Caractérisation de la couverture d'information : Une approche computationnelle fondée sur les asymétries

par

Erick VELAZQUEZ-GODINEZ

THÈSE PRÉSENTÉE À L'ÉCOLE DE TECHNOLOGIE SUPÉRIEURE
COMME EXIGENCE PARTIELLE À L'OBTENTION
DU DOCTORAT EN GÉNIE
Ph. D.

MONTRÉAL, LE 26 JUIN 2017

ÉCOLE DE TECHNOLOGIE SUPÉRIEURE
UNIVERSITÉ DU QUÉBEC

PRÉSENTATION DU JURY

CETTE THÈSE A ÉTÉ ÉVALUÉE

PAR UN JURY COMPOSÉ DE:

Mme Sylvie Ratté, Directrice de Thèse
Département de génie logiciel et des TI à l'École de technologie supérieure

M. Mohamed Cheriet, Président du Jury
Département de génie de la production automatisé à l'École de technologie supérieure

M. Luc Duong, membre du jury
Département de génie logiciel et des TI à l'École de technologie supérieure

M. Jean-Guy Meunier, Examinateur Externe Indépendant
Département de philosophie à l'Université du Québec à Montréal

ELLE A FAIT L'OBJET D'UNE SOUTENANCE DEVANT JURY ET PUBLIC

LE 14 JUIN 2017

À L'ÉCOLE DE TECHNOLOGIE SUPÉRIEURE

REMERCIEMENTS

En premier lieu, je tiens à exprimer toute ma gratitude à ma directrice de thèse, Mme. *Sylvie* Ratté. Votre passion pour la linguistique a su piquer ma curiosité ; cela m'a amené à découvrir un nouveau monde. Je vous remercie pour votre appui sans faille et votre patience tout au long du processus de rédaction.

I also wish to thank Professor *Frank* de Jong for his insightful and valuable feedback throughout my research project. I conducted the first part of my research project in collaboration with his team from Ares Hogeschool Wageningen located in Wageningen, the Netherlands.

Pour la deuxième partie de ma recherche, j'ai collaboré avec M. *Pierre-André* Ménard, que je tiens aussi à remercier infiniment pour sa disponibilité.

Je remercie également M. *Christian* Desrosiers pour tous les conseilles qu'il apporté à cette recherche.

My colleagues at the LiNCS and the LiVE have played an important part throughout my research project ; thank you for reminding me that I was not alone. I feel incredibly grateful to all of them. Thank you *Laura* Hernandez, *Kuldeep* Kumar and *Faten* Mhiri for your great feedbacks. I also wish to thank *Alpa* Shah and *Remi* Martin for having proofread all my drafts. Thank you *Otilia* Alejandro, *Athefee* Manafi, *Mingli* Zhang, *Ruben* Dorado and *Ruth* Reategui for your valuable enthusiasm and support. I am especially grateful to *Edgar* Garcia Cano, who lent me his laptop after mine randomly crashed. Thank you guys for having been such an incredible and supportive team !

Par ailleurs, je tiens à remercier Mme. *Jocelyne* Caron pour la révision linguistique qu'elle a réalisée sur ce document. Je remercie également Mme. *Kathleen* Pineau et Mme. *Sylvie* Gervais pour leurs corrections et commentaires sur la section des résultats. J'exprime aussi toute ma gratitude à M. *David* Bertet pour ses commentaires concernant mon approche philosophique dans le chapitre de méthodologie.

Agradezco también a mi familia, mi tía *Lola*, mi *Mamá* y mis hermanas *Judith* y *Lupita*, por su comprensión y por su cariño.

Je veux aussi remercier mes amis *François* Tessier et *Jean* Goneau pour leur présence tout au long de ce processus ; pour votre soutien, merci beaucoup.

À *Justin* Garcia et *François* Hébert : merci m'avoir permis de me ressourcer quand j'en avais le plus besoin. Je remercie aussi la famille Hébert, qui m'ont accueilli dans leur maison lorsque j'ai commencé la rédaction de ma thèse. Je tiens aussi à remercier *Will* Buckwell pour avoir été mon camarade d'études ces derniers mois.

Enfin, je tiens à remercier le CONACyT pour son soutien financier lors des quatre premières années de mon doctorat. Je remercie également ma directrice de thèse pour le financement qu'elle m'a octroyé pour la dernière année de mon doctorat grâce à ses fonds de recherche.

CARACTÉRISATION DE LA COUVERTURE D'INFORMATION : UNE APPROCHE COMPUTATIONNELLE FONDÉE SUR LES ASYMÉTRIES

Erick VELAZQUEZ-GODINEZ

RÉSUMÉ

De nos jours, la production accélérée d'information demande à toute personne d'adopter des stratégies de sélection d'information, d'exclusion d'information répétée et même de fusion d'information, afin de construire un panorama complet d'une thématique. Ces stratégies correspondent bien au processus de couverture d'information qui devient un exercice de plus en plus quotidien, mais aussi de plus en plus complexe. Des techniques de *Traitement Automatique de Langue Naturelle* (TALN) tentent de réaliser la couverture d'information de façon automatique. Dans cette thèse, nous abordons la couverture d'information avec une approche computationnelle basée sur les asymétries. Nous avons appliqué notre analyse en deux scénarios différents :

Dans le premier scénario, nous avons analysé la couverture d'information dans les dissertations d'étudiants en vérifiant la présence des concepts qui proviennent des sources bibliographiques officielles telles que suggérées dans le syllabus du cours. Nous réalisons cette analyse à l'aide d'un coefficient de couverture qui utilise de l'information lexico-sémantique. Cette caractéristique hybride nous permet de capturer les différentes formes de surface lexicale qu'un étudiant peut utiliser pour exprimer un même concept. Pour déterminer si les concepts d'un livre sont couverts dans le contenu des dissertations, nous mettons en œuvre une stratégie d'alignement de texte. Notre approche est en mesure de détecter une dissertation avec un faible degré de couverture d'information parmi un groupe de dissertations qui ont une meilleure couverture. Pour corroborer les interprétations de nos résultats, nous avons conduit une évaluation qualitative avec les enseignants du cours. Cette évaluation a fait constater que les résultats de nos analyses coïncident avec les notes octroyées aux dissertations. Conséquemment, la couverture des concepts dans les dissertations d'étudiants permet d'expliquer la note qui est attribuée aux dissertations par les enseignants.

Dans le deuxième scénario, nous avons analysé la couverture d'information dans les textes journalistiques de type narratif. Dans ce type de texte, des événements, qui se produisent dans le monde, sont racontés et discutés par les journalistes. Les événements deviennent notre intérêt dans ce cas. Un événement présente une structure, celle-ci peut trouver sa forme dans les réponses des questions : qui a fait quoi ? À qui ? Où ? Et quand ? Afin de capturer le plus d'information concernant un événement, nous avons conçu un coefficient de couverture d'information basé sur des patrons linguistiques linéaires. Ces patrons, bien que simples, essaient de capturer la structure d'un événement. Nous avons aussi utilisé une stratégie de pondération des patrons afin de privilégier un patron en particulier. Nous abordons la couverture d'information, dans ce cas, avec une approche de détection de la nouvelle information, qui correspond à l'information non couverte par les autres sources. Dans l'évaluation quantitative, notre approche asymétrique est en mesure de performer aussi bien que les mesures symétriques de l'état de l'art. En plus,

notre approche offre l'avantage d'expliquer l'origine de la nouvelle information grâce à la stratégie de pondération des patrons.

