation indique toutefois certains marqueurs tels que « c'est-à-dire » et « à savoir », mais ces marqueurs rares ne sont pas discriminants entre ces deux relations (ils marquent aussi bien élaboration que e-élaboration). Il précise également que pour l'e-élaboration, le segment attaché est « souvent une apposition (syntagme nominal ou adjectival), ou bien une relative explicative (non déterminative) ». De telles réalisations ont été mises en évidence par Prévot *et al.* (2009). Par ailleurs Vergez-Couret et Adam (2012) notent que le syntagme gérondif peut exprimer plusieurs relations de discours, incluant l'élaboration mais pas l'e-élaboration.

Enfin, nous avons montré en 7.2.3 que la relation d'élaboration présente une plus forte cohésion lexicale que la relation d'e-élaboration, avec un score moyen de 97.5 contre 60.2. Cette différence peut être vue comme le reflet de la distinction sémantique entre ces deux relations de discours. Notamment, nous pensons que l'élaboration implique une plus forte cohésion lexicale car elle relie deux propositions, alors que l'e-élaboration relie une proposition à une entité (elle n'élabore pas toute la proposition mais seulement une entité particulière). Cette distinction est illustré par les exemples (22) et (23)

(22) [Un soir, il faisait un temps horrible,]$_{16}$ [les éclairs se croisaient,]$_{17}$ [le tonnerre grondait,]$_{18}$ [la pluie tombait à torrent.]$_{19}$

(23) [Pourquoi a-t-on abattu Paul Mariani, [cinquante-cinq ans]$_4$, [attaché au cabinet de M. François Doubin]$_5$?]$_6$

Dans l'exemple (22), les segments [17-19] élaborent l'état décrit par le segment 16 : « faire un temps horrible » ; il s'agit donc d'une élaboration. Des liens de cohésion lexicale connectent effectivement les deux segments : les liens entre « temps » et « éclair », « tonnerre », « pluie ». Dans l'exemple (23), les segments enchâssés [4-5] élaborent une entité du segment englobant 6, Paul Mariani (et non pas l'événement « abattre Paul Mariani »). Aucun lien de cohésion lexicale ne peut être identifié entre ces segments, qui décrivent l'âge et la profession de l'entité « Paul Mariani », et le segment 6, qui mentionne que ce dernier a été abattu.

7.3.2 Notre proposition : intégrer la cohésion lexicale à un système d'apprentissage automatique

Dans cette section, nous proposons d'intégrer les différents indices permettant de distinguer entre élaboration et e-élaboration (dont la cohésion lexicale) dans un système de classification automatique, afin de prédire les confusions entre ces deux relations chez les annotateurs naïfs.

7.3.2.1 Méthodologie

Pour cette expérience, nous utilisons les 42 textes du corpus ANNODIS ayant été corrigés par les experts, et disposant donc de deux niveaux d'annotation : l'annotation dite « naïve » et l'annotation dite « experte ». Dans ces 42 textes, nous avons sélectionné tous les segments annotés élaboration ou e-élaboration dans l'annotation naïve et dans l'annotation experte. Parmi l'ensemble des cas d' élaboration / e-élaboration selon les annotateurs naïfs (1245 cas), nous restreignons donc nos données à ceux qui sont effectivement des élaboration / e-élaboration selon la référence établie par les experts (746 cas). En effet, nous n'avons défini d'indices que pour ces deux relations ; pour intégrer plus de relations à notre système, il faudrait implémenter de nouveaux indices dédiés à ces relations.

Chaque cas est constitué de deux segments (éventuellement complexes), et peut donc être exprimé sous la forme $< S_a, S_b >$. Pour chaque paire de segments, nous avons calculé les indices suivants :

- un score de cohésion (7.2.1)
- des indices syntaxiques, implémentés à l'aide de patrons : S_b est-il une proposition relative, une apposition nominale ou adjectivale (est-il entre parenthèses ?), ou encore un segment gérondif ?
- des indices structuraux : longueur des segments en nombre de mots et en nombre de sous-segments (1 pour un segment non complexe), positions relatives des deux segments (S_b peut être enchâssé dans S_a ou le suivre).

Ces indices sont résumés dans le tableau 7.7.

Nous avons ensuite traité les données obtenues en utilisant l'outil dédié à l'apprentissage automatique Weka. Plus précisément, nous avons utilisé l'implémentation Weka du modèle Random Forest, précédemment décrit dans la section 4.5.1 (page 131). Pour l'évaluation des résultats, présentée dans la prochaine section, nous avons procédé à une validation croisée à 10 échantillons (ce procédé a également été décrit dans la section 4.5.1).

7.3.2.2 Résultats

Le tableau 7.8 rappelle le bilan de l'annotation naïve lorsque l'on limite le corpus ANNODIS aux relations d'élaboration et d'e-élaboration. En guise de première comparaison, nous donnons dans le tableau 7.9 les résultats obtenus par notre système sans utiliser l'annotation naïve. La confusion e-élaboration→élaboration est

Indices	Description	Valeurs
Sc	*cf.* section 7.2.1	$Sc \in \mathbb{R}^+$
rel	S_b est une proposition relative	booléen
app	S_b est une apposition nominale/adjectivale	booléen
ger	S_b est un syntagme gérondif	booléen
par	S_b apparaît entre parenthèses	booléen
emb	S_b est enchâssé dans S_a	booléen
w_{Sa}	nombre de mots de S_a	$w_{S1} \in \mathbb{N}^*$
w_{Sb}	nombre de mots de S_b	$w_{S2} \in \mathbb{N}^*$
w_{tot}	$w_{Sa} + w_{Sb}$	$w_{tot} \in \mathbb{N}$
s_{Sa}	nombre de segments de S_a	$s_{S1} \in \mathbb{N}^*$
s_{Sb}	nombre de segments de S_b	$s_{S2} \in \mathbb{N}^*$
s_{tot}	$s_{Sa} + s_{Sb}$	$s_{tot} \in \mathbb{N}$

TABLEAU 7.7 – Indices implémentés

	elab	e-elab	← Naïfs
elab	302	70	
e-elab	158	216	

↑ Experts

Exactitude : 69.4%

TABLEAU 7.8 – Matrice de confusion pour l'annotation naïve

sensiblement réduite, tandis que la confusion e-élaboration→élaboration augmente légèrement, et l'exactitude passe à 73.4% Il est néanmoins important de noter que cette situation expérimentale ne correspond à aucun cas pratique, puisqu'on isole ici attachement des segments et annotation en relations de discours, tâches qui sont effectuées de manière indissociable par les annotateurs. En effet, même si notre système ne tient pas compte ici des relations de discours assignées par les annotateurs, il en bénéficie car l'ensemble des indices qu'il utilise sont calculés à partir de l'attachement choisi par ces mêmes annotateurs. Il faut donc se garder de confondre ces résultats avec ceux que l'on obtiendrait en remplaçant totalement l'annotation naïve : pour remplir un tel rôle, il faudrait que notre système traite également le problème de l'attachement des segments.

	elab	e-elab	← Auto.
elab	289	83	
e-elab	115	259	

↑ Experts

Exactitude : 73.4%

TABLEAU 7.9 – Matrice de confusion pour la classification automatique

Les résultats les plus intéressants sont ceux obtenus en prenant en entrée l'annotation naïve. Pour les produire, nous avons ajouté les relations de discours assignées par les annotateurs naïfs comme un indice supplémentaire utilisé par notre système. La matrice de confusion 7.10 résume ces résultats. On peut constater que notre système permet d'améliorer fortement l'annotation qu'il prend en entrée. La confusion e-élaboration→élaboration est réduite de 27%, passant de 158 occurrences incorrectement annotées à 115 ; la confusion élaboration→e-élaboration est réduite de 6% (de 70 à 66 erreurs). Le nombre de cas correctement annotés passe ainsi de 518 à 565 (+9%). Ces améliorations se reflètent sur le taux d'exactitude, qui est ici très bon : 75.7%.

	elab	e-elab	← Naïfs+auto.
elab	306	66	
e-elab	115	259	

↑ Experts

Exactitude : 75.7%

TABLEAU 7.10 – Matrice de confusion

Afin d'évaluer l'impact des différents types d'indices utilisés, nous avons testé notre système, toujours en prenant comme entrée l'annotation naïve, et en utilisant à chaque fois des jeux d'indices restreints : uniquement les indices de cohésion lexicale, puis les indices syntaxiques, puis les indices structurels. Nous présentons les résultats obtenus dans le tableau 7.11, en précisant pour chaque catégorie son apport à la *baseline* que représente l'annotation naïve.

Indices utilisés	Exactitude
Annotation naïve	69.4%
Score de cohésion lexicale	72.3% (+2.9%)
Indices syntaxiques	71.7% (+2.3%)
Indices structurels	69.7% (+0.3%)
All	75.7% (+6.3%)

TABLEAU 7.11 – Impact des différentes catégories d'indices

Le faible apport des indices structurels (+0.3%) peut s'expliquer par le fait qu'ils découlent directement de l'attachement effectué et donc de la relation assignée par l'annotateur. C'est pourquoi, malgré le fait qu'ils permettent de mettre au jour des différences marquées entre élaboration et e-élaboration, ils n'améliorent pas très sensiblement l'annotation prise en entrée. Les indices linguistiques (lexicaux ou syntaxiques) amènent une augmentation des performances bien plus importante (respectivement +2.9% et +2.3%). On peut néanmoins remarquer que les quatre indices linguistiques (*rel*, *app*, *ger* et *par*) associés ne font pas aussi bien que l'indice basé sur la cohésion lexicale. Cela peut s'expliquer par différentes observations :

- leur couverture est moindre ;
- dans la mesure où ces indices sont mentionnés dans le manuel d'annotation, on peut supposer qu'ils sont mieux pris en compte par les annotateurs (et donc par la *baseline*).

La cohésion lexicale telle qu'exprimée par notre indice basé sur le voisinage distributionnel est finalement le meilleur indicateur pour distinguer élaboration et e-élaboration. On peut enfin souligner la complémentarité des différentes catégories d'indices, mise en évidence par l'apport de performance du système lorsqu'il utilise tous les indices, supérieur à la somme des apports de chaque catégorie d'indices prise isolément.

7.3.3 Comment intégrer notre approche à une campagne d'annotation ?

Nous discutons dans cette section du rôle que pourrait revêtir une approche telle que la nôtre dans le cadre d'une campagne d'annotation comme ANNODIS. Dans le cadre du projet ANNODIS, une stratégie d'annotation itérative, en plusieurs passes, a été adoptée, afin de construire un corpus annoté de la manière la plus fiable possible. Dans une telle stratégie, une approche automatique telle que celle que nous avons présentée pourrait endosser différents rôles, selon le moment où l'on choisit de la faire intervenir :

(a) en amont de toute annotation, pour fournir une première pré-annotation : il s'agirait alors de remplacer les annotateurs naïfs en effectuant une première passe, destinée à être revue par les experts ;

(b) en aval de l'annotation, pour l'améliorer : il s'agirait ici de se substituer aux experts en corrigeant certaines erreurs d'annotation ;

(c) comme une phase intermédiaire, en ne se substituant ni aux annotateurs naïfs ni aux annotateurs experts, mais en ajoutant une nouvelle passe qui s'intercalerait entre l'annotation naïve et l'annotation experte.

Le système que nous avons proposé ne pourrait pas être utilisé, à l'heure actuelle, pour remplir le rôle (a). En effet, comme nous l'avons souligné, il ne peut se passer de l'annotation naïve puisqu'il ne traite pas la question de l'attachement. De plus, il ne reconnait qu'un ensemble très restreint de deux relations de discours. Toutefois, le développement pas à pas d'un système intégrant de plus en plus de relations jusqu'à pouvoir fournir une pré-annotation complète est une perspective à long terme de notre approche. Par ailleurs, les résultats obtenus et l'exploration des indices, de leur valeur informationnelle et de leur prise en compte effective par les annotateurs pourraient permettre d'enrichir le manuel d'annotation, ce qui constitue un autre type d'intervention en amont de l'annotation.

Par contre, notre système peut directement être intégré à la campagne d'annotation ANNODIS pour remplir les rôles (b) et (c), qui supposent une première annotation humaine. Cela répond à un besoin potentiel, puisqu'à l'heure actuelle, seuls 86 textes parmi les 130 textes du corpus ANNODIS « micro » ont bénéficié d'une double annotation. Les 44 textes restants ne bénéficient que de l'annotation naïve.

Notre système peut directement être exploité pour améliorer l'annotation naive (b) pour ces textes n'ayant pas été revus pas les experts, puisqu'il permet de corriger de manière entièrement automatisée un nombre important de cas, comme nous l'avons vu dans les résultats présentés en section 8.3.4 (de 69.4% à 75.7% d'exactitude pour la distinction entre les deux relations).

Enfin, nous discutons dans la suite de cette section des modalités selon lesquelles notre système pourrait être exploité pour assister l'annotation par les experts, en réduisant la charge de travail de ces derniers (c). Cette problématique est légèrement différente de celle de la situation (b) : il s'agit ici de valider automatiquement une partie des annotations naïves, jugées fiables, afin de ne présenter à l'expert qu'un sous-ensemble de cas jugés douteux. Cette différence de problématique entraîne quelques changements méthodologiques. Plutôt que d'utiliser l'annotation naïve comme un indice parmi d'autres dans le système d'apprentissage, on cherche cette fois, pour un sous-ensemble de cas ayant reçu la même annotation, à mettre de côté les plus fiables de ces cas, afin de ne pas les resoumettre à l'annotateur expert.

Prenons l'exemple du sous-ensemble des e-élaborations selon l'annotation naïve. Le tableau 7.8 nous a permis de constater que les annotateurs naïfs identifient les e-élaborations avec une assez bonne précision, puisque sur les 286 cas que contiennent nos données, 216 sont bien des e-élaborations selon l'annotation experte (on a donc un taux d'erreur d'environ 25%). Il paraît donc inutile que l'intégralité de ces cas soient revus par les experts, d'où l'utilité de faire intervenir une phase de classification automatique. Les résultats de cette classification automatique prendront la

forme présentée par le tableau 7.12, avec $a + b + c + d = 286$ (le nombre de cas d'e-élaboration selon les annotateurs naïfs).

	elab	e-elab	← Annot. auto.
elab	a	b	(annot. naïve=e-elab)
e-elab	c	d	
↑ Exp.	$a + c$	$b + d$	
	↙	↘	
	cas soumis aux experts	cas validés (taux d'erreur accepté : $b/(b + d) < 10\%$)	

TABLEAU 7.12 – Format de la matrice de confusion pour l'annotation experte assistée

La classification automatique a ici pour objectif de :
- réduire le temps de travail des experts, c'est-à-dire réduire le nombre $a + c$;
- tout en maintenant un nombre de faux positifs (nombre b) le plus bas possible, car ces derniers ne seront pas revus pas les experts ; viser un taux d'erreur inférieur à 10% paraît raisonnable étant donnée la difficulté de la tâche ; la quantité de faux négatifs (nombre c) a par contre moins d'importance, puisqu'ils se situent dans les cas soumis aux experts.

En d'autres termes, il s'agit ici de favoriser la précision sur le rappel dans la tâche de validation automatique des e-élaborations.

Concrètement, cela peut être mis en oeuvre en faisant appel à une matrice de coûts (Witten *et al.*, 2011, pp. 166-168), c'est-à-dire en indiquant au classifieur automatique que certaines erreurs « coûtent » plus que d'autres. La matrice que nous avons utilisée est donnée dans le tableau 7.13. Le coût associé aux cases a et d est de 0 (ce sont les « bonnes réponses ») ; les deux types d'erreurs se voient assigner des coûts différents : 1 pour les faux négatifs et 10 pour les faux positifs ; la confusion élaboration→e-élaboration est donc 10 fois plus coûteuse que la confusion e-élaboration→élaboration. La valeur de la pondération (10) a été fixée empiriquement de manière a maintenir le taux d'erreur en dessous de 10%, en accord avec les contraintes énoncées précédemment.

	elab	e-elab
elab	0	10
e-elab	1	0

TABLEAU 7.13 – Matrice de coûts utilisée

Le tableau 7.14 montre les résultats obtenus par l'application de cette stratégie. Parmi les 286 e-élaborations de l'annotation naïve, 159 sont validées automatiquement ; le taux d'erreur est seulement de 8.2%. La réduction de la charge de travail pour les experts est de $159/286 = 55.6\%$.

	elab	e-elab	← annot. auto.
elab	57	13	(annot. naïve=e-elab)
e-elab	70	146	
↑Expert	127	159	
	↙	↘	
	cas soumis aux experts	cas validés (erreur : 8.2%)	

TABLEAU 7.14 – Matrice de confusion pour l'annotation experte assistée

Ces résultats montrent que nos travaux sur la corrélation entre cohésion lexicale et structure rhétorique peuvent être mis à profit dans des approches automatiques de la structure du discours.