Mots clés: Couverture d'information, Théorie d'asymétrie, Analytique d'aprentisage, Mesure de couverture

CHARACTERIZATION OF INFORMATION COVERAGE : A COMPUTATIONAL APPROACH BASED ON ASYMMETRIES

Erick VELAZQUEZ-GODINEZ

ABSTRACT

Nowadays, accelerated production of information requires people to adopt strategies to select information, to exclude repeated information and even to merge information, to build a complete panorama of a topic. These strategies fit well with the process of coverage of information, which is becoming an everyday task, but also a complex exercise. *Natural Language Processing* (NLP) techniques attempt to achieve automatically the coverage of information. In this thesis, we address the coverage of information with a computational approach based on asymmetries. We applied our analysis in two different scenarios :

In the first scenario, we analyzed the coverage of information in students' dissertations by verifying the presence of terminology from the official bibliographic references as suggested in the syllabus of the course. We performed this analysis using a hybrid asymmetric coverage coefficient that uses lexical and semantic information. This hybrid characteristic allows us to capture the different forms of lexical surface that a student can use to express the same concept. To determine if the concepts of a book are covered in the content of a dissertation, we implemented a text-alignment strategy. Our approach can detect a dissertation containing low degree of coverage of information among a group of dissertations that have a better coverage. To corroborate the interpretations of our results, we conducted a qualitative evaluation with the course's teachers. This evaluation revealed that the results of our analyzes coincided with the grades given to the dissertations. Consequently, the coverage of concepts in student dissertations helps to explain the grades that teachers attributed to the dissertations.

In the second scenario, we analyzed the coverage of information in narrative journalistic texts. In this type of texts, events, which occur in the world, are told and discussed by journalists. Events become our interest in this case. An event presents a structure, which can find its form in the answers to the questions : who did what ? To whom ? Where ? And when ? In order to capture the most information about an event, we designed an information coverage coefficient based on linear linguistic patterns. These patterns, although simple, try to capture the structure of an event. We also used a strategy of weighting patterns to highlight a particular pattern. We addressed the coverage of information, in this case, with a strategy of novelty detection, which corresponds to information not covered by other sources. In the quantitative evaluation, our asymmetric approach is able to perform as well as the symmetric measures of the state of the art. In addition, our approach offers the advantage of explaining the origin of the new information because of the strategy of weighting of the patterns.

Keywords: Information coverage, Asymmetry theory, Learning analytics, Coverage measure

TABLE DES MATIÈRES

Page

LISTE DES TABLEAUX

Page

LISTE DES FIGURES

LISTE DES ABRÉVIATIONS, SIGLES ET ACRONYMES

N.T.	Notre traduction.
TALN	Traitement Automatique des Langues Naturelles.
Ma-E	Macro-Événement.
Mi-E	Micro-Événement.
ML	Machine Learning.
LA	Analytique d'apprentissage, (*Learning Analytics*, en anglais).
S–R	Le sujet de comparaison est couvert par le référent.
R–S	Le référent est couvert par le sujet de comparaison.
TA	Théorie d'Asymétrie.
VSM	Modèle vectoriel, (*Vector space Model*, en anglais).
SVM	Machine à vecteurs de support (*support vector machine*, en anglais).
HA	Tête d'argument, (*head argument*, en anglais).
SimWN	Similarité WordNet.
RDF	*Resource Description Framework*.
LSA	Analyse sémantique latente (*Latent semantic analysis*, en anglais).
ACSO	Apprentissage collaboratif supporté par ordinateur, (*Computer-Supported Collaborative Learning*, en anglais.)
LDA	Allocation de Dirichlet latente (*latent Dirichlet allocation*, en anglais).
TREC	Text REtrieval Conference.
RI	Références individuelles.
RG	Références générales.
RS	Références spécialisées.
RC	Références considérées.
ACHM	*Asymmetric Coverage Hybrid Measure*.
PVM	Processus de vérification par les membres (*Member check* en anglais.)

INTRODUCTION

Dans cette thèse, nous nous intéressons au concept de *couverture d'information* en l'envisageant selon un prisme computationnel au sens de Ratté (1995) [1]. Le terme *couverture* doit être compris ici comme l'ensemble des informations propres à une thématique particulière. Toute personne confrontée à l'analyse de vastes quantités de données textuelles – journaliste, lecteur curieux, étudiant, etc. – doit développer des aptitudes pour sélectionner les sources, exclure l'information répétée, fusionner les détails afin de créer en quelque sorte mais avec plus ou moins de succès, le panorama complet de la thématique qui l'intéresse. Dans ce sens, nous présentons le processus de couverture comme le moyen de *« récupérer et synthétiser l'information pour lui donner un sens, pour l'interpréter »*, afin que l'utilisateur, quel qu'il soit, puisse appréhender et analyser l'information dans sa totalité, mais aussi dans ses différences subtiles. Naturellement, la première stratégie qui nous vient à l'esprit pour aborder la couverture d'information serait la comparaison du contenu de deux sources de texte. Notre hypothèse est qu'il est envisageable de considérer cette comparaison en adoptant un point de vue asymétrique. C'est dans cet esprit ou acception que doit être lue cette thèse.

Quand nous parlons d'information, de quelle information parlons-nous ? Nous envisageons ici tout type de communication écrite. Pour les besoins de cette thèse, nous limitons à deux type de contextes appartenant aux textes journalistiques et aux dissertations d'étudiants dans le cadre d'un cours universitaire. À première vue, les deux contextes semblent dépareillés. En effet, la nouvelle journalistique répond à des nécessités distinctes de celles imposées à des étudiants dans un cours. Cependant, les deux formes répondent à un besoin, celui de saisir les différences et les recouvrements entre un écrit original (que celui-ci soit imposé par la chronologie, ou par un modèle à suivre) et un écrit subséquent. Nous avons conçu les expériences en fonction de ces deux contextes distincts justement afin d'en démontrer le potentiel explicatif et interprétatif.

1. C'est-à-dire que nous présentons un mécanisme représentatif de la couverture d'information.

Quant au texte original, de quoi parle-t-on au juste ? Dans le contexte journalistique, le texte original peut se résumer au premier texte divulguant la nouvelle, mais ce n'est pas le seul point de vue pouvant être adopté, ni le plus complet. Ce premier texte, à partir duquel tous les autres seront comparés, peut aussi être celui que choisit un utilisateur précis, selon ses allégeances politiques ou sa position éditoriale (ref. Chapitre 1, section 1.5). Dans le contexte universitaire, bien sûr, lorsqu'on parle de recouvrement entre un texte original et un écrit subséquent produit par un étudiant, on pense immédiatement au plagiat. Ici, nous excluons cet aspect, puisque ce qui nous concerne ici est la mesure d'un certain apprentissage. Nous référons ici à cette capacité, durant sa formation, qu'acquiert un étudiant à bien couvrir et exprimer les concepts des écrits fondateurs de sa discipline (ref. Chapitre 1, section 1.4).

Dans cette thèse, nous ne prenons pas position sur la nature de ce premier écrit en vertu duquel tous les autres seront examinés (pour la démonstration, nous utiliserons, dans un cas, l'antériorité chronologique, et dans l'autre, les textes fondateurs de la discipline). Lorsque nous décrivons le texte original, nous en parlons comme étant le référent. Les textes subséquents, produits par d'autres nouvelles ou par des étudiants deviennent, ainsi, dans ce contexte précis, des sujets de comparaison. La relation entre un texte original et les textes subséquents (que ceux-ci soient déterminés selon un ordre chronologique, un intérêt, une politique, une imposition (p. ex. syllabus de cours) sont envisagés dans une relation de comparaison asymétrique, le premier texte étant, de par sa nature, celui à partir duquel tous les autres seront comparés, mais aussi auquel ils contribueront afin de produire, pour l'analyste, à la production d'un assemblage complet d'information.

On pourrait se demander pourquoi recourir à une comparaison non égale en soi entre deux communications écrites. De fait, l'hypothèse générale soutenant cette thèse s'appuie sur notre propre capacité comme être humain à faire des comparaisons. En effet, en science cognitive, la comparaison comporte deux éléments, un référent et un sujet de comparaison (Tversky, 1977;

Tversky & Gati, 1978). Le référent est l'objet qui a le plus de caractéristiques, et ces dernières devront être appariées à celles du sujet de comparaison. En ce sens, la direction de comparaison est déterminante pour établir le degré de similarité entre le référent et le sujet. Si les objets de comparaison changent de rôle, le degré de similarité sera aussi altéré. Tversky (1977) affirme que la comparaison est un processus cognitif asymétrique et qu'il devrait être envisagé comme un appariement de caractéristiques entre deux objets plutôt qu'un calcul de distance.