7.4 Bilan et perspectives

Ce chapitre s'est intéressé à la relation entre cohésion lexicale et structure hiérarchique du discours, en s'appuyant sur les données annotées lors du projet ANNODIS « ascendant ».

Dans un premier temps, nous avons montré que la cohésion lexicale accompagne la structure hiérarchique du discours. Deux UDE présentent une cohésion lexicale significativement supérieure si elles sont directement attachées par une relation de discours, sans que cela ne puisse être attribué à un effet de la distance séparant ces deux UDE. D'autre part, la cohésion lexicale est significativement corrélée à la nature des relations de discours, certaines relations présentant une plus forte cohésion que d'autres ; nous avons donné des pistes d'interprétation à ces résultats, notamment le fait qu'on trouve parmi les relations les plus cohésives des relations connues pour être non-marquées. Un éventuel système de calcul automatique des attachements et de classification automatique en relations rhétoriques devrait prendre en compte la cohésion lexicale. Nous n'avons pas pour objectif de concevoir un tel système, qui devrait également intégrer toute une batterie d'indices et de marqueurs concernant toutes les relations de discours, la cohésion lexicale ne pouvant constituer qu'un indice parmi d'autres. De plus, certaines relations sont très peu représentées dans le corpus ANNODIS, de taille assez réduite, ce qui rend plus difficile d'envisager un apprentissage automatique performant sur ce corpus.

C'est pourquoi, dans un second temps, nous nous sommes intéressées à une problématique plus restreinte : la distinction entre deux relations de discours, élaboration et e-élaboration. Ces deux relations sont très fréquentes dans le corpus ANNODIS, ce qui nous a permis de disposer de données suffisante pour l'expérience proposée. De plus, il s'agit de deux relations très souvent confondues par les annotateurs ; leur distinction automatique présente donc un enjeu réel. Nous avons montré comment la cohésion lexicale peut participer à un système d'apprentissage automatique

chargé de prédire des erreurs d'annotation entre les relations de discours élaboration et e-élaboration, et comment un tel système peut être exploité dans le cadre d'une campagne d'annotation telle que celle d'ANNODIS.

Cette expérience tire profit de différentes couches d'annotation du corpus, ce qui est rare. Les perspectives sont nombreuses.

- Sur la base des textes du corpus ANNODIS qui ne bénéficient que d'une annotation naïve, il faudrait évaluer si le système que nous avons proposé permet un réel gain de temps lors de l'annotation experte (en d'autres termes, vérifier si la réduction de charge de travail se traduit effectivement par une réduction de temps de travail).
- Le succès de l'approche que nous avons mise en œuvre incite à poursuivre le travail entamé en intégrant pas à pas d'autres relations.

Finalement, la cohésion lexicale apparaît comme un indice pour l'identification de la relation d'élaboration, fonctionnant ici par contraste avec une relation impliquant une cohésion lexicale plus faible, la relation d'e-élaboration. Ce résultat est d'autant plus intéressant que l'élaboration est une relation apparaissant généralement non marquée, et dont la détection automatique constitue donc un défi. Dans le prochain chapitre, nous nous intéressons plus précisément à cette relation d'élaboration et à la façon dont elle peut être signalée par la cohésion lexicale.

Chapitre 8

L'Élaboration, une relation de discours signalée par la cohésion lexicale ?

Sommaire

Dans ce chapitre, nous nous concentrons sur la relation d'élaboration, qui a déjà fait l'objet d'une expérience dans le chapitre 7. Nous savons suite aux différents résultats présentés dans ce ce dernier chapitre que la relation d'élaboration est caractérisée par une forte cohésion lexicale. Cela nous incite à l'étudier plus précisément. Notre objectif est d'observer en détail comment se manifestent les liens de cohésion lexicale impliqués dans la relation d'élaboration, afin d'aboutir à une détection de certaines de ses réalisations en corpus, en nous affranchissant ainsi des données annotées lors de la campagne ANNODIS. Une telle détection non-supervisée est une problématique bien plus difficile que celle de la distinction entre deux relations (élaboration et e-élaboration) présentant des fonctionnements divergeant de manière marquée. Nous adoptons ainsi ici une démarche très ciblée, qui vient compléter le panorama d'approches du discours de cette troisième partie, et montre encore une fois que les indices lexicaux mis au jour par les voisins distributionnels permettent d'aborder une large gamme de phénomènes discursifs.

8.1 Contexte

Dans cette section, nous décrivons plus précisément la relation d'élaboration, relation particulièrement difficile à détecter et dont la description dans la SDRT fait appel à la cohésion lexicale (section 8.1.1) ; nous présentons ensuite le contexte scientifique dans lequel nous avons effectué les travaux présentés dans ce chapitre (section 8.1.2).

8.1.1 La relation d'élaboration

8.1.1.1 Définition

La relation d'élaboration est une relation subordonnante établie entre deux segments si et seulement si le second segment apporte des détails supplémentaires sur l'éventualité (état ou événement) décrite dans le premier segment. L'extrait du Petit Prince ci-dessous (1) est un exemple d'élaboration (tiré de Vergez-Couret, 2010, p. xv).

(1) Au matin du départ il mit sa planète bien en ordre. Il ramona soigneusement ses volcans en activité. (...) Il ramona donc également le volcan éteint. (...) Le petit prince arracha aussi, avec un peu de mélancolie, les dernières pousses de baobabs.

Dans cet exemple, la première phrase décrit un événement : « mettre sa planète bien en ordre ». Les phrases suivantes élaborent cette première phrase, car elles apportent des détails en décrivant des événements qui participent à la mise en ordre de la planète (« ramoner ses volcans en activité », « ramoner le volcan éteint » et « arracher

les dernières pousses de baobabs »). On parlera désormais de segment élaboré (pour le premier segment) et de segment(s) élaborant(s) (pour le(s) segment(s) fournissant des détails supplémentaires sur ce premier segment).

On peut remarquer qu'aucun marqueur discursif ne permet dans l'exemple (1) d'appuyer l'inférence de la relation d'élaboration. La présence d'un marqueur faciliterait l'identification automatique ; mais comme nous allons le voir, ce type de marqueur fait cruellement défaut pour la relation d'élaboration.

8.1.1.2 La relation d'élaboration : une relation particulièrement difficile à détecter

Scott et Souza (1990); Knott (1996); Knott *et al.* (2001) ont observé que la relation d'élaboration s'accompagne rarement de signaux de surface ; Knott (1996) qualifie cette relation de « non marquée ». Des travaux récents ont malgré tout identifié des marqueurs ou des indices syntaxiques et typo-dispositionnels jouant un rôle plus ou moins fort dans l'interprétation de la relation d'élaboration. Il s'agit notamment de :

- certains adverbes paradigmatisants : « notamment » (Vergez-Couret, 2009), « particulièrement », « en particulier », « précisément », etc. ;
- certaines structures syntaxiques : Fabricius-Hansen et Behrens (2001) ont étudié la conjonction de subordination allemande « *indem* », traduite par « *ved å* » en norvégien et par by + V-ing en anglais ; en français, la structure équivalente est le gérondif (Vergez-Couret et Adam, 2012) ;
- certaines structures textuelles telles que les structures énumératives (Bras *et al.*, 2008).

Mais ces indices restent rares et généralement ambigus, comme le montre par exemple Bras (2007) pour le marqueur discursif « d'abord » (« d'abord » peut marquer la première étape d'élaboration mais aussi celle d'explication). Une identification automatique en utilisant des marqueurs du discours est donc difficile, et les stratégies mises en oeuvre pour le repérage de cette relation restent très basiques. Dans les travaux de Marcu (2000), la relation d'élaboration est inférée sur des critères non-linguistiques (basés sur la taille des paragraphes et des sections) et *à défaut* de pouvoir attribuer une autre relation de discours : dans un paragraphe court ne contenant pas de marqueur de discours, la relation établie entre les phrases de ce paragraphe sera la relation d'élaboration.

8.1.1.3 La relation d'élaboration en SDRT : une relation signalée par la cohésion lexicale ?

La SDRT (*Segmented Discourse Representation Theory*) (Asher, 1993; Asher et Lascarides, 2003a) est une théorie représentationnelle dynamique du discours, qui cherche à rendre compte des liens entre contenu sémantique des propositions et structure globale du discours. C'est une théorie formelle qui s'efforce de décrire dans

un cadre logique les mécanismes mis en oeuvre du point de vue du lecteur pour inférer une relation entre deux segments de discours. Ces mécanismes sont traduits au moyen de règles d'inférence qui font intervenir différents types d'informations : marqueurs du discours, sémantique lexicale et grammaticale, connaissances du mondes, principes pragmatiques.

Dans le cadre théorique de la SDRT, l'une des règles d'inférence de la relation d'élaboration est basée sur de l'information au niveau lexical. Selon la SDRT, l'élaboration est inférée de manière non monotone (c'est-à-dire défaisable) s'il existe une relation de sous-type discursif entre les éventualités décrites. C'est le cas dans l'exemple (2).

(2) [Max ate a lovely meal.]a [He ate up salmon.]b [He devoured lots of cheese.]c [And then he savoured a chocolate cake.]d (Asher et Lascarides, 2003a, p. 282)

Ici, les événements décrits dans les segments b, c et d (« to eat up salmon », etc.) sont en relation de sous-type discursif avec l'événement décrit en a (« to eat a lovely meal »). On parle d'inférence défaisable car elle peut être annulée si une autre inférence monotone est établie ; ainsi, dans l'exemple (3), la présence du marqueur « then » contraint l'interprétation de la relation discursive de narration.

(3) [Max ate a lovely meal.]a [And then he ate up salmon.]b

Toujours dans le cadre de la SDRT, le lexique est considéré comme une source d'information importante (mais pas exclusive) pour inférer le sous-type discursif. En effet, celui-ci est souvent reflété, au niveau lexical, par la relation de sous-type lexical, équivalente à la relation d'hyponymie. Ainsi, dans l'exemple (2), les relations de sous-types discursifs précédemment mentionnées sont reflétées par les relations lexicales « to eat up » / « to devour » / « to savour » $>$ « to eat » et « salmon » / « cheese » / « chocolate » $>$ « meal ». On peut toutefois dès à présent remarquer que les relations lexicales impliquées sont complexes : on peut difficilement considérer « salmon » et « chocolate » comme des sous-types de « meal ». Il s'agit plutôt de sous-types de « food », lui-même sémantiquement associé à « meal ».

La SDRT fait ainsi l'hypothèse que la relation d'élaboration pourrait être marquée par des éléments de cohésion lexicale, et plus précisément par la présence de la relation lexicale d'hypéronymie/hyponymie.

8.1.2 Contexte du travail réalisé

Dans les sections qui suivent, nous présentons des travaux menés dans le but d'explorer l'hypothèse d'un marquage lexical de la relation d'élaboration. Nous y relatons le cheminement suivi, partant d'une phase d'exploration qualitative des aspects lexicaux impliqués dans la relation d'élaboration (8.2) et aboutissant à une phase expérimentale évaluant quantitativement l'apport des voisins dans un cas

particulier de détection automatique de la relation d'élaboration.

Ces travaux poursuivent les travaux menés en relation avec le projet ANNODIS « ascendant », avec un élargissement vers d'autres données (l'idée est de reporter les observations faites sur les données annotées d'ANNODIS à des données nouvelles, toujours extraites de Wikipédia). Ils ont donné lieu à une présentation au colloque MAD'10 (Multidisciplinary approaches to discourse, Adam et Vergez-Couret, 2010). Ils ont été menés en étroite collaboration avec Marianne Vergez-Couret, membre du projet ANNODIS et spécialiste de la SDRT et de la relation d'élaboration[1]. Nous avons ainsi pu bénéficier de son expertise en sémantique du discours.

8.2 Relation d'élaboration et cohésion lexicale : exploration qualitative

Dans un premier temps, nous avons choisi de mener une observation qualitative des phénomènes lexicaux mis en œuvre dans la relation d'élaboration. Nous nous sommes pour cela basées sur les données exploratoires du corpus ANNODIS « ascendant », que nous caractérisons dans la section 8.2.1. Nous nous sommes interrogées sur les aspects lexicaux impliqués dans la relation d'élaboration, en relation d'une part avec le *Subtype* de la SDRT (8.2.2), et d'autre part avec le voisinage distributionnel (8.2.3). Il s'agit donc d'une approche onomasiologique : on part de la relation de discours pour découvrir comment elle est signalée dans les textes, en portant ici une attention toute particulière à ses aspects lexicaux.

8.2.1 Données utilisées

L'étude que nous décrivons ici a été réalisée dans le courant de l'année 2009, alors que le projet ANNODIS était en cours. Nous n'avons donc pas bénéficié de la version finale du corpus développé lors de ce projet, mais de la version courante du corpus dit « exploratoire » (*cf.* section 7.1.2), constitué et annoté en parallèle de la réalisation d'un manuel d'annotation destiné aux annotateurs du corpus final. Les caractéristiques principales de la version du corpus exploratoire sur lequel nous nous sommes appuyées sont résumées dans le tableau 8.1.

Comme nous l'avons vu dans la section 7.1.2 page 210, la relation d'élaboration est l'une des plus attribuées ; elle se caractérise par un accord inter-annotateur très faible. De plus, même en se limitant aux cas d'accord entre les annotateurs, l'analyse effectuée par Vergez-Couret (2010) ne concorde que dans 32 cas sur 84 avec la double annotation élaboration. Cela montre à quel point la relation d'élaboration difficile à repérer, même pour des annotateurs formés.

1. M. Vergez-Couret a soutenu en 2010 une thèse intitulée : « Étude en corpus des réalisations linguistiques de la relation d'Élaboration ».

Nombre de textes	47
Nombre de textes doublement annotés	18
Nombre de mots	13465
Nombre d'UDEs	1541
Nombre d'élaborations [2]	230

TABLEAU 8.1 – Caractérisation du corpus utilisé

Comme mentionné ci-dessus, le corpus sur lequel nous avons travaillé pour ces analyses a été annoté parallèlement à la rédaction du manuel d'annotation (afin d'aider cette rédaction), et se caractérise logiquement par un accord inter-annotateur plus faible que le corpus final, et conséquemment par des annotations moins fiables. C'est pourquoi, afin de travailler sur des données les plus fiables possibles, et dans la mesure où leur quantité importait peu pour cette première exploration essentiellement qualitative, nous avons choisi de :

- extraire dans ce corpus toutes les élaborations bénéficiant d'un accord entre les deux annotateurs,
- puis faire une post-sélection d'un sous-ensemble restreint d'élaborations fiables parmi l'ensemble des relations extraites.

Au final, nous avons travaillé sur un petit corpus d'une quarantaine d'élaborations. Dans un premier temps, nous avons tenté d'identifier les relations lexicales pertinentes dans l'interprétation de ces élaborations (8.2.2) ; dans un second temps, nous avons projeté les voisins sur ces données afin de mettre en rapport nos analyses avec les liens lexicaux ainsi mis au jour (8.2.3).