Adoptons maintenant le prisme informatique pour examiner cette comparaison entre un texte et les écrits subséquents. En traitement automatique des langues naturelles (TALN), la plupart des stratégies qui calculent la similarité textuelle sont basées sur des approches géométriques où le concept de distance est utilisé pour calculer la similarité entre des phrases ou d'autres unités de texte (des paragraphes, ou encore des documents complets). En raison de l'utilisation d'une fonction de distance, le référent et le sujet sont traités au même niveau. De plus, la plupart de ces approches utilisent très peu de connaissances linguistiques, de sorte que des propriétés fondamentales du langage (par ex. l'asymétrie sujet-prédicat) sont ignorées. Cela est essentiellement dû au fait que plusieurs des mesures proposées en TALN sont très souvent basées sur des représentations erronées des textes, car elles considèrent les mots comme s'il s'agissait de pixels dans une image. Ces représentations planes évacuent ainsi les configurations à l'intérieur desquelles apparaissent les mots, configurations qui devraient être prises en compte lors de tout calcul de similitud textuelle (Mihalcea *et al.*, 2006; Roth, 2014).

Le thème est tout de même vaste. En effet, il existe de multiples manières de comparer des textes, chacune utilisant des informations plus ou moins complexes à analyser (mots, syntaxe, style, etc.). Puisque notre thèse concerne le TALN, nous avons opté pour une méthode nous permettant de faire le pont entre la théorie linguistique, la théorie cognitive et l'analyse automatique, en tenant compte des contraintes technologiques. C'est donc une approche d'ingé-

nierie qui guide notre cheminement, et celle-ci contraint *ipso facto* les domaines dans lesquels s'inscrivent nos propositions et les données que nous devons mesurer.

Comme nous le mentionnions précédemment, deux domaines nous intéressent plus particulièrement dans le cadre de cette thèse : la couverture d'information dans le contexte de l'Analytique de l'apprentissage [2] (désormais LA) et la mesure de la couverture d'information dans le contexte journalistique.

Nous commençons notre analyse dans le contexte de la LA. Dans son ensemble, la LA vise à recueillir, mesurer et analyser automatiquement des données (textuelles ou non) sur les apprenants et leurs divers contextes d'apprentissage. Notre intervention ici concerne la couverture de concepts dans les dissertations d'étudiants en les comparant aux sources bibliographiques officielles telles que suggérées dans un syllabus de cours. Nous traitons ainsi les sources bibliographiques comme le référent et les documents écrits par les étudiants comme les sujets de comparaison.

Notre analyse est effectuée en utilisant une mesure de couverture asymétrique qui combine l'information sémantique et lexicale pour déterminer comment les concepts dans les références bibliographiques sont abordés dans les documents des étudiants. Pour déterminer si les concepts d'un livre sont couverts par les paragraphes dans le document d'un étudiant, nous mettons en œuvre une stratégie d'alignement de texte. Cette approche distingue les productions plus fortes des plus faibles en mesurant le degré de couverture des concepts apparaissant dans les sources citées dans le syllabus ou par les étudiants eux-mêmes.

Notre deuxième contexte constitue une application naturelle du concept de couverture. En effet, la notion de couverture fait partie des trois types de biais (Saez-Trumper *et al.*, 2013) lorsqu'on

2. Learning Analytics en anglais.

parle des nouvelles (D'Alessio & Allen, 2000; Park *et al.*, 2009). Les textes qui nous intéressent ici sont des textes journalistiques de type narratif. Nous excluons ainsi les textes d'opinion.

Dans un premier temps, on pourrait concevoir une mesure de la couverture comme étant la quantité d'information générée par un événement en particulier. Cette première définition nous oblige à mieux définir ce qu'on entend par «*événement* » d'une part, et par «*quantité* d'information», d'autre part. Pour la première notion, nous adoptons à la fois une définition micro, fondée sur la linguistique de la phrase et une définition macro, basée sur la structure générique d'une nouvelle. La définition d'un micro-événement se limite ainsi au domaine de la phrase et se base sur les éléments permettant de répondre aux questions : qui fait quoi ? à qui ? où ? et quand ? La définition d'un macro-événement reprend les mêmes questions, mais dans le contexte élargi de la situation qui a donné lieu à la nouvelle (les deux discussions sont présentées dans la section 1.5 du chapitre 1). Une nouvelle est ainsi envisagée comme une suite de micro-événements qui composent la trame du macro-événement sous-jacent. Puisque les réponses aux questions précédentes (qui fait quoi ? à qui ? où ? et quand ?) correspondent, pour chaque phrase, aux relations grammaticales, nous utilisons ces éléments pour construire un coefficient de couverture pondéré fondé sur des patrons linguistiques qui tentent de capturer certaines relations grammaticales dans chaque phrase (ref. section 3.3.4).

Déterminer si une nouvelle joue le rôle de référent ou de sujet de comparaison pour la source de nouvelles pose plusieurs questions : quels sont les critères pour déterminer ces rôles ? La quantité d'information ? La fiabilité de la source ? Le contenu ? La date de parution d'un texte ? Le choix du lecteur ? Il est clair que la réponse à chacune de ces questions appartient aussi au lecteur. Dans le cadre de cette recherche, nous avons fait le choix de la chronologie pour assigner le rôle de référent à une nouvelle. Ainsi, une première nouvelle racontant un événement sera traitée comme le référent, sans tenir compte de son origine éditoriale. Toute autre nouvelle générée par la suite sera traitée comme un sujet de comparaison.

– Présentation de la thèse

Dans les deux contextes, notre rôle consiste, dans un premier temps, à comprendre les données sur lesquelles seront appliquées des mesures. Conséquemment, le chapitre 1 a ainsi pour but de présenter un bref récapitulatif sur les fondements théoriques et historiques de cette thèse. D'abord, nous révisons l'ensemble de concepts liés à l'analyse des dissertations d'étudiants dans le contexte du LA, ensuite ceux qui sont reliés à l'analyse des textes journalistiques. Finalement, nous aborderons le concept de mesure et plus particulièrement, celui de symétrie sur lequel se fonde la grande majorité des mesures de similarité textuelle en TALN. Ce chapitre se termine par un positionnement théorique qui encadre notre recherche.

Le chapitre 2 présente la problématique et les objectifs de cette recherche.

Le chapitre 3 présente d'abord les postulats et les cadres interprétatifs de notre études, appuyés sur des fondements philosophiques. Par la suite, la méthodologie est divisée en deux sections servant à expliquer notre démarche pour chacune de nos propositions.

Le chapitre 4 est aussi divisé en deux grandes sections : la première présente les résultats de l'analyse sur les textes des étudiants, la seconde, celle des textes journalistiques.

La discussion et la comparaison de notre travail avec d'autres études similaires forment le contenu du chapitre 5.

Finalement dans dernière section, nous exposons la conclusion pour récapituler le travail de cette thèse.

CHAPITRE 1

REVUE DE LA LITTÉRATURE

Dans ce premier chapitre, nous présentons la revue de la littérature sur laquelle notre recherche s'appuie. Notre intérêt de recherche principal étant la couverture d'information, à la section 1.1 nous présentons un modèle qui caractérise les éléments de la couverture d'information. À partir de ce modèle, nous appliquons le concept de couverture d'information dans deux scénarios : La production de textes d'étudiants et la production de textes journalistiques.

La revue de la littérature concernant le premier scénario couvre les travaux en linguistique appliquée qui justifient le rôle de la langue dans le processus de l'apprentissage chez l'être humain. Nous présentons aussi certains travaux en la LA qui effectuent de manière automatique l'analyse des textes d'étudiants. L'ensemble de ces travaux sont présentés à la section 1.4

Essentiellement, un journal contient des nouvelles, qui présentent une structure narrative où des événements sont présentés de façon chronologique. Nous présentons la définition d'événement d'un point de vue linguistique à la section 1.5. Cette section nous permet de prendre position par rapport à la définition des événements et la façon dont nous les interprétons.