8.2.2 Relations lexicales impliquées

Rappelons que selon la SDRT, il est possible d'inférer la relation d'élaboration à partir d'une relation de sous-type entre les types d'éventualités, qui peut se refléter par une relation de sous-type lexical, c'est-à-dire par de l'hyponymie. Nous avons souhaité examiner des réalisations attestées de l'élaboration, afin d'observer les relations lexicales qui y sont impliquées. Cette observation a eu lieu sur le corpus d'exemples dont nous avons décrit la constitution dans la section précédente. À cette étape, nous n'avons pas projeté les *Voisins de Wikipédia* sur ce corpus d'exemples. En effet, si notre objectif à terme est de déterminer si l'on peut concilier l'hypothèse de la SDRT selon laquelle l'élaboration est marquée lexicalement avec une détection automatique *via* notre ressource, dans un premier temps, nous avons souhaité analyser les aspects lexicaux impliqués dans la relation d'élaboration sans nous cantonner ni à la relation de sous-type présupposée par la SDRT d'une part, ni aux relations lexicales détectables par la projection de notre ressource d'autre part. Toutefois, pour rester dans un souci d'automatisation possible, nous ne considérons dans ce qui suit que des unités monolexicales (des mots graphiques).

Nous présentons ci-dessous trois exemples d'élaboration que nous commentons

(8.2.2.1), avant d'en tirer quelques éléments de synthèse (8.2.2.2).

8.2.2.1 Quelques exemples...

(4) [Un véhicule a effectué une spectaculaire sortie de route, hier vers 18h15, sur l'A36.]1 [La voiture circulait dans le sens Mulhouse-Montbéliard]2 [lorsqu'après être passée à hauteur du 35e RI,]3 [elle a quitté la chaussée sur sa droite.]4 [Frôlant le début d'une glissière de sécurité,]5 [le véhicule a gravi le talus,]6 [basculé de l'autre côté,]7 [traversé un champ,]8 [est entré dans un secteur boisé,]9 [pour finalement plonger vers le centre Leclerc dans une zone à pic.]10

Dans l'exemple (4), trois liens lexicaux permettent d'inférer la relation discursive $Subtype_D(\pi_1, \pi_4)$: les liens `véhicule`/`voiture`, `sortie`/`quitter` et `route`/`chaussée`. Si le premier de ces liens est bien un cas d'hyperonymie et peut donc être rapproché du *Subtype* de la SDRT, le lien `route`/`chaussée` relève quant à lui plutôt de la méronymie ; enfin, la paire `sortie`/`quitter`, très importante pour établir la relation de sous-type entre les deux événements, est intercatégorielle, donc difficilement caractérisable en termes de relations lexicales classiques.

(5) (...) [qui rappelle la vocation des bénévoles de l'association :]32 [être un soutien pour la paroisse,]33 [apporter une petite contribution financière aux travaux grâce aux manifestations et aux dons,]34 [accomplir de multiples tâches et démarches touchant aux bâtiment paroissiaux,]35 [contribuer à la convivialité entre les paroissiens.]36

Dans l'exemple (5), les événements des segments 33 à 36 sont tous des sous-types discursifs de "vocation des bénévoles de l'association". Au niveau lexical, cela se traduit par des liens de proximité sémantique (entre "vocation" et des mots comme "soutien", "apporter", "accomplir", "tâche" ou encore "contribuer"), auxquels on peut difficilement assigner une relation lexicale classique. Ces liens, qui s'établissent entre catégories de type différent, s'établissent dans ce discours, et n'apparaîtront sans doute pas dans une ressource traditionnelle.

(6) [C'est alors que le cyclone commence à augmenter de vitesse,]11 [en passant à 30 km/h.]12

Dans l'exemple (6), le segment 12 « en passant à 30 km/h » élabore le segment 11. C'est plus particulièrement la relation entre « km/h » et « vitesse » (les km/h sont une unité permettant d'exprimer une vitesse) qui permet d'inférer cette élaboration.

8.2.2.2 Quelques éléments de synthèse

La première remarque que l'on peut faire est que les relations lexicales impliquées dans la relation d'élaboration ne se limitent pas à la relation de sous-type lexical (c'est-à-dire d'hyponymie). Dans l'exemple classique d'Asher et Lascarides (2003a) (ex. (2)), la relation de sous-type discursif entre des événements comme « to eat a lovely meal » et « to devour salmon » est reflétée directement par des relations de sous-type lexical : « to devour » est un sous-type lexical de « to eat », et « salmon » est un sous-type lexical de « meal ». Dans les données réelles observées, on rencontre rarement de telles correspondances. Ainsi, dans l'exemple (4), la relation de sous-type discursif qui existe entre les événements « effectuer une spectaculaire sortie de route » et « quitter la chaussée sur la droite » n'est pas directement reflétée par des relations de sous-type lexical (il est difficile de considérer « quitter » comme un sous-type lexical d'« effectuer » par exemple). Pour retrouver des relations relevant du sous-type, il faudrait se situer à une granularité supérieure au mot (on pourrait considérer que « quitter » est un sous-type d'« effectuer une sortie » par exemple). Si, dans un souci d'automatisation de la détection de ces relations au niveau lexical, on préfère éviter de prendre en compte des expressions polylexicales, il faut dès lors s'ouvrir à des relations lexicales plus larges, comme celle entre « sortie » et « quitter » ou « route » et « chaussée ».

Notre deuxième remarque, qui découle partiellement de la première, est que les relations lexicales impliquées dans la relation d'élaboration peuvent concerner différentes catégories et positions morpho-syntaxiques. Celles qui paraissent les plus importantes sont celles établissant un lien entre les verbes ou entre les objets des verbes des deux segments considérés. Mais on trouve également des relations pertinentes en dehors de ce schéma, comme avec la paire « quitter » / « sortie » (verbe du segment élaboré / nom appartenant à l'objet du verbe du segment élaborant) ou dans l'exemple (5) avec les paires « vocation » / « accomplir » et « vocation » / « contribuer » (nom en position syntaxique d'objet dans le segment élaboré / verbes des segments élaborants).

Ainsi, les liens lexicaux impliqués dans l'interprétation de la relation d'élaboration sont d'après nos observations assez peu contraints : ils ne se limitent pas à la relation de sous-type lexical, et peuvent apparaître à différentes positions syntaxiques. Pour autant, il ne faut pas considérer tous les liens de cohésion lexicale reliant un segment élaboré et un segment élaborant comme des manifestations de la relation discursive d'élaboration. Selon nous, une distinction importante existe entre les liens lexicaux accompagnant l'élaboration – il est normal que deux segments entretenant une relation de discours quelle qu'elle soit (ou plus généralement deux segments proches quelconques dans un texte donné) entretiennent des liens de cohésion lexicale – et les liens lexicaux permettant d'inférer l'élaboration. Cette différence fine peut être établie en montrant que certains liens de cohésion lexicale peuvent être modifiés sans incidence au niveau discursif, alors que d'autres sont indispensables à l'interprétation. Ainsi, si l'on modifie l'exemple (6) en remplaçant

« à 30 km/h » par « sur Madagascar » (*cf.* ex. (7)), la relation discursive rattachant le segment 12 au segment 11 devient une relation d'Arrière-plan plutôt que d'élaboration.

(7) [C'est alors que le cyclone commence à augmenter de vitesse,]11 [en passant sur Madagascar.]12

Nos deux premières remarques amènent à considérer les voisins distributionnels comme une bonne option *a priori* dans l'optique d'un repérage automatique de l'élaboration. Notre dernière remarque incite à bien préciser les modalités selon lesquelles ils seraient exploités. En effet, comme nous le montrons dans les chapitres qui précèdent, la projection des voisins distributionnels permet de mettre au jour de manière large des phénomènes de cohésion lexicale dans les textes ; il paraît dès lors délicat de distinguer liens essentiels et liens accompagnants. Nous posons et illustrons plus précisément ce problème dans la section qui suit.

8.2.3 Utiliser des relations de voisinage distributionnel pour détecter l'élaboration ?

Comme nous l'avons souligné ci-dessus, la variété des liens lexicaux impliqués dans la relation d'élaboration dépasse les relations de sous-type lexical et peut-être rapprochée de la variété des liens observés dans les voisins. Nous n'avons pas quantifié les différents types de relations lexicales observées (mais de toute façon elles ne sont pas quantifiées dans notre base de voisins non plus). On peut donc espérer capter les liens impliqués dans l'élaboration en projetant les voisins. Mais nous avons également souligné que tous les liens lexicaux accompagnant la relation d'élaboration ne sont pas impliqués dans l'interprétation discursive. Ces liens périphériques de cohésion lexicale ont de grandes chances d'être également captés par les voisins (problème qui se poserait à moindre mesure avec une ressource constituée uniquement d'hypéronymes). Il est risqué de s'appuyer sur ces liens pour détecter l'élaboration : les phénomènes de cohésion lexicale couvrent l'intégralité des textes, et concernent toutes les relations de discours. Nous avons d'ailleurs montré dans la section 7.2.3 que l'élaboration n'est pas la seule relation qui est associée à des phénomènes de forte cohésion lexicale.

Nous illustrons ces considérations avec la figure 8.1, qui reprend l'exemple (4). Nous y représentons les liens de voisinage existants entre le segment 1, « Un véhicule a effectué une spectaculaire sortie de route, hier vers 18h15, sur l'A36. », et tout ce qui suit.

Dans cet exemple, on observe des liens pertinents pour inférer la relation d'élaboration, c'est-à-dire les liens déjà mentionnés `véhicule`/`voiture`, `sortie`/`quitter` et `route`/`chaussée`. On observe également d'autres liens qui participent à la cohésion lexicale de l'ensemble, sans intervenir dans l'interprétation d'élaboration, comme les

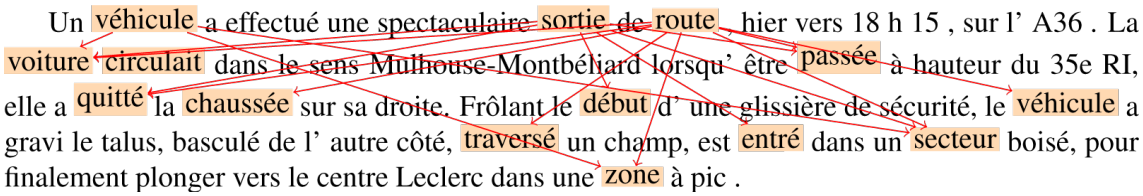


FIGURE 8.1 – Projection des voisins sur l'exemple (4)

liens `route`/`voiture` ou `route`/`véhicule`. Enfin, certains liens ne sont pas pertinents dans ce contexte donné, comme ici `route`/`traverser`. On peut aussi remarquer que la cohésion lexicale mise au jour concerne également les segments qui ne sont pas en relation d'élaboration avec le segment 1 : les segments 2 et 3, « La voiture circulait dans le sens Mulhouse-Montbéliard]2 [lorsqu'après être passée à hauteur du 35e RI,]3 » (relation d'Arrière-plan). Ainsi, comme nous le prévoyions, la projection des voisins permet ainsi de détecter les liens essentiels à l'interprétation discursive, mais également des liens périphériques de cohésion lexicale ainsi que des liens accidentels.

Nous percevons donc une conciliation possible entre l'hypothèse sous-jacente à la règle d'inférence de la SDRT (l'élaboration est marquée par la cohésion lexicale) et l'éventualité d'un repérage automatique de l'élaboration en utilisant les voisins distributionnels : les indices lexicaux pertinents vont au-delà de ce qui est présupposé par la SDRT et se rapprochent en cela de ce qui est détectable par les voisins ; mais les voisins doivent être canalisés pour être exploitables.

Dans le chapitre 7, nous avons présenté une expérience d'identification de l'élaboration utilisant le voisinage distributionnel ; dans cette expérience, l'élaboration était identifiée par contraste avec une relation établissant peu de cohésion lexicale, la relation d'e-élaboration ; cette stratégie a permis d'éluder la question de la sélection des liens de cohésion lexicale pertinents pour l'inférence de l'élaboration. Dans un objectif de détection de l'élaboration en corpus, nous envisageons deux stratégies possibles pour restreindre les liens de voisinage projetés.

- La première consisterait à cibler les liens de voisinage, c'est-à-dire à rechercher des liens dans des positions syntaxiques très spécifiques, au sein de constructions prédéfinies.
- La seconde consisterait à combiner le voisinage distributionnel avec d'autres marqueurs de l'élaboration ; il s'agirait donc encore une fois d'une détection de l'élaboration par contraste avec d'autres relations concernées par le même marqueur mais étant généralement moins lexicalement cohésives.

Dans la section 8.3, nous présentons une expérience pratique de détection de l'élaboration qui allie ces deux stratégies, en proposant de combiner le voisinage distributionnel avec la construction gérondive. En effet, comme nous allons le voir en 8.3.1, le gérondif est un marqueur ambigu de l'élaboration (illustré précédemment par l'exemple (6)). De plus, dans le cadre d'une identification automatique, il offre de nombreux avantages, de par son statut de construction syntaxique : il peut être repéré fiablement, et l'identification et la délimitation des segments à relier sont

déterminés par la syntaxe (le segment à attacher est le syntagme gérondif et le segment cible la proposition principale, *cf.* 8.3.1) ; tout cela nous permettra de nous focaliser sur l'apport des indices lexicaux.

8.3 Détecter fiablement des cas d'élaboration par combinaison d'indices

Suite à la phase exploratoire décrite en 8.2, nous avons choisi de combiner les voisins distributionnels avec un marqueur ambigu de l'élaboration : le gérondif (8.3.1). Nous avons pour cela proposé deux modes de combinaison de ces indices et extrait toutes leurs occurrences au sein du corpus WikipédiaFR2007 (8.3.2), afin d'observer leur fiabilité (8.3.3 et 8.3.4). Il s'agit cette fois d'une approche sémasiologique : on part d'un marqueur potentiel et on observe ses réalisations.

8.3.1 Le gérondif : un marqueur ambigu de l'élaboration

Vladimir. – Qu'est-ce qu'on fait maintenant ?
Estragon. – On attend.
Vladimir. – Oui, mais en attendant.
Samuel Beckett, En attendant Godot

Le gérondif est une forme verbale non conjuguée, admettant toutes les expansions du verbe (compléments et modifieurs), mais dont le sujet n'est pas exprimé. Le gérondif et ses compléments et modifieurs forment un syntagme gérondif, qui présente le fonctionnement d'un complément adverbial. Il est généralement rattaché à un prédicat verbal, comme dans l'exemple (8). On appelle construction gérondive le couple formé par la proposition principale et le syntagme gérondif.

(8) Elle est sortie de la ville en longeant la mer. (Halmøy, 2003, p. 72)

Sur le plan syntaxique, la construction gérondive établit donc un lien de subordination entre deux verbes. Ce lien peut être associé à différentes valeurs sémantiques ; ces valeurs ne sont pas véhiculées par le gérondif lui-même mais dépendent des relations sémantiques en jeu entre les deux verbes, ainsi éventuellement que d'autres éléments contextuels (Halmoy, 1982). Sur le plan discursif, le syntagme gérondif peut, selon l'analyse de Vergez-Couret (2010, pp. 218-230), être attaché à la proposition principale par plusieurs relations de discours : élaboration, résultat, explication, conditionnel et circonstance accompagnante (cette dernière relation étant proposée par Vergez-Couret (2010)). Le gérondif peut donc être considéré comme un marqueur ambigu de l'élaboration.

Nous faisons l'hypothèse qu'en combinant le gérondif avec le voisinage distributionnel, il est possible de mettre au point un marqueur fiable de l'élaboration. Nous illustrons cette hypothèse avec les exemples (9), extraits de WikipédiaFR2007, pour

lesquels nous mettons en rapport une analyse SDRT et la présence ou l'absence de liens de voisinage distributionnel.

(9) United Fruit Company *investit* dans le pays, **en achetant** des parts dans le chemin de fer, l'électricité et le télégraphe.

(10) Dans la ville de Koriko, Kiki, accompagnée de son chat noir Jiji, *distribue* des colis **en volant** sur son balai.

En (9), la proposition principale introduit l'événement « investir dans le pays » et le syntagme gérondif introduit l'événement « acheter des parts ». Le type de l'événement du syntagme gérondif est un sous-type de celui de la proposition principale ; la relation d'élaboration est donc inférée. Cette relation discursive se reflète au niveau lexical par la proximité sémantique entre les verbes « investir » et « acheter », proximité captée par l'analyse distributionnelle puisque ce couple est présent dans les *Voisins de Wikipédia*. En (10), la proposition principale et le syntagme gérondif décrivent deux procès se déroulant simultanément : le syntagme gérondif pourrait être remplacé par la forme « tout en volant sur son balai » ; la relation de Circonstance accompagnante est ici inférée. La relation s'établissant entre les deux procès est fortuite, et n'est pas captée par la projection des voisins distributionnels : « distribuer » et « voler » ne sont pas voisins.