En raison de la nature informatique de cette recherche, nous présentons quelques modèles qui identifient les événements dans les textes de nouvelles. Puisque la partie fondamentale de notre recherche est le concept de comparaison, nous en discutons les aspects cognitifs, linguistiques et aussi technologiques à la section 1.2.

Nous terminons ce chapitre avec un résumé de problèmes et des critères qui circonscrivent notre contribution.

1.1 La couverture d'information

D'une façon très générique, la couverture est définie, par le Trésor de la langue française, comme « *Ce qui, matériellement, sert à couvrir, à recouvrir ou à envelopper quelqu'un ou*

quelque chose [1] ». Nous envisageons la couverture en soi comme un processus comportant trois composants (figure 1.1) . La première composante est l'objet de référence principal en A dans la figure 1.1 ; il se définit par un ensemble de traits distinctifs qui objectivement le caractérise et qui peut évoluer dans le temps. L'observateur (figure 1.1.B) forme le second composant. C'est par lui que les caractéristiques de l'objet de référence principal sont filtrées et mises en relief. La couverture est entièrement dépendante de cet observateur puisque ce n'est que par lui que se crée le texte qui constitue la projection en C (figure 1.1) de l'objet de référence principal. Cette projection forme donc le troisième composant du processus de couverture. On notera que le processus crée nécessairement une asymétrie entre l'objet de référence et la projection produite par l'observateur. Le filtre par lequel l'observateur analyse l'objet de référence représente donc les intérêts propres de cet observateur et conditionne ainsi le processus de projection.

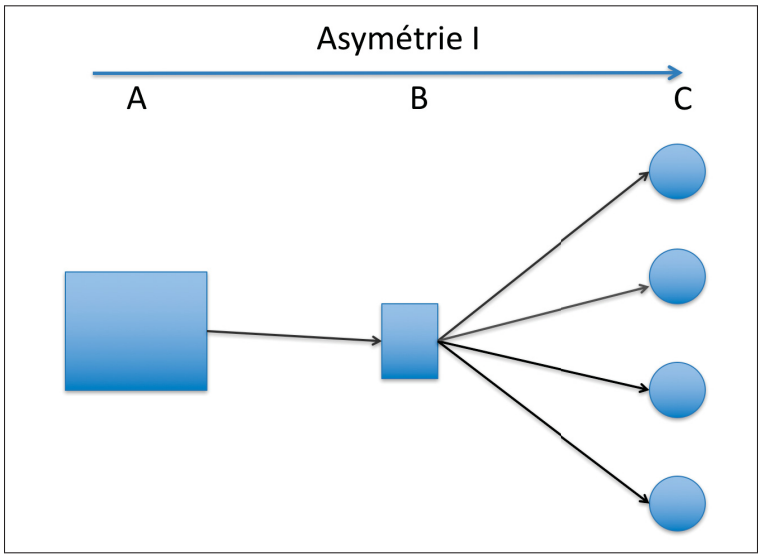


Figure 1.1 Diagramme général de la couverture d'information.

L'*observateur* (en B, fig.1.1) pourrait décider ce qu'il veut couvrir à tous les moments. Une projection dépend principalement des caractéristiques de la référence auxquelles il a accès et de l'observation qu'il fait. Cette projection pourrait refléter ses opinions, ses points de vue et les perceptions qu'il éprouve en observant la référence. La couverture est donc asymétrique,

1. Consultation faite en ligne : http://atilf.atilf.fr/ le 11 novembre 2016.

car l'observateur peut omettre délibérément dans ses projections certaines caractéristiques de l'objet de référence. L'observateur pourrait revenir, en tout temps, reprendre et vérifier l'état des caractéristiques de l'objet de référence pour les couvrir. Dans ce contexte, l'intérêt de l'observateur est prendre en compte les caractéristiques de l'objet de référence, les relations et les interactions qui existent entre elles pour les reporter dans une projection.

Finalement, la composante *projection*, fig 1.1-C, est la couverture "matérielle" de la référence faite par l'observateur. C'est la projection qui permet à un lecteur de connaître l'objet de référence. Cette connaissance est limitée à ce que l'observateur a transmis dans la projection.

Partant de cette définition générique, examinons maintenant comment ces concepts se matérialisent dans nos deux cas de figure : les productions écrites d'étudiants (figure 1.2) et les textes de nouvelles (figure 1.3).

1.1.1 Couverture des productions étudiantes

Nous nous plaçons maintenant dans le contexte où un étudiant doit lire un ensemble de livres ou d'articles scientifiques afin de produire un texte traitant d'une thématique spécifique (figure 1.2). Dans ce cas, l'objet de référence principal est constitué par la documentation que l'étudiant doit lire (figure 1.2.A). La thématique choisie par l'étudiant (qui devient ici l'observateur) impose à ce dernier un filtre par lequel il analysera la documentation (figure 1.2.B). L'étudiant réalise donc l'observation des documents en y cherchant des caractéristiques qui pourront correspondre à la thématique choisie afin de produire un texte qui constitue la projection en C (1.2). Il peut donc choisir d'éliminer certains documents, mais aussi d'augmenter cet ensemble avec des références supplémentaires qu'il considère importantes pour la thématique choisie.

Ainsi, dans ce contexte précis, l'objet de référence est formé par un ensemble de documents liés à une thématique spécifique. Cet ensemble se compose d'articles ou de livres suggérés par le professeur et auxquels l'étudiant-observateur pourra ajouter des documents complémentaires.

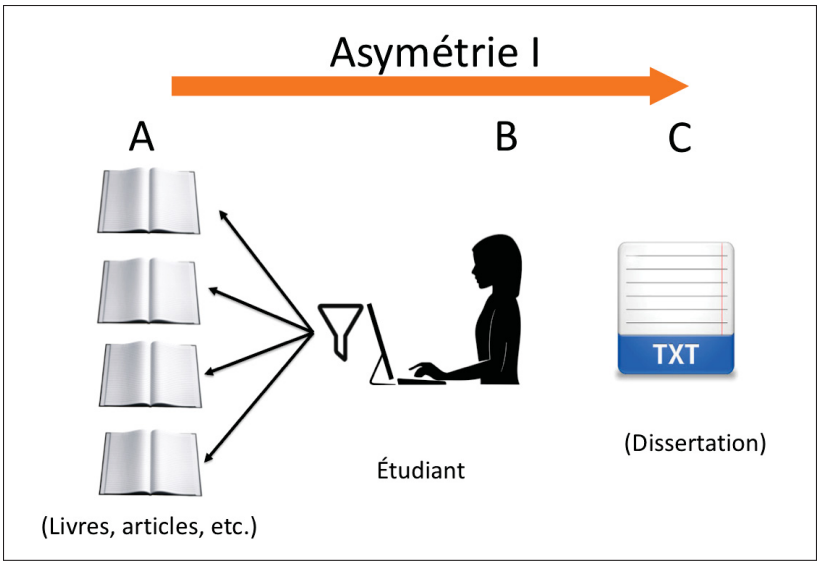


Figure 1.2 Diagramme de couverture pour plusieurs références et une seule projection.

1.1.2 Couverture des textes journalistiques

Dans le contexte des textes de nouvelles (figure 1.3), l'objet de référence est un événement dans le monde en A (figure 1.3) que les journalistes doivent couvrir dans un ou plusieurs textes. Chaque journaliste devient un observateur en B (figure 1.3) ; le biais [2] de ce journaliste (ses opinions, sa ligne éditoriale, etc.) devient ce filtre par lequel il analysera la situation. Chaque observateur-journaliste peut donc produire, pour le même événement, un ensemble de textes qui constituent autant de projections. Cet ensemble de projections en C (fig. 1.3) permet ainsi de couvrir l'objet de référence.

Dans ce contexte particulier, une deuxième asymétrie apparaît entre les projections elles-mêmes. En effet, la première projection devient en quelque sorte une projection de référence sur laquelle pourra s'appliquer une évaluation de la couverture en la comparant aux écrits subséquents. De manière complémentaire, chaque nouvelle projection couplée à l'évaluation de la couverture peut ainsi servir à compléter et à augmenter cette projection de référence et ainsi

2. Nous adoptons le terme bias pour se rapprocher au terme anglais.

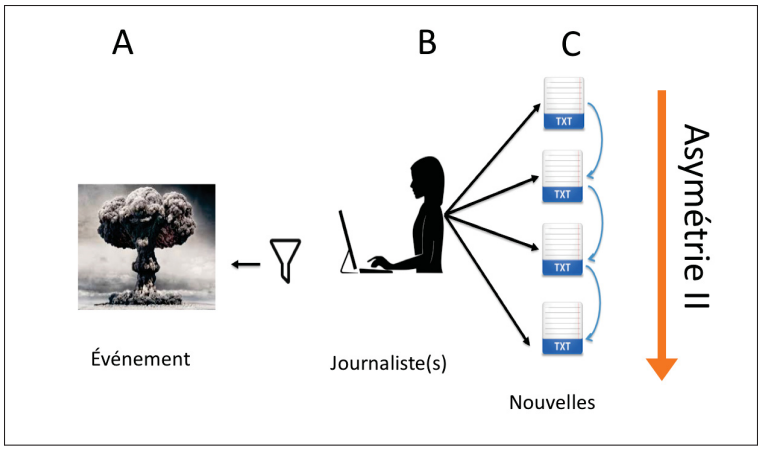


Figure 1.3 Diagramme de couverture pour une référence et plusieurs projections.

donner un portrait plus fidèle du véritable objet de référence dans le monde. La couverture ici implique l'analyse des multiples projections entre elles, en partant d'une projection spécifique.

1.2 La comparaison : la symétrie et l'asymétrie

En science cognitive, la reconnaissance, l'apprentissage et le jugement sont des exemples de processus mentaux cognitifs où les humains catégorisent des stimuli en termes de similarité. En ce sens, nous supposons que tout objet à comparer fait partie d'un ensemble d'objets qui partagent des caractéristiques communes. La similarité a souvent été abordée en philosophie et en psychologie comme une relation symétrique (Tversky, 1977). Cependant, Tversky (1977) et Tversky & Gati (1978) ont démontré que la similarité pour l'homme est, dans la plupart des cas, une relation asymétrique. Particulièrement, Tversky (1977) indique que la similarité asymétrique est observée dans les tâches de production, comme la reconnaissance des formes et l'association de mots.

L'asymétrie est également présente dans la relation entre les prédicats et les arguments puisque tout changement dans la structure du prédicat implique une interprétation différente de l'événement (Di Sciullo, 2013).

En informatique, les auteurs se sont intéressés à la similarité pour comparer des images, des mots, des phrases et des textes. Dans ce contexte, la similarité est toujours une mesure symétrique ; cela s'explique par l'utilisation de modèles géométriques qui dominent la plupart des approches. En TALN , un modèle géométrique fréquemment utilisé pour calculer la similarité des mots est l'analyse sémantique latente (LSA) ; celle-ci est basée sur sur un modèle vectoriel ou *Vector Space Model* (VSM) en anglais. Nous analysons de plus près les concepts de symétrie et d'asymétrie puisqu'ils sous-tendent l'approche que nous proposons.

1.2.1 La symétrie / l'asymétrie

Le concept de symétrie [3] a été employé depuis l'antiquité, principalement dans le domaine de l'architecture et de l'art (peinture, sculpture et aussi musique [4]). Les mathématiques modernes ont formalisé le concept de symétrie géométrique en termes d'un ensemble de transformations

3. Du mot grec συμμετρία, symmetria.
4. Pour plus d'informations sur l'histoire de l'utilisation de la symétrie et de l'asymétrie, voir l'annexe V.

géométriques possibles : la translation, la rotation, et la réflexion (Mitchell, 1990). Dans un plan cartésien, par exemple, ces opérations peuvent se réaliser à partir d'un point de repère, d'un axe ou encore sur l'arête d'une figure géométrique.

L'idée d'une symétrie naturelle [5] a aussi mené à son application pour décrire des processus cognitifs tels la comparaison. Pour ce faire, le concept géométrique de distance a dû être emprunté. La distance est utilisée pour déterminer le degré de similarité entre deux objets, A et B, qui ont été projetés dans un espace de coordonnées. Si les deux objets A et B sont près l'un de l'autre dans l'espace, ils sont envisagés comme deux objets similaires.

Par contre, si les objets A et B sont éloignés l'un de l'autre, ils seront envisagés comme étant deux objets différents qui ne partagent pas les mêmes caractéristiques. La distance est donc une fonction symétrique, car la distance entre les objets A et B est la même qu'entre B et A.

Cependant, Tversky (1977) postule que la similarité est une relation asymétrique et qu'elle est mieux décrite comme une correspondance (entre des ensembles de caractéristiques ou un processus d'appariement) plutôt qu'un calcul de distance entre deux points. Dans sa proposition, Tversky (1977) considère que chaque élément à comparer détient un rôle différent. C'est ainsi qu'il distingue le référent et le sujet de comparaison. Le référent est l'objet de comparaison qui détient les caractéristiques ou les stimuli les plus proéminents. Le choix de l'objet qui jouera le rôle de référent dépend de l'importance qui est attribuée aux caractéristiques de l'objet. Le sujet de comparaison est généralement l'objet ayant des caractéristiques moins proéminentes, (Tversky, 1977). Il existe donc une direction dans la comparaison qui dépend de la proéminence des caractéristiques des objets à comparer, (Tversky, 1977). Pour mieux comprendre la différence entre référent et sujet de comparaison, Tversky mentionne que les jugements de similarité peuvent être envisagés comme une extension d'énoncés en langue naturelle qui exprime la similarité, tel que : *A est comme B* ; où *A* est le sujet de comparaison et *B*, le référent. Dans ce sens, nous aurons aussi des énoncés en langue naturelle comme : *Le fils ressemble à son père*, ou *le portrait de Jean ressemble à Jean*. L'inversion de l'ordre dans ces énoncés

5. Cette notion de symétrie naturelle est basée sur la composition du corps humain présentée dans l'homme de Vitruve (Vitruvius, 2009) par Léonardo da Vinci.

ne nous semblerait pas naturelle. Le choix des énoncés en langue naturelle est associé avec la symétrie/asymétrie en jugement de similarité, (Tversky, 1977).

Formellement, Tversky (1977) définit la similarité de la façon suivante :

$$S(A,B) = F(A \cap B, A - B, B - A) \tag{1.1}$$

Où $F()$ est une fonction de similarité, $A \cap B$ représente les caractéristiques communes entre A et B. $A - B$ représente les caractéristiques qui appartiennent seulement à A. $B - A$ représente les caractéristiques qui appartiennent seulement à B. L'ensemble de ces relations de caractéristiques entre A et B déterminent leur valeur de similarité. Cette formalisation est illustrée sous la forme d'un diagramme à la figure 1.4.

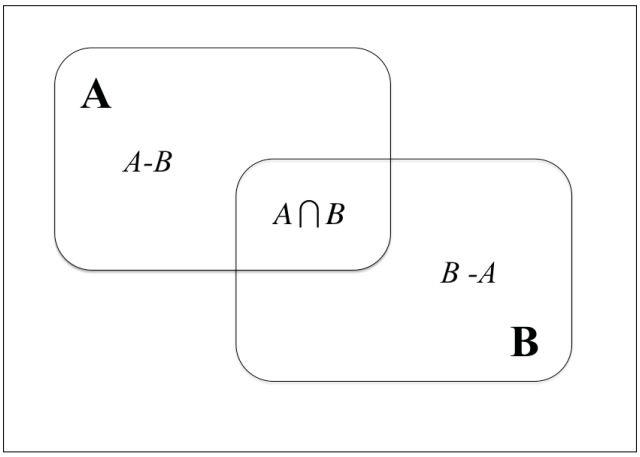


Figure 1.4 Diagramme de similarité de Tversky (1977).