8.3.2 Combiner le gérondif avec le voisinage distributionnel : mise en oeuvre

Nous présentons ici comment nous avons mis en œuvre la combinaison du gérondif et des voisins distributionnels. La figure 8.2 résume la procédure décrite dans cette section.

8.3.2.1 Marqueurs envisagés

Comme souligné dans la section 8.3.1, le gérondif est un marqueur faible de l'élaboration ; par la suite, nous notons ce marqueur ***G***. Notre hypothèse est qu'en combinant le gérondif avec des liens de voisinage ciblés, il est possible de former des marqueurs plus fiables de la relation d'élaboration. Cette combinaison peut se faire de plusieurs manières, selon les positions syntaxiques des items lexicaux entre lesquels on pose la contrainte d'un lien de voisinage. Nous présentons ici les deux marqueurs que nous avons envisagés.

Le gérondif, par rapport aux marqueurs lexicaux (ou *cue phrases*) tels que « plus particulièrement », présente ici l'avantage d'être une structure syntaxique, ce qui favorise son association avec des liens lexicaux ciblés. La stratégie la plus évidente consiste ainsi à contraindre la présence d'un lien de voisinage entre le gérondif lui-même et le verbe de la proposition principale. Ce marqueur est illustré par l'exemple (11) ; dans cet exemple, la proximité sémantique entre les verbes « envoyer » et

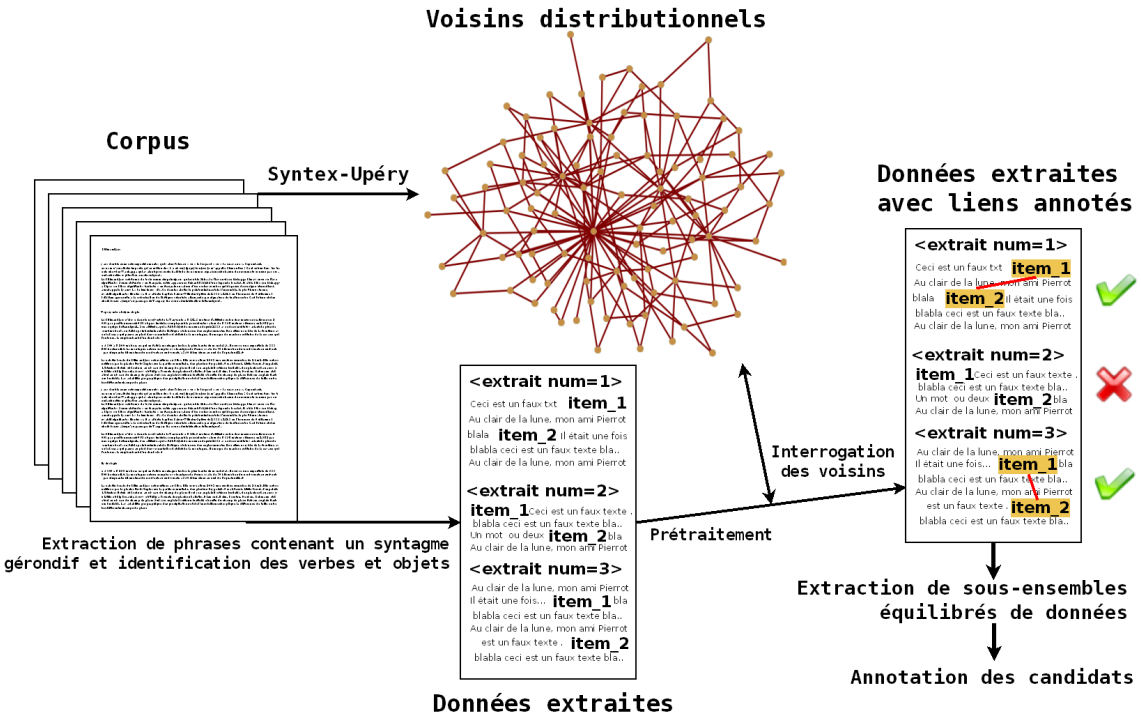


FIGURE 8.2 – Procédure d'interrogation des voisins sur des données particulières

« contribuer », saisie par la présence d'un lien de voisinage, contribue selon nous à signaler la relation d'élaboration liant le syntagme gérondif à la proposition principale. Nous notons désormais ce marqueur ***G***+VV : l'élaboration est détectée si les Verbes (de la proposition principale et du syntagme gérondif) sont Voisins.

(11) [...] et les villages contribuaient également à ce grand projet religieux ***en envoyant*** des vivres.

L'autre stratégie que nous avons souhaité tester a consisté à imposer, en plus du lien entre gérondif et verbe de la proposition principale, un lien de voisinage entre les objets de ces deux verbes. Le marqueur obtenu est illustré par l'exemple (12) ; ici, deux liens de voisinage aident selon nous à inférer l'élaboration : celui reliant les verbes « élargir » et « englober », et celui reliant les noms « empire » et « mondes ». Nous notons désormais ce marqueur ***G***+VV+OV : l'élaboration est détectée si les Verbes de la proposition principale et du syntagme gérondif sont Voisins, *ET* si les Objets des verbes sont reliés par au moins un lien de Voisinage.

(12) Les Skrulls [...] élargissent leur empire ***en englobant*** dans celui-ci les mondes moins avancés qu'ils rencontrent.

Les trois marqueurs considérés sont donc de plus en plus restrictifs : les cas marqués par ***G***+VV+OV constituent un sous-ensemble de ceux marqués par ***G***+VV, eux-

mêmes inclus parmi les cas marqués par ***G***. Dans le cas de ***G***+VV+OV, la contrainte ajoutée est double, puisqu'il faut que les deux verbes concernés possèdent des objets pour que ceux-ci puissent être reliés par un lien de voisinage, ce qui n'est bien sûr pas toujours le cas. Si notre hypothèse se vérifie, chacun de ces marqueurs devrait donc permettre d'extraire des candidats à l'élaboration avec une précision croissante ; nous présentons dans la section qui suit la mise en oeuvre de cette extraction.

8.3.2.2 Extraction des candidats

Nous avons choisi de projeter nos marqueurs dans le corpus WikipédiaFR2007, déjà caractérisé dans la section 2.2.1 (p. 58).

Dans la mesure où nous recherchons une construction syntaxique spécifique, nous avons choisi de travailler sur la version analysée syntaxiquement de ce corpus (c'est-à-dire sur l'analyse syntaxique effectuée par SYNTEX en amont de la construction des voisins), plutôt que sur une version annotée morpho-syntaxiquement comme précédemment. Nous montrons ci-dessous (13) un extrait de l'analyse syntaxique fournie par SYNTEX pour l'exemple (12). [3]

(13) <TXT> [...] les Skrulls [...] élargissent leur empire en englobant dans celui -ci les mondes moins avancés qu' ils rencontrent .
<ETIQ> (...)
`Det??|le|les|9|DET;10|`
`NomPrXXInc|Skrull|Skrulls|10||DET;9`
(...)
`VCONJP|élargir|élargissent|15||OBJ;17`
`DetMS|leur|leur|16|DET;17|`
`Nom?S|empire|empire|17|OBJ;15|DET;16`
`Prep|en|en|18||NOMPREP;19`
`Ppr|englober|englobant|19|NOMPREP;18|PREP;20,OBJ;24`
(...)
`Det??|le|les|23|DET;24|`
`Nom?P|monde|mondes|24|OBJ;19|DET;23,ADJ;26`
`Adv|moins|moins|25|ADV;26|`
`PpaMP|avancer|avancés|26|ADJ;24|ADV;25`
(...)

À partir de cette analyse syntaxique, il est possible de détecter les gérondifs à partir du motif SYNTEX ci-dessous, que l'on peut lire « la préposition "en" régit un verbe au participe présent *via* la relation `NOMPREP` » :

`Prep|en|en|`*n*`|.*|NOMPREP;`*n+1* `Ppr|.*|.*|`*n+1*`|NOMPREP;`*n*`|.*`

3. Rappelons qu'une étiquette posée par SYNTEX contient les informations suivantes : catégorie|lemme|forme|numéro|régi par|recteur de (*cf.*2.2.2 p. 58)

En parcourant les étiquettes de mot régi en mot régi, il est également possible de repérer et de délimiter les objets des différents verbes.

Une fois extraits tous les gérondifs, on peut vérifier la présence (ou non) de liens de voisinage aux positions spécifiées par les différents marqueurs. La démarche est différente de celle que l'on avait jusqu'ici adoptée : on ne projette pas *a priori* tous les liens de voisinage dans un environnement donné (texte ou portion de texte), mais on sélectionne des items lexicaux particuliers (ici, par leur position syntaxique au sein de la phrase, mais on pourrait également envisager d'autres critères) et on vérifie s'ils entretiennent ou non un lien de voisinage. Comme nous l'avons montré en 4.5.2 (page 134), cette démarche différente laisse envisager un filtrage moindre des voisins ; ainsi, nous avons uniquement placé des seuils sur les rangs des voisins, tel qu'expliqué dans la section 4.5.2.3.

G	***G*+VV**	***G*+VV+OV**
18571	375	193

TABLEAU 8.2 – Nombre de candidats extraits par la projection de chaque marqueur

Le tableau 8.2 indique le nombre de candidats extraits par la projection de chaque marqueur sur le corpus WikipédiaFR2007. On peut noter que les candidats extraits par les deux marqueurs faisant appel au voisinage distributionnel sont en très faible nombre par rapport au nombre de gérondifs extraits ; rappelons en outre que la relation d'élaboration a de très nombreuses réalisations, qui ne se limitent pas aux constructions gérondives. On est donc ici très loin d'une détection de la relation d'élaboration en général, puisque l'on ne pourra se prévaloir que de la détection de certains cas bien particuliers de cette relation (on remarquera notamment qu'on ne détecte ici que des élaborations intraphrastiques). Mais ce faible rappel ne constitue pas selon nous un achoppement à cette étude ; mettre au point un marqueur fiable de l'élaboration, même à très faible couverture, constitue un enjeu important :

- dans la problématique de cette thèse, puisque cela encouragera à utiliser des indices basés sur la cohésion lexicale ;
- pour les recherches sur cette relation de discours peu étudiée car souvent considérée comme non-marquée 8.1.1.2.

De plus, si la combinaison du gérondif avec les voisins distributionnels s'avère performante, cela laissera envisager de combiner ces derniers avec d'autres marqueurs ambigus de l'élaboration, tels que « Premièrement », « Dans un premier temps », etc., si l'objectif est d'atteindre une plus forte couverture dans la détection automatique de cette relation.

8.3.3 Annotation des candidats

Une fois définis les marqueurs, afin d'évaluer si effectivement ils signalent de manière fiable la relation d'élaboration, il faut procéder à une annotation des données qu'ils permettent d'extraire. En vue de cette annotation, nous avons sélectionné de

manière aléatoire un sous-ensemble de phrases parmi celles extraites avec nos différents marqueurs, en équilibrant les différentes catégories (***G***, ***G***+VV et ***G***+VV+OV). L'annotation a été effectuée par deux annotateurs experts [4]. Nous avons choisi de constituer une référence la plus fiable possible en procédant à une annotation en deux phases :

- dans un premier temps, chaque annotateur annote la totalité des données sélectionnées ;
- dans un second temps, les cas de désaccord sont à nouveau soumis aux annotateurs (réunis) afin de constituer la référence.

Cette annotation s'est effectuée *via* une interface d'annotation que nous avons développée et que nous décrivons dans la section qui suit (8.3.3.1).

8.3.3.1 Développement d'une interface d'annotation

Afin de faciliter l'annotation, nous avons choisi de développer une interface Web en PHP / MySQL. Cette interface a permis de procéder aux deux phases de l'annotation, puis, une fois celle-ci terminée, d'explorer les résultats.

L'ordre dans lequel les extraits sont présentés aux annotateurs a été fixé de manière aléatoire, mais il est le même pour les deux annotateurs. Les candidats extraits par ***G***, ***G***+VV, et ***G***+VV+OV sont donc mélangés et il n'est pas possible de les identifier. Après s'être identifié, l'annotateur accède à l'annotation. Les candidats lui sont soumis un à un avec une simple question : « Le gérondif marque-t-il ici une élaboration ? ». La figure 8.3 permet de visualiser un candidat tel qu'il est présenté à l'annotateur. Chaque réponse vient compléter la base de données MySql (*cf.* figure 8.4) répertoriant l'ensemble des candidats et les informations qui leur sont associées. La tâche d'annotation peut être interrompue et reprise à loisir, et a ainsi pu se dérouler sur plusieurs jours.

La première phase d'annotation se termine quand plus de 100 candidats ont été annotés pour chaque marqueur. Une fois l'annotation effectuée, c'est *via* cette même interface que se font la constitution de la référence et l'exploration des résultats. Nous présentons ces fonctionnalités et les résultats des deux phases d'annotation dans les sections qui suivent.

8.3.3.2 Bilan de la première phase d'annotation

Le tableau 8.3 présente le nombre de candidats doublement annotés pour chaque marqueur. La matrice de confusion obtenue suite à l'annotation des ces 314 candidats est donnée par le tableau 8.4. Les nombres en gras représentent les cas d'accord : les deux annotateurs s'accordent sur 223 cas d'élaboration et 57 cas de non-élaboration, ce qui correspond à un taux d'accord de 89%. Le coefficient Kappa (Cohen, 1960) (qui permet d'annuler l'effet du déséquilibre entre les classes en tenant compte de

4. M. Vergez-Couret et nous-même.

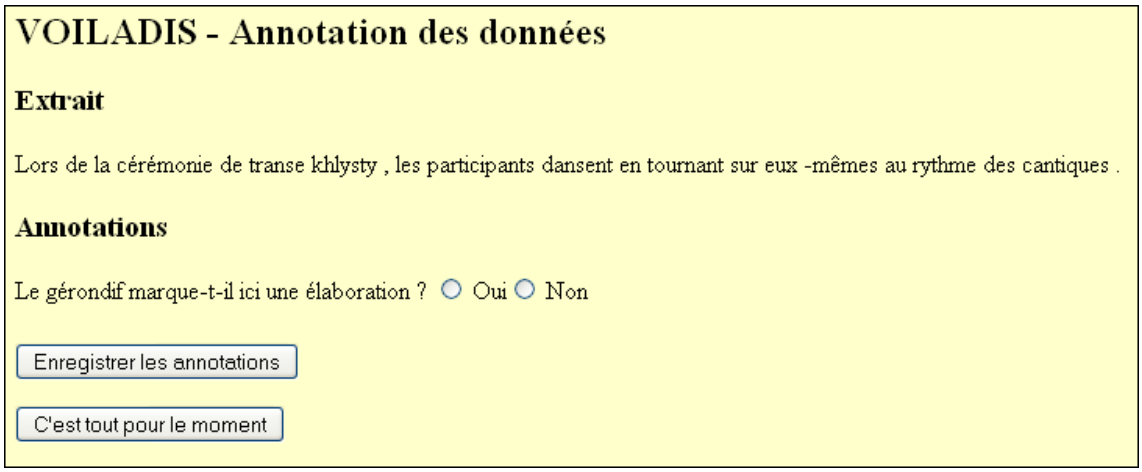


FIGURE 8.3 – Interface développée - Phase d'annotation

G	*G*+VV	*G*+VV+OV	TOT.
108	101	105	314

TABLEAU 8.3 – Nombre de candidats annotés au terme de la première phase d'annotation

l'accord aléatoire), calculé à partir de ces données, s'élève à 0.70. Ce score correspond à un accord inter-annotateurs « fort », selon l'échelle proposée par Landis (1977). Tout en restant le reflet d'une tâche difficile, il permet d'envisager que cette tâche soit automatisable.