Tversky (1977) indique également que la notion de similarité symétrique ne doit pas être rejetée complètement ; elle est valable dans de nombreux contextes, et dans beaucoup d'autres, il s'agit d'une approximation utile. Par ailleurs, il souligne que la similarité symétrique ne peut être acceptée comme un principe universel de similarité en psychologie. En outre, Tversky (1977)

montre que la notion de similarité asymétrique a été observée dans les tâches de comparaison où les gens comparent deux objets pour déterminer leur degré de similarité.

Leyton (1992) utilise également les concepts de symétrie et d'asymétrie dans sa théorie sur la perception et la cognition. Il présente la symétrie comme un élément nécessaire à toute activité cognitive quotidienne. Leyton (1992) a recours au problème de récupération du processus pour expliquer sa théorie, qu'il présente de la façon suivante : supposons qu'un individu observe un état, qui est appelé *moment présent*. Une certaine caractéristique structurelle de ce moment permet à la personne de reculer dans le temps et déduire les processus qui ont mené à ce moment présent. Le problème de récupération du procès représente donc les efforts qu'une personne doit faire pour récupérer les processus passés relativement à un moment de repère.

Comme solution à ce problème, Leyton (1992) présente deux principes :

- le principe de symétrie : Une symétrie dans le présent est comprise comme ayant existé depuis toujours. La symétrie est l'absence de processus-mémoire.
- le principe d'asymétrie : Une asymétrie dans le présent est interprétée comme provenant d'une symétrie passée. L'asymétrie est la mémoire qu'un processus laisse sur un objet.

Ainsi, Leyton (1992) considère la mémoire qu'un processus laisse sur un objet comme l'élément principal pour identifier la symétrie et l'asymétrie. Reprenons ici l'exemple de Leyton (1992) pour mieux le comprendre : Supposons qu'un réservoir de gaz reste stable dans une chambre et que le gaz ait atteint son équilibre dans un premier temps, en A (fig 1.5). Si l'on trace un axe vertical juste au milieu du réservoir, pour chaque position dans le réservoir, la concentration de gaz est équivalente. Maintenant, au temps 2, supposons que nous utilisons un aimant sur le côté gauche. Cet aimant entraîne le déplacement du gaz ainsi qu'une augmentation dans les particules du gaz du même côté du réservoir en B (1.5). La distribution du gaz est devenue asymétrique. Leyton (1992) mentionne que si une personne rentre dans la chambre, elle pourrait conclure qu'il y a eu un changement qui a provoqué une concentration de gaz sur le côté gauche de la chambre, cela même si la personne n'a pas vu le mouvement. Comme tel, l'asymétrie agit donc comme une mémoire du mouvement. Si dans un temps 3, en C (fig 1.5) le gaz atteint encore une fois l'équilibre dans le réservoir, et qu'une personne qui n'a pas

encore été dans la chambre y accède, elle ne pourrait pas dire que quelque chose s'est passée. Ainsi, le gaz revenu en état symétrique a effacé toute mémoire de l'événement passé. Alors la symétrie dans le présent ne permet pas de déduire une différence dans le passé, (Leyton, 1992).

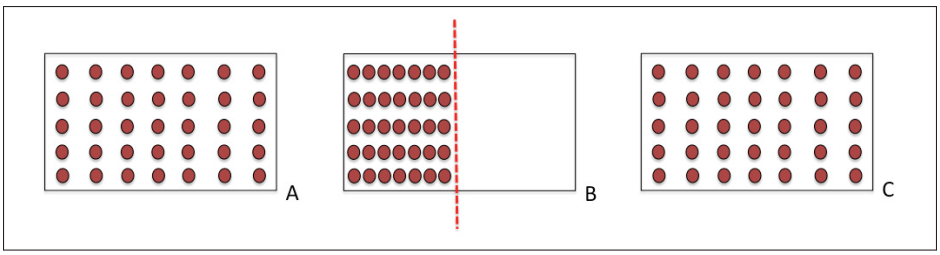


Figure 1.5 Reproduction du diagramme d'un réservoir de gaz en trois temps, A, B, et C. Example tiré de Leyton (1992, p. 8).

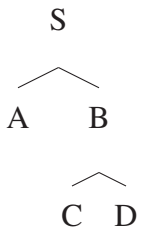
Leyton (1992) applique aussi sa théorie en linguistique et en art. En ce qui concerne la linguistique, Leyton (1992) affirme que les arguments en grammaire générative, qui justifient l'existence des opérations de mouvement dans les structures syntaxiques, sont basés sur les principes de symétrie et d'asymétrie. Le principe d'asymétrie agit en distinguant les capacités dans l'information positionnelle des syntagmes dans la structure hiérarchique des constituants. Le principe de symétrie, quant à lui, est instancié par le principe de projection. Le principe de projection dit que toute structure syntaxique doit provenir du lexique, dans le sens qu'elle réalise les exigences de la sous-catégorisation et des rôles dictés par le lexique. Toutes les propriétés lexicales sont maintenues, sans distinction, dans tous les niveaux syntaxiques (Leyton, 1992).

Dans les théories syntaxiques actuelles, qui sont basées sur la grammaire générative, les expressions linguistiques peuvent être représentées en termes de graphes orientés (Di Sciullo, 2013). En syntaxe, les concepts de symétrie et d'asymétrie sont identifiés dans les relations structurelles de la phrase. Ces relations peuvent être la préséance, la domination et la *C-command*[6] (Carnie, 2015). La *C-command* est peut-être l'une des relations structurelles de la phrase les plus importantes (Carnie, 2015). Les concepts de domination et de C-command furent origina-

6. C signifie constituant.

lement présentés par Reinhart (1976). Un nœud *C-commande* [7] ses sœurs et toutes les filles et les petites-filles de ses sœurs. Nous aurons deux types de *C-command* :

- *C-command* symétrique : Relation entre deux nœuds soeurs. Un nœud *A C-commande* symétriquement *B* si *A C-commande B* et *B C-commande A*, (Reinhart, 1976; Carnie, 2015). L'exemple de l'arbre syntaxique ci-dessous montre une relation de C-command symétrique entre les nœuds A et B.
- *C-command* asymétrique : Relation entre un nœud tante et ses nièces et les descendantes de celle-ci. Le nœud *A C-commande* asymétriquement *B* si *A C-commande B* mais *B* ne *C-commande* pas *A*. (Reinhart, 1976; Kayne, 1994; Carnie, 2015). Cette relations peut être observée dans l'arbre syntaxique ci-dessous où A C-commade asymétriquement les nœuds C et D, ces derniers ne *C-commandent* pas le nœud A.



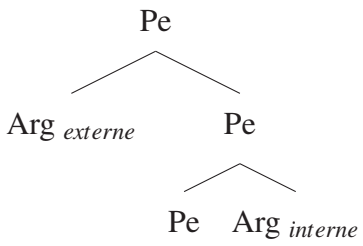
Le concept d'asymétrie a été largement discuté en linguistique, par exemple, dans la Théorie d'Asymétrie (TA), originalement proposée en morphologie par Di Sciullo (2005). La TA tient compte du fait qu'un changement des relations asymétriques dans un objet morphologique provoque soit un charabia ou une interprétation sémantique différente, (Di Sciullo, 2005). Par exemple, prenons le mot *prototypical*, en anglais : si nous séparons les morphèmes de ce mot : proto-typic-al ; en altérant leur ordre on obtient *al-typic-proto, *typic-al-proto, etc, exemple tiré de Di Sciullo (2005). Les relations syntaxiques sont aussi asymétriques ; une inversion des constituants n'entraîne pas un «charabia [8] », mais l'altération des relations sémantiques et de l'information, Di Sciullo (2013). Par exemple : *John killed Mary* aura un sens différent de *Mary killed John*. À propos de la sélection des arguments du prédicat, (Di Sciullo, 2013) mentionne :

7. Nous ajoutons un C majuscule comme préfixe au verbe commander pour indiquer le même sens que *C-command en anglais.*

8. Nous référons au terme anglais *gibberish.*

> « *Argument structure relations are asymmetric in the sense that a predicate asymmetrically selects an argument, whereas the inverse relation does not hold : an argument does not asymmetrically select a predicate*[9]. »

L'arbre syntaxique suivant[10] présente la structure d'un prédicat dyadique avec son argument externe (habituellement appelé sujet) et son argument interne (habituellement appelé objet) :



Nous constatons la relation asymétrique sur cet arbre en appliquant les concepts de *C-command* asymétrique. Selon Di Sciullo (2013), la structure des prédicats dénote des événements et le noyau d'un événement peut être modifié par des adjoints, c'est à dire par d'autres arguments internes au prédicat, par exemple des adjoints de localisation spatiale ou temporelle. Ainsi, l'asymétrie est également une propriété de la localisation spatiale d'un événement (Di Sciullo, 2013). La TA prédit correctement qu'il devrait exister une asymétrie entre le point d'origine d'un événement et son point final (Di Sciullo, 2013).