	élaboration	non élaboration	TOT.
élaboration	**223**	21	244
non élaboration	13	**57**	70
TOT.	236	78	314

TABLEAU 8.4 – Matrice de confusion

8.3.3.3 Seconde phase d'annotation : constitution de la référence

Lors de la seconde phase d'annotation, les cas de désaccord (au nombre de 34, *cf.* tableau 8.4) sont à nouveau présentés aux annotateurs réunis. Ceux-ci peuvent alors convenir d'une nouvelle annotation, ou maintenir leur désaccord. La figure 8.4 présente un extrait de la base de données MySQL utilisée. On peut y voir quatre cas de désaccord entre les deux annotateurs (identifiés par les préfixes « cle » et « mar »), aux lignes 2, 5, 7 et 8. Ces cas ont donc été réannotés lors de la seconde phase d'annotation. Trois d'entre eux ont finalement, après discussion, suscité un accord ; le dernier (ligne 7) n'a pas abouti à un tel accord.

			id	texte	gerv	objv	cle_elab	cle_etat	mar_elab	mar_etat	ref_elab	ref_etat
☐ 1	✎	✕	10-18021_1-18021_6	<item num="1" lemme="diminuer" cat="VINF">diminuer...	1	1	1	annote	1	annote	1	annote
☐ 2	✎	✕	10-45796_23-45796_24	La ligne n' évita pas l' effondrement de la France...	1	0	0	annote	1	annote	1	annote
☐ 3	✎	✕	1-10586_10-10586_11	Lors de la cérémonie de transe khlysty , les parti...	0	0	1	annote	1	annote	1	annote
☐ 4	✎	✕	1-10975_9-10975_11	Les villes qui se situent sur ses rives <item pos=...	0	0	0	annote	0	annote	0	annote
☐ 5	✎	✕	1-11475_6-11475_9	Les humains influencés par Khorne <item pos="1" le...	0	0	0	annote	1	annote	1	annote
☐ 6	✎	✕	1-11786_29-11786_33	Selon les anales historiques ismaéliennes , à la f...	0	0	0	annote	0	annote	0	annote
☐ 7	✎	✕	1-12061_2-12061_13	Elle <item pos="1" lemme="être" cat="VCONJSp">est<...	0	0	0	annote	1	annote	*NULL*	annote
☐ 8	✎	✕	1-12079_15-12079_21	On peut s' y rendre par le train de la ligne du TG...	0	0	1	annote	0	annote	1	annote

FIGURE 8.4 – Interface développée - Extrait de la base de données MySql

Au terme de la seconde phase d'annotation, une décision a été prise pour 26 cas de désaccord initial. Dans les 8 cas restants, le désaccord a été maintenu. Ces cas ont été exclus de la référence constituée. Le tableau 8.5 présente les nombres de candidats pour chaque marqueur au terme de cette seconde phase d'annotation (c'est-à-dire les nombres de candidats ayant suscité un accord soit lors de la première phase d'annotation, soit après discussion).

G	***G***+VV	***G***+VV+OV	TOT.
102	100	104	306

TABLEAU 8.5 – Nombre de candidats annotés au terme de la seconde phase d'annotation

Parmi les 8 cas dits de « désaccord », on trouve principalement des candidats pour lesquels deux interprétations concurrentes sont possibles. Ainsi, l'exemple (14) peut selon nous susciter deux lectures :

- soit on considère que les actions « défendre ses vaisseaux » et « attaquer ceux des adversaires » ne peuvent absolument pas participer à l'action « ramasser des pièces d'or », et que le syntagme gérondif fournit ici un arrière-plan au procès décrit par la proposition principale (relation de Circonstance accompagnante) ;
- soit on se replace dans le contexte d'un jeu (le contexte élargi, auquel pouvaient accéder les annotateurs, nous apprend que l'on se trouve ici dans une description des règles du jeu « Korsar »). On peut alors imaginer que, dans le contexte de ce jeu, l'action « ramasser des pièces d'or » s'effectue directement

en défendant ses vaisseaux et en attaquant ceux des adversaires, et qu'il y a donc une relation d'élaboration entre ces événements.

(14) Le but est de ramasser le plus de pièces d'or en utilisant ses corsaires pour défendre ses vaisseaux et attaquer ceux des adversaires.

L'interface que nous avons développée permet de visualiser plus en détail les données annotées *via* une interrogation par formulaire. La figure 8.5 montre ainsi la requête posée pour rechercher tous les exemples exclus de la référence au terme de la seconde phase d'annotation. Il est également possible *via* ce formulaire de croiser

FIGURE 8.5 – Interface développée - Visualisation des données

les résultats de l'annotation avec les différents marqueurs ayant servi à extraire les candidats ; on peut ainsi examiner qualitativement les réalisations de chaque marqueur. Avant de mener ces observations plus fines, nous présentons dans la section qui suit les résultats obtenus concernant la précision globale de chaque marqueur.

8.3.4 Résultats

Le tableau 8.6 synthétise nos résultats. Pour chaque marqueur, sont précisés :
- le nombre de candidats extraits dans le corpus ;
- le nombre de candidats annotés dans la référence ;
- parmi ces candidats, le nombre de candidats pour lesquels les annotateurs ont identifié une relation d'élaboration, et le nombre de candidats pour lesquels une autre interprétation a été préférée ;
- la précision du marqueur : $\frac{\text{nb d'elab.}}{\text{nb de candidats annotés}}$;
- l'intervalle de confiance.

	Extraits	**Annotés**	**Elab.**	**Non Elab.**	**Précision**	**Inter. conf.**
G	18571	102	62	40	**60.8%**	9.45%
G+VV	375	100	81	19	**81.0%**	6.59%
G+VV+OV	193	104	99	5	**95.2%**	2.8%

TABLEAU 8.6 – Résultats : fiabilité des différents marqueurs

Ces résultats confirment tout d'abord que le gérondif constitue un marqueur ambigu de l'élaboration, puisque seuls 60.8% des candidats annotés sont des élaborations. Nous avons fait le choix d'équilibrer les trois catégories de candidats pour l'annotation, malgré les différences très importantes de taille entre ces trois catégories. Ainsi, nous avons annoté plus de la moitié de la classe ***G***+VV+OV, mais une très faible proportion de ***G*** (un peu plus d'un pour 200). Cela se traduit par un intervalle de confiance très élevé pour cette catégorie (près de 10%). Toutefois, la différence de performance entre ***G*** et nos marqueurs ***G***+VV et ***G***+VV+OV est suffisamment importante pour assurer que la combinaison avec le voisinage distributionnel apporte une amélioration significative de la précision avec laquelle l'élaboration est détectée.

En effet, nos deux marqueurs obtiennent de très bons résultats. Le premier, ***G***+VV, signale l'élaboration dans 81% des cas. Le second est particulièrement fiable puisqu'il signale l'élaboration dans plus de 95% des cas (avec un intervalle de confiance inférieur à 3%). Ces résultats sont extrêmement encourageants dans la perspective d'une détection automatique de la relation d'élaboration.

Nous avons observé les cas dans lesquels ce marqueur échouait. Dans certains cas, on peut attribuer l'échec au lien de voisinage qui n'est pas pertinent. Ainsi, dans l'exemple (15), le lien `marcher/incendier` est difficilement interprétable dans ce contexte. Dans d'autres cas, un marqueur est présent qui pourrait annuler l'inférence d'élaboration. Ainsi, dans l'exemple (16), on trouve le connecteur « mais », qui est marqueur fort de la relation discursive de contraste. Nos résultats pourraient donc être encore améliorés en prenant en compte des marqueurs d'autres relations de discours.

(15) Ils *marchent* la campagne en *incendiant* toutes les habitations.

(16) Le roi d'Espagne lui accorda une décoration qu'il accepta, *mais* en refusant la pension qui y était attachée.

8.4 Bilan et perspectives

Partant de l'hypothèse de la SDRT selon laquelle la relation d'élaboration pourrait être inférée à partir d'éléments relevant de la cohésion lexicale (8.1), ce chapitre a présenté une exploration qualitative de cette hypothèse sur des données attestées (8.2) et une expérience pratique de détection automatique de la relation d'élaboration (8.3). Le bilan de ces travaux est très positif, que ce soit concernant les recherches sur la relation d'élaboration ou concernant les objectifs annoncés de cette thèse, à savoir valider l'utilisation des voisins distributionnels pour aborder la structuration du discours.

8.4.1 Sur la relation d'élaboration

À l'issue de l'expérience relatée dans ce chapitre (section 8.3), nous avons à nouveau confirmé que la relation d'élaboration est bien marquée par la cohésion lexicale, et avons montré une modalité particulière de ce marquage. L'emploi d'indices lexicaux permet de sélectionner de manière fiable, parmi un ensemble de cas présentant la même construction syntaxique, des exemples d'élaboration. D'un point de vue théorique, nous confirmons donc l'hypothèse de la SDRT et l'illustrons de données réelles. D'un point de vue pratique, nous avons proposé des modalités d'utilisation d'une ressource lexicale pour la détection de l'élaboration, tout en en montrant les limites. De nombreuses perspectives s'offrent à nous dans l'optique d'une détection plus large de la relation d'élaboration.

- De nombreux marqueurs ambigus de la relation d'élaboration ont été identifiés. Nous pouvons tenter de les combiner avec les voisins distributionnels afin de détecter de manière fiable des cas dans lesquels ces marqueurs identifient la relation d'élaboration.
- Concernant les cas fréquents d'élaboration dans lesquels aucun marqueur discursif n'est présent, nous pouvons tenter une détection en ciblant certaines relations lexicales, notamment entre les verbes et les objets des segments concernés.

Le fait d'identifier de façon fiable, même en faible quantité, des exemples d'élaboration constitue un apport important dans l'étude de cette relation car, en plus de la validation des indices ayant été utilisés pour la détection, cela permet également de disposer de nouvelles données attestées pour l'étude de la relation.

8.4.2 Sur l'utilisation des voisins distributionnels dans des approches ascendantes du discours

Via les expériences décrites dans ce chapitre et dans le chapitre précédent, nous avons validé l'utilisation d'indices lexicaux dans le cadre d'une approche *bottom-up* du discours, et l'utilisation des voisins pour capter ces indices. Nous avons défini dans ce chapitre de nouvelles modalités d'exploitation des voisins distributionnels pour aborder le discours. Alors qu'on les avait jusqu'ici employés pour détecter la cohésion lexicale « en général », nous avons notamment proposé ici de les utiliser de manière ciblée, très fine, pour détecter certaines relations lexicales particulières apparaissant comme le corrolaire d'une relation de discours.

Là encore, les perspectives sont nombreuses. Nous nous sommes intéressée ici à la relation d'élaboration car sa détection constituait un défi, dans la mesure où elle est considérée comme non-marquée, et pouvait donc tout particulièrement bénéficier d'une prise en compte du niveau lexical. Mais la prise en compte d'informations lexicales peut selon nous être profitable à d'autres niveaux de l'analyse relationnelle de la structuration discursive, comme mis au jour par l'exploration relatée dans le chapitre 7. Il faudrait ainsi définir des modalités d'exploitation des voisins distributionnels pour traiter d'autres aspects de la structure du discours.

Conclusion générale

> Une œuvre n'est jamais nécessairement finie, car celui qui l'a faite ne s'est jamais accompli, et la puissance et l'agilité qu'il en a tirées, lui confèrent précisément le don de l'améliorer, et ainsi de suite... Il en retire de quoi l'effacer et la refaire.
>
> Paul Valéry

Dans cette thèse, nous avons présenté un ensemble de travaux visant à tester deux principaux aspects :

(a) l'adéquation d'une ressource distributionnelle à la problématique de l'identification de liens de cohésion lexicale ;

(b) la possibilité d'exploiter les indices ainsi repérés à tous les niveaux de l'étude de l'organisation textuelle.

Ces deux hypothèses de travail ont été vérifiées.

Concernant le premier point (a), nous avons montré que si une ressource distributionnelle s'avère très pertinente pour appréhender la variété des relations sémantiques impliquées dans la cohésion lexicale, il faut toutefois l'utiliser prudemment. De la construction d'une base distributionnelle à la production de sorties exploitables, nous avons montré pas à pas le parcours suivi et les traitements mis en place, avec un accent particulier sur les méthodes de filtrage des liens projetés.

Concernant le second point (b), nous avons fourni une série d'éclairages qui ont montré l'apport d'une prise en compte réfléchie de la cohésion lexicale pour une grande variété de problématiques liées à l'étude et au repérage automatique de l'organisation textuelle : segmentation thématique de textes, caractérisation des structures énumératives, étude de la corrélation entre lexique et structure rhétorique du discours et enfin détection de réalisations d'une relation de discours en particulier, la relation d'élaboration.

Étant donnée la variété des aspects traités, des bilans partiels ont été effectués à la fin de chaque chapitre. Dans cette conclusion, nous dressons un bilan plus global des apports des travaux présentés dans les deux parties « pratiques » de cette thèse (parties II et III) et évoquons les perspectives envisagées.

Les apports des travaux présentés dans la II^{e} partie concernent surtout :

(a) la réflexion sur l'analyse distributionnelle et la proposition d'une méthode pour l'évaluation de ressources distributionnelles : nous avons montré que l'annotation de liens projetés en contexte *via* l'interface que nous avons réalisée peut fournir des données fiables, pouvant être exploitées pour la caractérisation d'une base distributionnelle et l'exploration de l'impact de ses différents paramètres ;

(b) la caractérisation des sorties obtenues suite à la projection des voisins en texte : nos résultats permettent de mieux appréhender le type de sorties résultant d'une telle procédure, tant quantitativement – nombre de liens projetés et de couples

différents impliqués, répartition des liens (plus de liens « à proximité » que de liens « à distance »), couverture atteinte – que qualitativement par le biais de la phase d'annotation ;

(c) l'articulation des dimensions syntagmatique (*via* des critères contextuels) et paradigmatique (*via* le critère distributionnel) pour la détection de la cohésion lexicale : nous avons montré l'impact et la complémentarité des deux types d'indices, qui permettent de rendre compte de la possibilité d'instanciation de relations lexicales en contexte.

Les perspectives que nous souhaitons explorer concernent surtout ce dernier point. Nous aimerions combiner le voisinage distributionnel avec une palette plus large d'indices contextuels, calculés sur l'ensemble du corpus aussi bien que dans le texte particulier dans lequel on cherche à détecter les liens de cohésion lexicale ; ces indices contextuels seraient notamment basés sur l'application de patrons.

Concernant les travaux présentés dans la III^{e} partie de cette thèse, leur intérêt réside particulièrement dans :

(a) la variété des éclairages apportés sur l'organisation textuelle, et partant, la variété des modalités d'exploitation de la cohésion lexicale considérées : nous avons couvert des approches du discours allant du très global au très local ; la cohésion lexicale a été parfois envisagée de manière purement quantitative, *via* le calcul de scores de cohésion (notamment avec le travail sur la segmentation thématique), et d'autres fois de manière très qualitative et ciblée avec le repérage de liens se réalisant dans des contextes très précis dans lesquels il apparaissent comme le corolaire d'une relation de discours (travail sur l'élaboration). Ces deux stratégies peuvent bien sûr être combinées, et le travail sur les structures énumératives a amorcé une approche plus hybride, en exploitant à la fois des mesures quantitatives rendant compte de la force de leur cohésion lexicale et de la répartition des liens lexicaux, et une approche plus ciblée *via* la recherche de composantes en interaction avec la structure de SE.

(b) les retours qualitatifs systématiques sur les données traitées : pour les travaux qui font appel à la cohésion lexicale pour l'accès à la structure discursive, elle est un indice, souvent parmi d'autres, et ne fait généralement pas en elle-même l'objet d'investigations spécifiques. Nos recherches sont quant à elles ciblées sur la cohésion lexicale et sur les méthodes qui permettent de l'exploiter. Pour cette raison, nous avons toujours proposé d'examiner dans les textes les manifestations des phénomènes traités : « erreurs » du système de segmentation thématique, grande variabilité des scores de cohésion lexicale des structures énumératives, différences entre relations de discours, etc. Lorsque nous avons utilisé des indices lexicaux en combinaison avec d'autres indices (par exemple des indices syntaxiques pour la distinction entre élaboration et e-élaboration), nous avons évalué leur impact propre. Ce point de vue axé sur un mode de signalisation et sur des méthodes nous a permis de porter un regard particulier sur les tâches

abordées, notamment sur la tâche de segmentation thématique.

Partant de ce bilan, nos perspectives, en plus de celles mentionnées en fin de chaque chapitre, concernent l'exploitation de nos méthodes dans des approches de l'organisation textuelle émergeant de la prise en compte de la cohésion lexicale (à la manière par exemple des réseaux phrastiques de Hoey) et tirant parti de ses propriétés particulières, comme la possibilité de former des liens « longue distance ».

Bibliographie

Adam, C. et Morlane-Hondère, F. (2009). Détection de la cohésion lexicale par voisinage distributionnel : application à la segmentation thématique. *In Actes du colloque RECITAL, Senlis, France.*

Adam, C., Muller, P. et Fabre, C. (2010). Une évaluation de l'impact des types de textes sur la tâche de segmentation thématique. *In Actes de TALN'10*, Montréal, Canada.

Adam, C. et Vergez-Couret, M. (2010). Signalling elaboration : Combining gerund clauses with lexical cues. *In Proceedings of the 8th MAD (Multidisciplinary Approaches of Discourse) : Multidisciplinary Perspectives on Signalling Text Organisation*, pages 80–90.

Adam, C. et Vergez-Couret, M. (2012). Exploiting naive vs expert discourse annotations : an experiment using lexical cohesion to predict elaboration / entity-elaboration confusions. *In Proceedings of the Sixth Linguistic Annotation Workshop*, pages 22–30, Jeju, Republic of Korea. Association for Computational Linguistics.

Afantenos, S., Asher, N., Benamara, F., Bras, M., Fabre, C., Ho-dac, M., Draoulec, A. L., Muller, P., Péry-Woodley, M.-P., Prévot, L., Rebeyrolle, J., Tanguy, L., Vergez-Couret, M. et Vieu, L. (2012). An empirical resource for discovering cognitive principles of discourse organisation : the annodis corpus. *In Eighth Language Resources and Evaluation Conference (LREC 2012)*, Istanbul (Turkey).

Afantenos, S. D. et Asher, N. (2010). Testing sdrt's right frontier. *In Proceedings of the 23rd International Conference on Computational Linguistics*, COLING '10, pages 1–9, Stroudsburg, PA, USA. Association for Computational Linguistics.

Allan, J., Carbonell, J., Doddington, G., Yamron, J. et Yang, Y. (1998). Topic detection and tracking pilot study. final report. *In Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop.*

Asher, N. (1993). Reference to abstract objects in discourse. *In Studies in Linguistics and Philosophy*, volume 50. Kluwer Academic Publishers, Dordrecht, Amsterdam.

Asher, N. et Lascarides, A. (2003a). *Logics of conversation*. Cambridge :CUP.

Asher, N. et Lascarides, A. (2003b). *Logics of Conversation*. Cambridge University Press.

Baldridge, J. et Lascarides, A. (2005). Probabilistic head-driven parsing for discourse structure. *In Proceedings of the Ninth Conference on Computational Natural Language Learning (CoNLL-2005)*, pages 96–103, Ann Arbor, Michigan. Association for Computational Linguistics.

Baroni, M. et Lenci, A. (2008). Concepts and properties in word spaces. *Alessandro Lenci (ed.), From context to meaning : Distributional models of the lexicon in linguistics and cognitive science, Special issue of the Italian Journal of Linguistics*, 20(1):55–88.

Baroni, M., Trento, I., Lenci, A. et Pisa, I. (2009). One distributional memory, many semantic spaces. *Proceedings of the EACL 2009 Geometrical Models for Natural Language Semantics (GEMS), Athens, Greece*.

Barzilay, R. et Elhadad, M. (1997). Using lexical chains for text summarization. *In ACL'97/EACL'97 Workshop on Interlligent Scalable Text Summarization*, Madrid.

Barzilay, R., McKeown, K. R. et Elhadad, M. (1999). Information fusion in the context of multi-document summarization. *In Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, ACL '99, pages 550–557, Stroudsburg, PA, USA. Association for Computational Linguistics.

Beeferman, D., Berger, A. et Lafferty, J. (1997). Text segmentation using exponential models. *In Proceedings of the Second Conference on Empirical Methods in Natural Language Processing*, pages 35–46, Providence.

Beeferman, D., Berger, A. et Lafferty, J. D. (1999). Statistical models for text segmentation. *Machine Learning*, 34(1–3):177–210.

Beigman Klebanov, B. et Shamir, E. (2006). Reader-based exploration of lexical cohesion. *Language Resources and Evaluation*, 40:109–126. 10.1007/s10579-006-9004-6.

Beigman Klebanov, B. et Shamir, E. (2007). Reader-based exploration of lexical cohesion. *Language Resources and Evaluation*, 41:27–44. 10.1007/s10579-007-9015-y.

Bellot, P. et El-Beze, M. (2001). Classification et segmentation de textes par arbres de décsion. application à la recherche documentaire. *Techniqueet science informatiques*, 20(1):107–134.

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

BENBRAHIM, M. et AHMAD, K. (1994). Computer-aided lexical cohesion analysis and text abridgement. Rapport technique, University of Surrey.

BERZLÁNOVICH, I., EGG, M. et REDEKER, G. (2008). Coherence structure and lexical cohesion in expository and persuasive texts. *In Workshop on Constraints in Discourse III*.

BESTGEN, Y. (2009). Quel indice pour mesurer l'efficacité en segmentation de texte ? *In Conférence sur le traitement automatique des langues naturelles (TALN)*.

BESTGEN, Y. et PIÉRARD, S. (2006). Comment évaluer les algorithmes de segmentation automatiques ? essai de construction d'un matériel de référence. *Actes de TALN : Verbum ex machina, Louvain-la-neuve*, 6:407–414.

BLOOMFIELD, L. (1933). *Language*. New York : Henry Holt.

BOGURAEV, B. K. et NEFF, M. S. (2000). Discourse segmentation in aid of document summarization. *In HICSS'00*.

BOOKSTEIN, A., KULYUKIN, V. A. et RAITA, T. (2002). Generalized hamming distance. *Inf. Retr.*, 5(4):353–375.

BOUAUD, J., HABERT, B., NAZARENKO, A. et ZWEIGENBAUM, P. (2000). Regroupements issus de dépendances syntaxiques sur un corpus de spécialité : catégorisation et confrontation à deux conceptualisations du domaine. *In* CHARLET, J., ZACKLAD, M., KASSEL, G. et BOURIGAULT, D., éditeurs : *Ingénierie des connaissances : évolutions récentes et nouveaux défis*, chapitre 17, pages 275–290. Eyrolles, Paris.

BOURIGAULT, D. (2002). UPERY : un outil d'analyse distributionnelle étendue pour la construction d'ontologies à partir de corpus. *In Actes de la 9e conférence sur le Traitement Automatique de la Langue Naturelle*, pages 75–84, Nancy.

BOURIGAULT, D. (2007). Un analyseur syntaxique opérationnel : Syntex. Habilitation à Diriger les Recherches, Laboratoire CLLE-ERSS (UMR5263), CNRS & Universit'e Toulouse-Le Mirail.

BOURIGAULT, D. et FABRE, C. (2000). Approche linguistique pour l'analyse syntaxique de corpus. *Cahiers de Grammaire*, 25(25):131–151.

BOURIGAULT, D. et GALY, E. (2005). Analyse distributionnelle de corpus de langue générale et synonymie. *In 4^{es} Journées de la linguistique de corpus*, pages 163–174, Lorient.

Bras, M. (2007). French adverb d'abord and discourse structure. *In* Aurnague, M., Korta, K. et Larrazabal, J., éditeurs : *Language, Representation and Reasoning. Memorial Volume to Isabel GomezTxurruka*, pages 77–102. Universidad del País Vasco.

Bras, M., Prévot, L. et Vergez-Couret, M. (2008). Quelles relations de discours pour les structures énumératives ? *In* Durand, J., Habert, B. et Laks, B., éditeurs : *Actes du Premier Congrès Mondial de Linguistique Française, CMLF'08*, pages 1945–1964, Paris.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32.

Brunn, M., Chali, Y. et Pinchak, C. J. (2001). Text summarization using lexical chains. *In Proceedings of the Document Understanding Conference (DUC 2001)*, pages 135–140, Nouvelle Orléans.

Budanitsky, A. et Hirst, G. (2006). Evaluating wordnet-based measures of lexical semantic relatedness. *Computational Linguistics*, 32(1):13–47.

Bullinaria, J. A. (2008). Semantic categorization using simple word co-occurrence statistics. *In Proceedings of ESSLLI Workshop on Distributional Lexical Semantics*.

Bullinaria, J. A. et Levy, J. P. (2006). Extracting semantic representations from word co-occurrence statistics : A computational study. *Behavior Research Methods*, 39:510–526.

Callan, J. P., Croft, W. B. et Harding, S. M. (1992). The Inquery retrieval system. *In Proceedings of the Third International Conference on Database and Expert Systems Applications*, pages 78–83.

Châar, S. L., Ferret, O. et Fluhr, C. (2004). Filtrage pour la construction de résumés multi-documents guidée par un profil. *Traitement Automatique des Langues (TAL)*, 45(1):65–93.

Charolles, M. (1983). Coherence as a principle in the interpretation of discourse. *Text*, 3:71–97.

Charolles, M. (1995). Cohésion, cohérence et pertinence du discours. *Travaux de Linguistique*, 24:125–151.

Charolles, M. (1997). L'encadrement du dicours - univers, champs, domaines et espace. *Cahier de Recherche Linguistique*, 6:1–60.

Chen, H., Branavan, S., Barzilay, R. et Karger, D. (2009). Content Modeling Using Latent Permutations. *Journal of Artificial Intelligence Research*, 36:129–163.

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.

Choi, F. Y. Y., Wiemer-Hastings, P. et Moore, J. (2001). Latent semantic analysis for text segmentation. *In Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing*, pages 109–117, Pittsburgh.

Church, K. et Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):pp. 22–29.

Cilibrasi, R. et Vitanyi, P. (2007). The google similarity distance. *Knowledge and Data Engineering, IEEE Transactions on*, 19(3):370 –383.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20 (1):37–46.

Collins, C., Penn, G. et Carpendale, S. (2008). Interactive visualization for computational linguistics. HLT tutorials.

Cornish, F. (1999). *Anaphora, discourse, and understanding : Evidence from English and French.* Clarendon Press (Oxford England and New York).

Couto, J., Ferret, O., Grau, B., Hernandez, N., Jackiewicz, A., Minel, J.-L. et Porhiel, S. (2004). RÉgal, un système pour la visualisation sélective de documents. *La présentation d'information sur mesure, Numéro Spécial de RIA*, pages 481–514. Paris, C. et Colineau, N. (éditeurs invités.

Cruse, D. A. (1986). *Lexical Semantics.* Cambridge University Press.

Curran, J. R. (2003). *From distributional to semantic similarity.* Thèse de doctorat, Institute for Communicating and Collaborative Systems School of Informatics University of Edinburgh.

de Beaugrande, R.-A. (1997). *New foundations for a science of text and discourse : Cognition, communication, and the freedom of access to knowledge and society.* Norwood, NJ : Ablex.

Dias, G., Alves, E. et Lopes, J. G. P. (2007). Topic segmentation algorithms for text summarization and passage retrieval : an exhaustive evaluation. *In Proceedings of the 22nd national conference on Artificial intelligence - Volume 2*, AAAI'07, pages 1334–1339. AAAI Press.

Dias, G., Moraliyski, R., Cordeiro, J. et Doucet, A. (2010). Automatic discovery of word semantic relations using paraphrase alignment and distributional lexical semantics analysis. *Natural Language Engineering*, 1 (1):1–30.

Diestel, R. (2010). *Graph Theory*. Springer.

Dubois, J. et Dubois-Charlier, F. (1970). Principes et méthode de l'analyse distributionnelle. *Langages*, 5(20):3–13.

Ducrot, O. (1972). *Dire et ne pas dire*. France : Hermann, Paris.

Eisenstein, J. et Barzilay, R. (2008). Bayesian unsupervised topic segmentation. *In EMNLP '08 : Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 334–343, Morristown, NJ, USA. Association for Computational Linguistics.

Evert, S. et Krenn, B. (2001). Methods for the qualitative evaluation of lexical association measures. *In In Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, pages 188–195.

Evert, S. et Krenn, B. (2005). Using small random samples for the manual evaluation of statistical association measures. *Computer Speech & Language*, 19(4):450–466.

Evert, S. et Lenci, A. (2009). Distributional semantic models : Theory and empirical results. *In Advanced course at ESSLLI'09*.

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*, Louvain.

Fabricius-Hansen, C. et Behrens, B. (2001). Elaboration and related discourse relations viewed from an interlingual perspective. *In Proceedings from Third Workshop on Text Structure*, Austin, Texas. version papier + version electronique.

Fang, Z. et Cox, B. (1998). Cohesive harmony and textual quality : An empirical investigation. *In National Reading Conference Yearbook*, volume 47, pages 345–353.

Fayol, M. (1997). On acquiring and using punctuation : A study of written french. *In* Costermans, J. et Fayol, M., éditeurs : *Processing Interclausal Relationships : Studies in the Production and Comprehension of Text*, pages 157–178. Lawrence Erlbaum Associates : Mahwah, New Jersey.

Fellbaum, C., éditeur (1998). *Wordnet : An Electronic Lexical Database (Language, Speech, And Communication)*. The MIT Press.

Ferret (2010). Similarité sémantique et extraction de synonymes à partir de corpus. *In Proceedings of TALN 2010*.

FERRET, O. (1998). *ANTHAPSI : un système d'analyse thématique et d'apprentissage de connaissances pragmatiques fondé sur l'amorçage.* Thèse de doctorat, Université de Paris Sud.

FERRET, O. (2002). Segmenter et structurer thématiquement des textes par l'utilisation conjointe de collocations et de la récurrence lexicale. *In TALN*, Nancy.

FERRET, O. (2006a). Approches endogène et exogène pour améliorer la segmentation automatique de documents. *Traitement Automatique des Langues (TAL)*, 47:111–135.

FERRET, O. (2006b). Découvrir les thèmes d'un document pour en améliorer la segmentation thématique. *In Conférence Internationale sur le Document Électronique (CIDE 9)*, pages 97–111.

FERRET, O. (2007). Finding document topics for improving topic segmentation. *In Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 480–487, Prague, Czech Republic. Association for Computational Linguistics.

FIRTH, J. R. (1957). Modes of meaning. *In Papers in Linguistics 1934-1951*, pages 190–215. London : Oxford University Press.

GALE, W., CHURCH, K. et YAROWSKY, D. (1992). One sense per discourse. *In In Proceedings of the 4th DARPA Speech and Natural Language Workshop, New-York*, pages 233–237.

GEORGESCUL, M., CLARK, A. et ARMSTRONG, S. (2006). An analysis of quantitative aspects in the evaluation of thematic segmentation algorithms. *In Proceedings of the 7th SIGdial Workshop on Discourse and Dialogue*, pages 144–151, Sydney, Australia. Association for Computational Linguistics.

GREFENSTETTE, G. (1994a). Corpus-derived first, second and third-order word affinities. *In Proceedings of Euralex*, pages 279–290.

GREFENSTETTE, G. (1994b). *Explorations in automatic thesaurus discovery.* Kluwer Academic Pub., Boston.

GRISHMAN, R. et KITTREDGE, R., éditeurs (1986). *Analyzing Language in Restricted Domains : Sublanguage Description and Processing.* New Jersey : Lawrence Erlbaum Associates.

GROSZ, B. J. et SIDNER, C. L. (1986). Attention, intentions and the structure of dicourse. *Computational Linguistics*, 12(3):175–204.

Guinaudeau, C., Gravier, G. et Sébillot, P. (2012). Enhancing lexical cohesion measure with confidence measures, semantic relations and language model interpolation for multimedia spoken content topic segmentation. *Comput. Speech Lang.*, 26(2):90–104.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P. et Witten, I. H. (2009). The weka data mining software : An update. *SIGKDD Explorations*, 11(1):..

Hall, M. A. (2011). Weka class InfoGainAttributeEval.

Halliday, M. (1985). *An Introduction to Functional Grammar*. Baltimore : Edward Arnold Press.

Halliday, M. et Hasan, R. (1976). *Cohesion in English*. Longman, London.