Chomsky (2005) propose que le développement du langage chez un individu est déterminé par la génétique et l'expérience ainsi que par des principes d'efficacité computationnelle. Di Sciullo (2016) présente les prémisses de la présence de l'asymétrie dans la faculté du langage en discutant deux principes d'efficacité computationnelle : *minimize symmetrical relations* et *minimize externalization*.

En particulier, le principe *minimize symmetrical relations* s'applique aussitôt que possible dans les dérivations syntaxiques et élimine les relations symétriques (Di Sciullo, 2016). Ce qu'il est nécessaire de comprendre ici, c'est que l'application du principe *minimize symmetrical rela-*

9. N.T. La relation de la structure argumentale est donc asymétrique dans le sens où un prédicat sélectionne un argument de façon asymétrique tandis que l'inverse ne peut pas avoir lieu. Un argument ne sélectionne asymétriquement pas son prédicat.

10. L'arbre syntaxique et l'exemple sont extraits de Di Sciullo (2013).

tions a pour effet d'éliminer les relations symétriques en déplaçant un ou plusieurs constituants. D'après la théorie de Leyton (1992), l'application de ce principe correspond ainsi au concept de mémoire, qui se trouve ici instancié par la trace des déplacements et agit comme une récupération du processus. On rejoint ainsi le principe asymétrique de Leyton, puisqu'on agit sur une symétrie «du passé» pour générer une asymétrie de la structure «dans le présent».

Le domaine des asymétries est vaste et encore plus complexe que ce que nous venons de présenter. Nous nous contentons de montrer que l'asymétrie est présente dans la configuration du langage, ce qui sert le propos de cette recherche.

En TALN, la comparaison de textes se fait aussi, la plupart du temps, avec des approches symétriques (Mihalcea *et al.*, 2006; Roth, 2014; Ferreira *et al.*, 2016). Ceci s'explique par l'utilisation de modèles d'espaces géométriques, comme *Vector Space Model*, et le modèle *bag-of-words* pour calculer la similarité cosinus. Cette approche exige la conversion du texte sous la forme d'un vecteur afin de calculer l'angle cosinus entre les vecteurs, ce qui, plus tard, sera interprété comme une similarité. Nous aborderons plus en détails cette mesure à la section 1.3.1. Si l'objectif d'une comparaison est de déterminer le degré de similarité entre deux textes, il importe donc de considérer les propriétés du langage. Pour les raisons exposées précédemment, l'asymétrie est une propriété structurelle du langage, une approche de comparaison asymétrique s'impose donc, correspondant d'avantage à la réalité du langage qu'une approche symétrique.

Les concepts de symétrie et d'asymétrie sont également utilisés dans la conception de techniques d'interaction humain-machine [11]. Dans ce contexte particulier, nous trouvons des manipulations symétriques et asymétriques sur les objets dans une interface. La manipulation symétrique se produit quand la main dominante et la main non dominante [12] partagent le même

11. Nous faisons référence ici aux techniques de manipulation sur des interfaces tactiles, soit un écran tactile ou un téléphone intelligent

12. Le terme de main dominante ou non dominante est utilisé pour désigner la préférence d'un utilisateur à se servir d'une main en particulier pour réaliser une tâche, soit la main droite ou la main gauche. La main dominante est celle qui réalise les détails les plus fins lors d'une tâche. Par conséquent, la fonction de la main non dominante est de fournir un support supplémentaire pour que la main dominante puisse réaliser la tâche. Pensons quand nous écrivons une lettre sur papier : la main dominante tient le stylo et la main non dominante tient le cahier pour que celui-ci ne bouge pas.

espace de travail dans un espace de temps dit coordonné. Le mouvement de zoom et de rotation d'une image est l"exemple le plus commun d'une manipulation symétrique (Velazquez Godinez, 2012). La manipulation asymétrique, moins répandue en techniques d'interaction, est la plus naturelle pour la réalisation des tâches comme l'écriture, ou le dessin à main levée (Velazquez Godinez, 2012) pour une discussion plus approfondie.

1.3 Les mesures de similarité textuelle

Le calcul de similarité entre les phrases d'un texte ou de plusieurs textes constitue une étape nécessaire en analyse automatique telle l'alignement de textes (Barzilay & Elhadad, 2003; Nelken & Shieber, 2006) ou la construction de résumés (Erkan & Radev, 2004).La similarité de phrases, est en fait, un processus qui combine des méthodes de similarité de mots pour exprimer la similarité entre deux segments de texte. Les mesures de similarité de phrases sont généralement classifiées en deux groupes : les mesures à base lexicale qui utilisent l'incidence des mêmes lemmes ou des mêmes formes de surface pour calculer la similarité, et les mesures à base de connaissances linguistiques, qui utilisent des relations sémantiques ou des relations syntaxiques pour le calcul de la similarité.

1.3.1 Mesures à base lexicale

Nous entendons par mesure à base lexicale toute mesure qui utilise uniquement la surface lexicale (mots) pour le calcul de similarité entre deux textes. Chaque mesure pourrait utiliser de façon différente l'information lexicale ; dans quelques cas, nous trouverons des mesures qui se contentent de vérifier si les mots dans les deux segments de textes sont identiques. Ceci sera utilisé pour exprimer une valeur qui représentera la similarité. Cette valeur varie généralement entre 0 et 1.

Dans ce groupe, l'approche la plus courante consiste à mesurer le cosinus de l'angle entre deux vecteurs. La similarité cosinus entre deux vecteurs est définie de la façon suivante :

$$cos(\overrightarrow{x}, \overrightarrow{y}) = \frac{\overrightarrow{x} \cdot \overrightarrow{y}}{\|\overrightarrow{x}\|\|\overrightarrow{y}\|} \tag{1.2}$$

Dans la formule 1.2, $\overrightarrow{x}$ et $\overrightarrow{y}$ sont la représentation vectorielle de deux segments de textes. La mesure a des propriétés symétriques, puisque l'ordre des vecteurs n'altère pas le résultat. De plus, nous pouvons interpréter le cosinus comme un coefficient de corrélation normalisé, ce qui est dénoté par les éléments $\|\overrightarrow{x}\|$ et $\|\overrightarrow{y}\|$, voir (Manning & Hirich, 1999, p. 301).

Une autre propriété intéressante de la similarité cosinus est que, si elle est appliquée sur des vecteurs normalisés, elle se comporte comme le fait la distance euclidienne. Cela permet de savoir si deux objets sont près d'un troisième (Manning & Hirich, 1999). Cette mesure de similarité a connu plusieurs améliorations. Par exemple,

- l'utilisation de n-grammes pour construire les vecteurs,
- la suppression des mots vides (*stop words*, en anglais),
- la suppression de quelques mots de contenu (adjectifs ou adverbes),
- la lemmatisation des mots,
- la désuffixation (stemming) des mots, etc.

Le développement récent des algorithmes de groupement spectral a contribué à la prolifération des mesures symétriques. En effet, ces algorithmes exigent que la matrice résultante sur laquelle s'appliquera les groupements soit obligatoirement symétrique (Von Luxburg, 2007).