Halmoy, J.-O. (1982). *Le gérondif. Eléments pour une description syntaxique et sémantique*. Thèse de doctorat, University of Trondheim.

Halmøy, O. (2003). *Le gérondif en français*. Collection l'Essentiel Français. Ophrys.

Harabagiu, S. et Maiorano, S. (1999). Knowledge-lean coreference resolution and its relations to textual cohesion and coherence. *In Proceedings of the Association for Computational Linguistics Workshop on Discourse/Dialogue Structure and Reference*, pages 29–38, University of Maryland.

Harris, Z. (1951). *Structural Linguistics*. University of Chicago Press.

Harris, Z. (1962). *String Analysis of Language Structure*. Mouton and Co. The Hague.

Harris, Z. (1968). *Mathematical Structures of Language*. New-York : JohnWiley & Sons.

Harris, Z., Gottfried, M., Ryckman, T., Mattick, P. J., Daladier, A., Harris, T. N. et Harris, S. (1989). *The Form of Information in Science : Analysis of an Immunology Sublanguage*, volume 104 de *Boston Studies in the Philosophy of Science*. Kluwer Academic Publishers.

Harris, Z. S. (1954). Distributional structure. *Word*, 10:146–194.

Harris, Z. S. (1988). *Language and information*. Columbia University Press (New York).

Harris, Z. S. (1990). La genèse de l'analyse des transformations et de la métalangue. *Langages*, 25(99):9–20.

HARRIS, Z. S. (1991). *A theory of language and information : A mathematical approach*. Clarendon Press (Oxford England and New York).

HASAN, R. (1984). Coherence and cohesive harmony. *In* FLOOD, J., éditeur : *Understanding Reading Comprehension : Cognition, language, and the structure of prose*, pages 181–219. Newark, Delaware : International Reading Association.

HEARST, M. A. (1994). Multi-paragraph segmentation of expository text. *In ACL'94*, Las Cruces, NM.

HEARST, M. A. (1997). TextTiling : segmenting text into multi-paragraph subtopic passages. *Computational Linguistics*, 23(1):33–64.

HERNANDEZ, N. (2004). *Détection et Description Automatique de Structures de Texte*. Thèse de doctorat, Université de Paris-Sud XI.

HINDLE, D. (1990). Noun classification from predicate-argument structures. *In Proceedings of the 28th annual meeting of the Association for Computational Linguistics (ACL-1990)*, pages 268–275.

HIRSCHMAN, L. et GRISHMAN, Ralph & Sager, N. (1975). Grammatically-based automatic word class formation. *Information Processing and Management*, 11:39–57.

HIRST, G. (1997). Context as a spurious concept. *In AAAI Fall Symposium on Context in Knowledge Representation and Natural Language*, Cambridge, MA. Superseded by February 2000 version.

HIRST, G. et ST-ONGE, D. (1998). Lexical chains as representation of context for the detection and correction of malapropisms. *In* FELLBAUM, C., éditeur : *WordNet : An Electronic Lexical Database and Some of its Applications*. The MIT Press, Cambridge, MA.

HO-DAC, L.-M. (2007). *La position initiale dans l'organisation du discours : une exploration en corpus*. Thèse de doctorat, Université Toulouse le Mirail.

HO-DAC, L.-M., FABRE, C., PÉRY-WOODLEY, M.-P. et REBEYROLLE, J. (2009). Des indices aux marqueurs : méthodes de découverte de marqueurs discursifs complexes. *In Linguistic and Psycholinguistic Approaches to Text Structuring*, Paris.

HO-DAC, L.-M., PÉRY-WOODLEY, M.-P. et TANGUY, L. (2010). Anatomie des Structures Énumératives. *In Actes de la conférence TALN 2010*, page (publication numérique), Montréal, Canada.

HOBBS, J. R. (1985). On the coherence and structure of discourse. Rapport technique Report No. CSLI-85-37, Center for the Study of Language and Information, Stanford University.

Hoey, M. (1991). *Patterns of Lexis in Text.* Oxford University Press, Oxford.

Hoey, M. (1994). Patterns of lexis in narrative : a preliminary study. *In* Tanskmann, S. et Warwik, B., éditeurs : *Topics and Comments : Papers from the Discourse Project*, volume 13, pages 1–40. Anglicana Turkuensia.

Hoey, M. (1997). The discourse's disappearing (and reappearing) subject : an exploration of the extent of intertextual interference in the production of texts. *In* Simms, K., éditeur : *Language and the Subject*, pages 245–264. Amsterdam : Rodopi.

Hollingsworth, B. et Teufel, S. (2005). Human annotation of lexical chains : Coverage and agreement measures. *In Proceedings of the SIGIR-05 Workshop ELECTRA : Methodologies and Evaluation of Lexical Cohesion Techniques in Real-world Applications.*

Ion, R., Ceauşu, A. et Irimia, E. (2011). An expectation maximization algorithm for textual unit alignment. *In Proceedings of the 4th Workshop on Building and Using Comparable Corpora : Comparable Corpora and the Web*, BUCC '11, pages 128–135, Stroudsburg, PA, USA. Association for Computational Linguistics.

Ji, X. et Zha, H. (2003). Domain-independent text segmentation using anisotropic diffusion and dynamic programming. *In Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, SIGIR '03, pages 322–329, New York, NY, USA. ACM.

Jobbins, A. C. et Evett, L. J. (1998). Text segmentation using reiteration and collocation. *In Proceedings of the 17th international conference on Computational linguistics - Volume 1*, ACL-COLING '98, pages 614–618, Stroudsburg, PA, USA. Association for Computational Linguistics.

Károly, M. (2002). *Lexical Repetition in Text.* Peter Lang, Frankfurt am Main.

Klavans, J. L., McKeown, K., Kan, M.-Y. et Lee, S. (1998). Resources for evaluation of summarization techniques. *In* Rubio, Gallardo, C. et Tejada, éditeurs : *Proceedings of the First International Conference on Language Resources and Evaluation (LREC'98)*, Granada, Spain.

Knott, A. (1996). *A Data-Driven Methodology for Motivating a Set of Coherence Relations.* Thèse de doctorat, Department of Artificial Intelligence, University of Edinburgh.

Knott, A., Oberlander, J., O'Donnell, M. et Mellish, C. (2001). Beyond elaboration : the interaction of relations and focus in coherent text. *In* Sanders, T., Schilperoord, J. et Spooren, W., éditeurs : *Text representation : linguistic and psycholinguistic aspects*, pages 181–196. Amsterdam : Benjamins.

KOZIMA, H. (1993). Text segmentation based on similarity between words. *In Proceedings of the 31st annual meeting on Association for Computational Linguistics*, pages 286–288, Columbus.

KUCERA, H. et FRANCIS, W. N. (1967). *Computational Analysis of Present-Day American English.* Brown University Press.

LAKOFF, G. (1987). *Women, Fire and Dangerous Things : What Categories Reveal About the Mind.* Chicago : University of Chicago Press.

LANDAUER, T. K. et DUMAIS, S. T. (1997). A solution to plato's problem : The latent semantic analysis theory of acquisition, induction and representation of knowledge. *Psychological Review*, 104:211–240.

LANDAUER, T. K., FOLTZ, P. et LAHAM, D. (1998). An introduction to latent semantic analysis. *Discourse Processes*, pages 259–284.

LANDIS, J.R.and Koch, G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33 (1):159–174.

LASCARIDES, A. et ASHER, N. (1993). Temporal interpretation, discourse relations and commonsense entailment. *Linguistics and Philosophy*, 16(5):437–493. http ://www.utexas.edu/cola/depts/philosophy/faculty/asher/papers/lash.ps.

LEGALLOIS, D. (2003). La base lexicale de la cohésion des textes non narratifs. *Les cahiers du CRISCO*, 7:7–26.

LEGALLOIS, D. (2004). Cohésion lexicale et réseaux phrastiques dans la construction du texte expositif. *In L'Unité texte*, pages 171–201. S.P.e.D. Klinger (Ed.).

LEIBNIZ, G. W. (1704). *Table de définitions.* in L. Couturat ed. 1904, Opuscules et fragments indédits de Leibniz , Paris.

LENCI, A. (2008). Distributional semantics in linguistic and cognitive research. *Lenci A. (ed.), From context to meaning : Distributional models of the lexicon in linguistics and cognitive science, special issue of the Italian Journal of Linguistics*, 20/1:1–31.

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

LINDSEY, R., VEKSLER, V., GRINTSVAYG, A. et GRAY, W. (2007). Be wary of what your computer reads : the effects of corpus selection on measuring semantic relatedness. *In 8th International Conference of Cognitive Modeling, ICCM.*

LUC, C. (2000). *Représentation et composition des structures visuelles et rhétoriques du texte. Approche pour la génération de textes formatés.* Thèse de doctorat, Université Paul Sabatier – Toulouse III.

Luc, C. (2001). Une typologie des énumérations basée sur les structures rhétoriques et architecturales du texte. *In TALN*, Tours.

Lund, K. et Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical cooccurrence. *Behaviour Research Methods*, 28:203–208.

Lyons, J. (1977). *Semantics, Volume 2*. Cambridge University Press.

Malik, R., Subramaniam, L. V. et Kaushik, S. (2007). Automatically selecting answer templates to respond to customer emails. *In Proceedings of the 20th international joint conference on Artifical intelligence*, IJCAI'07, pages 1659–1664, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Malioutov, I. et Barzilay, R. (2006). Minimum cut model for spoken lecture segmentation. *In ACL-44 : Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 25–32, Morristown, NJ, USA. Association for Computational Linguistics.

Mann, W. et Thompson, S. (1988). Rhetorical structure theory : Toward a functional theory of text organization. *Text*, 8:243–281.

Mann, W. C. et Thompson, S. A. (1987). Rhetorical structure theory : a theory of text organisation. Rapport technique, Technical report ISI/RS-87-190, Information Sciences Intitute.

Marathe, M. et Hirst, G. (2010). Lexical chains using distributional measures of concept distance. *In* Gelbukh, A., éditeur : *Proceedings, 11th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2010) (Lecture Notes in Computer Science 6008, Springer-Verlag)*, pages 291–302, Iaşi, Romania.

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

Martin, J. R. (1992). *English Text. System and Structure*. Amsterdam : John Benjamins.

Masson, N. (1995). An automatic method for document structuring. *In Proceedings of the 18th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, Seattle, Washington, USA.

Mathet, Y. et Widlöcher, A. (2009). La plate-forme glozz : environnement d'annotation et d'exploration de corpus. *In Actes de la Conférence Traitement Automatique du Langage Naturel*, Senlis.

McCarthy, R. (1988). Some vocabulary patterns in conversation. *In* Carter, M. et McCarthy, R., éditeurs : *Vocabulary and Language Teaching*, pages 181–200. London and New-York : Longman.

Mihalcea, R., Corley, C. et Strapparava, C. (2006). Corpus-based and knowledge-based measures of text semantic similarity. *In Proceedings of the 21st national conference on Artificial intelligence, AAAI06*, volume 1, pages 775–780. AAAI Press.

Mihalcea, R. et Moldovan, D. (2001). A highly accurate bootstrapping algorithm for word sense disambiguattion. *International Journal of Artificial Intelligence Tools*, 10:5–22.

Moens, M.-F. et Busser, R. D. (2001). Generic topic segmentation of document texts. *In Proceedings of the 24th ACM SIGIR Annual International Conference on Research and Development in Information Retrieval*, pages 418–419, New York.

Moore, J. et Wiemer-Hastings, P. (2003). Discourse in computational linguistics and artificial intelligence. *In* Graesser, A., Gernsbacher, M. et Goldman, S., éditeurs : *Handbook of Discourse Processes*, pages 439–486. Erlbaum, Mahwah, NJ.

Morlane-Hondère, F. et Fabre, C. (2010). L'antonymie observée avec des méthodes de TAL : une relation à la fois syntagmatique et paradigmatique ? *In* Association pour le Traitement Automatique des Langues (ATALA), éditeur : *Actes de TALN 2010*, page 6, Montréal, Canada. Article court.

Morlane-Hondère, F. et Fabre, C. (2012). Le test de substituabilité à l'épreuve des corpus : utiliser l'analyse distributionnelle automatique pour l'étude des relations lexicales. *In Actes du Congrès Mondial de Linguistique Française (CMLF) 2012,*, Lyon, France. Á paraître.

Morris, J. (2007). *Reader's Perceptions of Lexical Cohesion and Lexical Semantic Relations in Text.* Thèse de doctorat, University of Toronto, Faculty of Information Studies.

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, J. et Hirst, G. (2004). Non-classical lexical semantic relations. *In Proceedings of the HLT Workshop on Computational Lexical Semantics*, pages 46–51, Boston.

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).

Mortureux, M.-F. (1993). Paradigmes désignationnels. *Semen*, 8:123–141.

Nazarenko, A., Zweigenbaum, P., Bouaud, J. et Habert, B. (1997). Corpus-based identification and refinement of semantic classes. *In Proceedings of the 1997 American Medical Informatics Association (AMIA)*, pages 585–589. AMIA.

Neelameghan, A. (2001). Lateral relationships in multicultural, multilingual databases in the spiritual and religious domains : The om information service. *In* Bean, C. et Green, R., éditeurs : *Relationships in the organization of knowledge*, pages 185–198. Norwell, Mass. : Kluwer Academic Publishers.

Neelameghan, A. et Rao, I. K. R. (1976). Non-hierarchical associative relationships : Their types and computer-generation of rt links. *In Library Science with a Slant toward Documentation*, volume 13, pages 24–42.

Nyyssönen, H. (1992). Lexis in discourse. *In Nordic Research on Text and Discourse. NORDTEXT Symposium 1990*, pages 73–80, Espoo, Finland.

Okumura, M. et Honda, T. (1994). Word sense disambiguation and text segmentation based on lexical cohesion. *In Proceedings of 15th International Conference on Computational Linguistics (COLING)*, pages 755–761, Kyoto, Japan.

Otero, P. G. (2009). Comparing different properties involved in word similarity extraction. *In Progress in Artificial Intelligence, 14th Portuguese Conference on Artificial Intelligence, EPIA 2009*, pages 634–645.

Pantel, P. et Lin, D. (2000). An unsupervised approach to prepositional phrase attachment using contextually similar words. *In Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, ACL '00, pages 101–108, Stroudsburg, PA, USA. Association for Computational Linguistics.

Parsons, G. (1996). The development of the concept of cohesive harmony. *In* Berry, M., Butler, C.and Fawcett, R. et G., H., éditeurs : *Meaning and form : Systemic functional interpretation*, pages 585–599. Norwood, N.J : Ablex Publishing Corporation.

Passonneau, R. J. et Litman, D. J. (1997). Discourse segmentation by human and automated means. *Computational Linguistics*, 23(1):103–139. http ://www.research.att.com/ diane/cl97.ps.

Peirsman, Y., Heylen, K. et Speelman, D. (2007). Finding semantically related words in dutch. co-occurrences versus syntactic contexts. *In In Proceedings of the CoSMO workshop*, Roskilde, Denmark.

Péry-Woodley, M.-P. et Scott, D. (2006). Discours et Document : traitements automatiques. Numéro thématique. *revue T.A.L.*, 47(2):7–19.

PEVZNER, L. et HEARST, M. A. (2002). A critique and improvement of an evaluation metric for text segmentation. *Computational Linguistics*, 28:1–19.

PHILLIPS, M. K. (1985). *Aspects of Text Structure : An Investigation of the Lexical Organization of Text*. Amsterdam : North-Holland.

POIBEAU, T. et MESSIANT, C. (2008). Do we still Need Gold Standards for Evaluation ? *In Proceedings of the Language Resource and Evaluation Conference*, pages –, Maroc.

POLANYI, L. (1988). A formal model of the structure of discourse. *Journal of Pragmatics*, 12:601–638.

PONTE, J. M. et CROFT, B. W. (1997). Text segmentation by topic. *In Proceedings of the first European Conference on research and advanced technology for digital libraries*. U.Mass. Computer Science Technical Report TR97-18.

PRINCE, V. et LABADIÉ, A. (2007). Text segmentation based on document understanding for information retrieval. *In* KEDAD, Z., LAMMARI, N., MÉTAIS, E., MEZIANE, F. et REZGUI, Y., éditeurs : *Natural Language Processing and Information Systems*, volume 4592 de *Lecture Notes in Computer Science*, pages 295–304. Springer Berlin / Heidelberg.

PRÉVOT, L., VIEU, L. et ASHER, N. (2009). Une formalisation plus précise pour une annotation moins confuse : la relation d'élaboration d'entité. *Journal of French Language Studies*, 19(2):207–228.

PÉRY-WOODLEY, M.-P. (2000). *Une pragmatique à fleur de texte : approche en corpus de l'organisation textuelle*. Mémoire présenté pour l'obtention d'une Habilitation à Diriger des Recherches.

R DEVELOPMENT CORE TEAM (2012). *R : A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

ROZE, C., DANLOS, L. et MULLER, P. (2010). LEXCONN : a French lexicon of discourse connectives. *In Proceedings of the 8th* MAD *Multidisciplinary Perspectives on Signalling Text Organisation*, pages 114–125, Moissac.

RYCKMAN, T. (1990). De la structure d'une langue aux structures de l'information dans le discours et dans les sous-langages scientifiques. *Langages*, 25(99):21–38.

SAGER, N. (1986). *Analyzing Language in Restricted Domains : Sublanguage Description and Processing*, chapitre Sublanguage : Linguistic phenomenon, computational tool, pages 1–19. New Jersey : Lawrence Erlbaum Associates, grishman, r. and kittredge, r. édition.

Sager, N., Friedman, C. et Lyman, M. (1987). *Medical Language Processing : Computer Management of Narrative Data*. Addison-Wesley, Reading, MA.

Sahlgren, M. (2006). *The Word-Space Model : Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces*. Thèse de doctorat, Stockholm University.

Salton, G., Singhal, A., Buckley, C. et Mitra, M. (1996). Automatic text decomposition using text segments and text themes. *In* Press, A., éditeur : *Proceedings of Hypertext'96*, pages 53–65, New York.

Salton, G., Yang, C. S. et Yu, C. T. (1975). A theory of term importance in automatic text analysis. *Journal of the American Society for Information Science*, 26(1):33–44.

Sardinha, T. B. (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.

Saussure, F. d. (1916). *Cours de linguistique générale*. Lausanne and Paris : Payot, ed. c. bally and a. sechehay édition.

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. *In Proceedings of the Conference on New Methods in Language Processing*, Manchester, UK.

Schütze, H. (1992). Dimensions of meaning. *In Proceedings of the 1992 ACM/IEEE conference on Supercomputing*, Supercomputing '92, pages 787–796, Los Alamitos, CA, USA. IEEE Computer Society Press.

Schütze, H. (1993). Word space. *In Advances in Neural Information Processing Systems 5*, pages 895–902. Morgan Kaufmann.

Schütze, H. (1998). Automatic word sense discrimination. *Computational Linguistics*, 24(1):97–123.

Scott, D. et Souza, C. D. (1990). Getting the message across in rst-based text generation. *In* Dale, R., Mellish, C. et Zock, M., éditeurs : *Current Research in Natural Language Generation*, pages 47–73. Academic Press, London.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, 27:379–423 and 623–656.

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.

SINCLAIR, J. (1966). Beginning the study of lexis. *In* BAZELL, C., CATFORD, J. C., HALIDAY, M. A. K. et ROBINS, R. H., éditeurs : *In Memory of J. R. Firth*, pages 410–430. London : Longman.

SITBON, L. et BELLOT, P. (2005). Segmentation thématique par chaînes lexicales pondérées. *In Actes de la 12ème Conférence Traitement Automatique du Langage Naturel (TALN'05)*, pages 505–510, Dourdan.

SPERBER, D. et WILSON, D. (1986). *Relevance*. Blackwell, London.

SPORLEDER, C., LINLIN, L. et PALMER, A. (2010). Cohesive links with literal and idiomatic expressions in discourse : An empirical and computational study. *In Proceedings of the 8th MAD (Multidisciplinary Approaches of Discourse) : Multidisciplinary Perspectives on Signalling Text Organisation*, pages 17–20. Moissac, France.

STOKES, N. (2003). Spoken and written news story segmentation using lexical chains. *In Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology : Proceedings of the HLT-NAACL 2003 student research workshop - Volume 3*, NAACLstudent '03, pages 49–54, Stroudsburg, PA, USA. Association for Computational Linguistics.

STOKES, N. (2004). *Applications of Lexical Cohesion Analysis in the Topic Detection and Tracking Domain*. Thèse de doctorat, Department of Computer Science, University College Dublin.

STOKES, N., CARTHY, J. et SMEATON, A. (2002). Segmenting broadcast news streams using lexical chains. *In STarting AI Researchers Symposium (STAIRS 2002)*, pages 145–154.

SUBBA, R. et DI EUGENIO, B. (2009). An effective discourse parser that uses rich linguistic information. *In Proceedings of Human Language Technologies : The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 566–574, Boulder, Colorado. Association for Computational Linguistics.

TANGUY, L. (2012). *Complexification des données et des techniques en linguistique : contributions du TAL aux solutions et aux problèmes*. Mémoire présenté pour l'obtention d'une Habilitation à Diriger des Recherches.

TANSKANEN, S.-K. (2006). *Collaborating towards coherence : lexical cohesion in English discourse*. John Benjamins, Amsterdam.

TERRA, E. et CLARKE, C. L. A. (2003). Frequency estimates for statistical word similarity measures. *In Proceedings of the 2003 Conference of the North American*

Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1, NAACL '03, pages 165–172, Stroudsburg, PA, USA. Association for Computational Linguistics.

Turco, G. et Coltier, D. (1988). Des agents doubles de l'organisation textuelle, les marqueurs d'intégration linéaire. *Pratiques*, 57:57–79.

Turney, P. D. (2008). A uniform approach to analogies, synonyms, antonyms, and associations. *In Proceedings of the 22nd International Conference on Computational Linguistics - Volume 1*, COLING '08, pages 905–912, Stroudsburg, PA, USA. Association for Computational Linguistics.

van der Plas, L. (2008). *Automatic Lexico-Semantic Acquisition for Question Answering.* Thèse de doctorat, University of Groningen.

Vergez-Couret, M. (2009). Un marqueur, *plus particulièrement* de la relation d'élaboration. *In Linguistic and Psycholinguistic Approaches to Text Structuring*, Paris.

Vergez-Couret, M. (2010). *Étude en corpus des réalisations linguistiques de la relation d'Élaboration.* Thèse de doctorat, Université de Toulouse Le-Mirail.

Vergez-Couret, M. et Adam, C. (2012). Signalling elaboration : Combining french gerund clauses with lexical cohesion cues. *Discours*, 10:online. Special issue on Multidisciplinary Approaches to Discourse.

Vossen, P. (1998). *EuroWordNet : a multilingual database with lexical semantic networks.* Kluwer Academic Publishers, Norwell, MA, USA.

Wandmacher, T., Ovchinnikova, E. et Alexandrov, T. (2008). Does latent semantic analysis reflect human associations ? *In Lexical Semantics Workshop at ESSLLI 2008*, Hamburg.

Weeds, J. et Weir, D. (2005). Co-occurrence retrieval : A flexible framework for lexical distributional similarity. *Comput. Linguist.*, 31(4):439–475.

Weeds, J. E. (2003). *Measures and Applications of Lexical Distributional Similarity.* Thèse de doctorat, University of Sussex.

Wellner, B. et Pustejovsky, J. (2007). Automatically identifying the arguments of discourse connectives. *In Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 92–101, Prague, Czech Republic. Association for Computational Linguistics.

Wells, III, W. (1986). Efficient synthesis of gaussian filters by cascaded uniform filters. *IEEE Trans. Pattern Anal. Mach. Intell.*, 8(2):234–239.

WIDLÖCHER, A. (2008). *Analyse macro-sémantique des structures rhétoriques du discours : Cadre théorique et modèle opératoire.* Thèse de doctorat, Université de Caen.

WITTEN, I. H., FRANK, E. et HALL, M. A. (2011). *Data Mining : Practical Machine Learning Tools and Techniques with JAVA Implementations (Third Edition).* Morgan Kaufmann.

YANG, D. et POWERS, D. M. W. (2006). Word sense disambiguation using lexical cohesion in the context. *In Proceedings of the COLING/ACL on Main conference poster sessions*, COLING-ACL '06, pages 929–936, Stroudsburg, PA, USA. Association for Computational Linguistics.

Annexe A

Exemples de résultats de traitements sur le texte « Albanie »

A.1 Cliques extraites

Pour le texte « Albanie », 4953 cliques maximales ont été trouvées. La figure A.1 présente un histogramme de la taille des cliques maximales. La taille maximales des

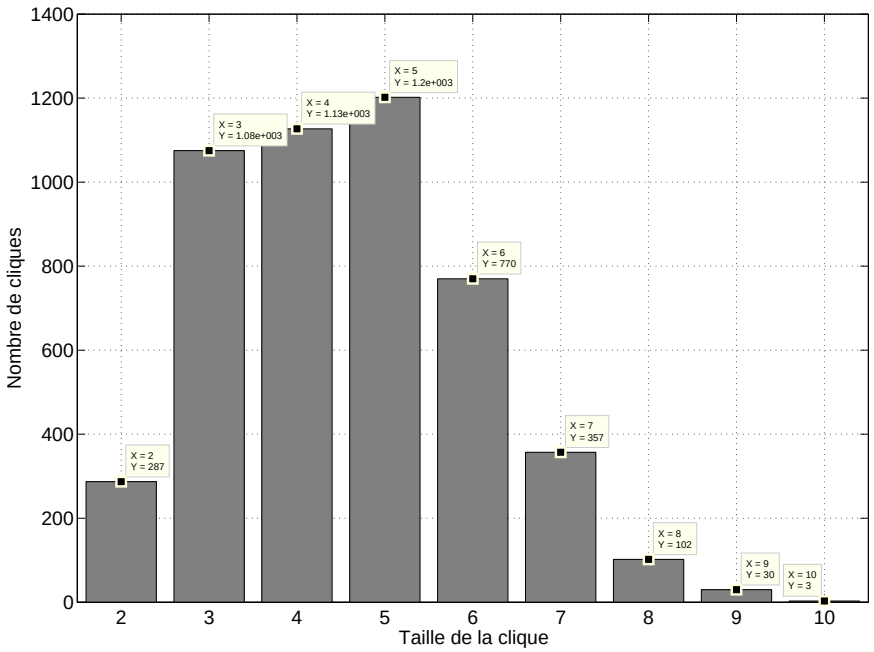


FIGURE A.1 – Histogramme de la taille des cliques maximales

cliques est de 10 tokens. On trouve assez peu de cliques de taille 2, car beaucoup

sont inclues dans des cliques plus longues. Suivent quelques exemples de cliques maximales de différentes tailles.

Clique de longueur 10 :
auteur_N, autorité_N, communauté_N, entreprise_N, famille_N, membre_N, nation_N, pape_N, peuple_N, soldat_N

Cliques de longueur 9 :
canal_N, continent_N, côte_N, lac_N, mont_N, montagne_N, plaine_N, plateau_N, terrain_N
association_N, auteur_N, chef_N, communauté_N, culture_N, entreprise_N, membre_N, nation_N, population_N
chiffre_N, consommation_N, durée_N, débit_N, longueur_N, masse_N, poids_N, pression_N, revenu_N

Cliques de longueur 8 :
charbon_N, denrée_N, hydrocarbure_N, minerai_N, pétrole_N, sel_N, viande_N, électricité_N
auteur_N, chanteur_N, journaliste_N, membre_N, ministre_N, pape_N, professeur_N, soldat_N
consommation_N, durée_N, flux_N, longueur_N, poids_N, pression_N, revenu_N, risque_N
assemblée_N, juge_N, maire_N, pape_N, préfet_N, président_N, secrétaire_N, sultan_N
destination_N, démographie_N, fleuve_N, géographie_N, montagne_N, nord-est_N, nord_N, république_N

Cliques de longueur 07 :
enfant_N, fille_N, mère_N, pape_N, parti_N, union_N, utilisateur_N
association_N, communauté_N, entreprise_N, gouvernement_N, musique_N, nation_N, province_N
altitude_N, consommation_N, durée_N, débit_N, envergure_N, longueur_N, poids_N
fille_N, juge_N, maire_N, mère_N, père_N, union_N, utilisateur_N
commune_N, lac_N, mont_N, montagne_N, plaine_N, plateau_N, terrain_N
conquête_N, destination_N, géographie_N, indépendance_N, langue_N, nord-est_N, peuple_N
adhésion_N, libération_N, mise_N, nomination_N, partage_N, rapprochement_N, transfert_N

Cliques de longueur 6 :
décembre_N, janvier_N, juillet_N, juin_N, mars_N, novembre_N

fille_N, mère_N, nation_N, parti_N, régime_N, tribu_N
déficit_N, effondrement_N, exode_N, inflation_N, mortalité_N,
pauvreté_N
centre_N, communauté_N, production_N, service_N, système_N, vie_N
homme_N, membre_N, pape_N, peuple_N, population_N, soldat_N
roi_N, royaume_N, régime_N, république_N, système_N, troupe_N
envahir_V, géographie_N, nord-est_N, nord_N, occupation_N,
territoire_N

A.2 Composantes connexes extraites

Cette annexe présente l'intégralité des composantes connexes de petite taille extraites du texte « Albanie » par la méthode détaillée dans la section 3.4.1.2.

violoncelliste_N violoniste_N
communiste_A conservateur_A socialiste_A républicain_A
annuel_A régulier_A publique_A
domaine_N secteur_N
échouer_V devenir_V
diviser_V démembrement_N
prise_N conquérir_V
officiel_A standard_A distinct_A indo-européen_A nomade_A
littéraire_A latin_A
catholique_A musulman_A
appellation_N ancien_N
législatif_A judiciaire_A administratif_A constitutionnel_A
union_N royaume_N
télévision_N radio_N
minerai_N pétrole_N charbon_N
préciser_V privilégier_V
seuil_N altitude_N
masse_N flux_N précipitation_N air_N
important_A divers_A propre_A différent_A principal_A
domination_N submerger_V
continental_A méditerranéen_A littoral_A côtier_A
éventail_N alternative_N
estimer_V diminuer_V
déjeuner_N repas_N
état_N secrétaire_N
byzantin_A ottoman_A

investissement_N vendre_V
besoin_N approvisionnement_N
humide_A froid_A sec_A pluvieux_A
montagneux_A fertile_A alluvial_A
capitalisme_N communisme_N
musique_N poète_N chanteur_N journaliste_N écrivain_N
accéder_V expliquer_V
âge_N garder_V période_N
entamer_V craindre_V tutelle_N
stalinien_A totalitaire_A
eau_N longueur_N rivière_N
reconnaître_V ajouter_V titre_N renommer_V devoir_V
fuite_N exode_N
piste_N aéroport_N voie_N livre_N route_N
noble_N bourgeoisie_N
dialecte_N parler_N
exportation_N importation_N
albanais_A serbe_A
migration_N migrer_V
pape_N prince_N ministre_N président_N
été_N hiver_N
long_A élever_V
orthodoxe_N sunnite_A chiite_A
sein_N interne_A
chasser_V reconquérir_V
longue_A long_N
milliard_N mètre_N
chroniqueur_N connaître_V géographe_N
énergétique_A alimentaire_A chimique_A mécanique_A
découpage_N concerner_V
parvenir_V réussir_V
extérieur_A intérieur_A
chômage_N mortalité_N fécondité_N
instaurer_V régler_V
italien_A actuel_A théorique_A
social-démocrate_A minoritaire_A
économique_A politique_A culturel_A social_A financier_A
inonder_V vivant_N dévaster_V concentrer_V priver_V
retard_N atteindre_V durée_N
islam_N christianisme_N
exister_V perdurer_V
trilogie_N reportage_N
ottoman_N turc_N

sel_N huile_N vinaigre_N
viande_N plat_N fruit_N
terrestre_A moyen_A fort_A faible_A maximal_A maximum_A bas_A
maire_N naître_V
situation_N troupe_N
vivant_A adulte_A fossile_A
janvier_N juin_N juillet_N novembre_N décembre_N