Certaines mesures asymétriques ont cependant été proposées dans le contexte de la détection du plagiat. Puisque la tâche vise à identifier les extraits identiques entre une source plagiée et une production donnée, la notion d'asymétrie est obligatoire. La mesure proposée par Shivakumar & Garcia-Molina (1995) constitue un cas de figure. En effet, les auteurs proposent un modèle de fréquences relatives (Relative Frequency Model), fondé sur une représentation de type VSM et l'utilisation de la similarité cosinus. Le modèle qu'ils proposent est ainsi capable

de détecter des sous-ensembles similaires dans les documents ; c'est aussi la première mesure asymétrique reportée dans la littérature.

$$subset(D_1, D_2) = \frac{\sum_{w_i \in c(D_1, D_2)} \alpha_i^2 * F_i(D_1) \cdot F_i(D_2)}{\sum_{i=1}^{N} \alpha_i^2 F_i^2(D_1)} \tag{1.3}$$

Où $F_i(D_1)$ et $F_i(D_2)$ sont les vecteurs de fréquences des documents D_1 et D_2. Le paramètre α est un vecteur de poids, chaque α_i étant associé à l'occurrence d'un *i*ème terme particulier. Cette formule s'apparente à celle qui définit la similarité cosinus ; le numérateur est également formé par le produit scalaire des deux vecteurs. La principale motivation pour Shivakumar & Garcia-Molina (1995) de modifier la formule originale du cosinus est le fait qu'elle considère la longueur de deux vecteurs dans le calcul, ce qui génère des valeurs faibles même si un document peut être un sous-ensemble d'un autre document [13]. La solution des auteurs consiste à considérer la normalisation seulement par le document D_1. Puisque le sens de la comparaison peut altérer le résultat, les auteurs sélectionnent la valeur maximale entre deux documents en alternant leur rôle :

$$sim = (R, S) = max\{subset(R, S) subset(S, R)\} \tag{1.4}$$

Shivakumar & Garcia-Molina (1995) conçoivent la similarité de documents comme un chevauchement entre les documents et non comme une similarité sémantique. De plus, Shivakumar & Garcia-Molina (1995) utilisent une approche basée sur VSM, ce qui signifie que les termes d'un document et l'ordre des mots dans les phrases ne sont pas pris en compte.

Toujours dans le contexte de la détection du plagiat, Brin *et al.* (1995) et Bao *et al.* (2003) présentent deux mesures de similarité asymétriques pour exprimer l'inclusion des sous-ensembles de documents. Pour améliorer la performance de leurs mesures, ils utilisent un modèle qui considère seulement les mots avec une haute fréquence dans le document. Ils appellent ce mo-

13. Cette propriété serait présente dans toutes les mesures symétriques.

dèle *Heavy Frequency Vector*. Ce modèle leur permet de proposer deux approches différentes. La première approche est appelée *Inclusion Proportion Model* :

$$Incl(A,B) = \frac{\mid F(A) \subset F(B) \mid}{\mid F(A) \mid} = \frac{\sum_{i,j=1}^{n} \alpha_i (F_i(A) \oplus F_j(B))}{2 \times \sum_{i=1}^{n} \alpha_i F_i(A)} \tag{1.5}$$

Où $F(A)$ et $F(B)$ sont les vecteurs des fréquences de mots de deux documents. $F_i(A)$ correspond au nombre d'occurrences du mot $i$ème dans A ; $F_j(B)$ correspond au nombre d'occurrences du mot $j$ème dans B. Le symbol $\oplus$ signifie une somme directe, qui est conditionnée à $w_i = w_j$. Si cette condition n'est pas satisfaite l'opération $F_i(A) \oplus F_j(B) = 0$. Ceci assure la continuité des mots formant un ensemble similaire entre deux documents. α est un vecteur de poids de mots. Finalement la valeur de la similarité est donnée par la fonction suivante :

$$S_{IPM}(A,B) = min\{1, max\{Incl(A,B), Incl(B,A)\}\} \tag{1.6}$$

La seconde proposition est celle de Bao *et al.* (2003) ; elle correspond au modèle qu'eux-mêmes appellent *Heavy Frequency Model*.

$$Subset_{HFM}(A,B) = \frac{\sum_{i,j=1}^{n} \{[F_i(A) \otimes F_j(B)] \times [1 - \mid P_i(A) - P_j(B) \mid]\}}{\sum_{i=1}^{n} F_i^2(A)} \tag{1.7}$$

Où $P(D)$ est le vecteur de proportion respective. $P(D) = F(D) / \mid D \mid$, à savoir que c'est un *Heavy Frequency Vector*. $\mid D \mid$ est la quantité de mots. L'opérateur $\otimes$ est un produit tensoriel conditionné à $w_i = w_j$. Le cas contraire, $F_i(A) \otimes F_j(B) = 0$. La valeur de similarité finale est encore donnée par une fonction de sélection :

$$S_{HFM}(A,B) = min\{1, max\{Subset_{HFM}(A,B), Subset_{HFM}(B,A)\}\} \tag{1.8}$$

Ces deux mesures, 1.5 et 1.7, sont conçues pour détecter le plagiat : elles se contentent d'exprimer le degré de chevauchement entre un document D_1 et un autre document D_2. Il est clair que pour la détection de plagiat, l'ordre syntaxique est moins important et des intérêts de similarité sémantique sont aussi moins présents dans ce contexte. Les trois mesures conçues pour la détection de plagiat mettent en œuvre une stratégie de similarité lexicale et les résultats sont bons. Ceci est dû à la nature propre du plagiat : deux textes à comparer qui partagent la même surface lexicale sont une copie l'un de l'autre.

Dice (1945) présente un coefficient pour mesurer la similarité des ensembles. Ce coefficient est assez souvent utilisé en traitement automatique des langues naturelles comme mesure de référence à côté de la similarité cosinus. Elle est considéré comme une mesure lexicale, car elle mesure l'inclusion des ensembles formés par deux phrases à comparer. La formule est :

$$Sim = (A,B) = \frac{2 \mid A \cap B \mid}{\mid A \mid + \mid B \mid} \tag{1.9}$$

Soit A et B deux phrases différentes qui partagent en partie la même d'information. Si nous disons que A est inclus en B, nous pourrons aussi conclure que B est inclus en A est complètement différent ($A \subset B \neq B \subset$ A). Une mesure de l'inclusion de $A \subset B$ devrait être différent à l'inclusion de $B \subset A$. Cependant, les mesures symétriques ne satisfont cela (Bao *et al.*, 2003).

Comme on peut le voir, la normalisation de la mesure 1.9 rend ce coefficient une mesure symétrique.

1.3.2 Mesures à base taxinomique

Une mesure taxinomique utilise des connaissances linguistiques pour calculer la similarité entre deux concepts. Ces informations linguistiques dénotent des relations sémantique/lexicales comme la synonymie qui peut être utilisée pour le calcul de similarité entre deux segments de texte. L'utilisation de telles connaissances linguistiques a comme but d'améliorer la performance des mesures basées sur des concepts purement lexicaux projetés des espaces géo-

métriques comme SVM [14]. Ces mesures permettraient donc de capturer des relations lexico-sémantiques entre des mots.

Puisque les mesures taxinomiques utilisent de l'information pour chaque mot dans les segments de texte à comparer, nous allons d'abord présenter les mesures de similarité de mots. Ces mesures sont utilisées dans une stratégie de similarité de texte pour finalement donner une valeur globale de similarité.

WordNet Miller (1995) est une ressource largement utilisée pour créer ce genre de mesures. Il s'agit d'une base de données lexicale avec une structure de graphe. WordNet organise les noms, les verbes, les adjectifs et les adverbes en ensemble de synonymes, appelés *synsets*, chacun représentent un concept lexical. WordNet relie les *synset* par des relations sémantiques telles que :

- La synonymie : qui est la relation de base de WordNet, car elle est utilisée pour représenter le sens des mots et pour grouper les ensembles de synonymes. Il s'agit d'une relation symétrique.
- L'antonymie : il s'agit aussi d'une relation symétrique entre les formes de mots. Elle est spécialement utilisée pour organiser la signification des adjectifs et des adverbes.
- L'hyponymie : et son inverse, l'hyperonymie sont des relations transitives entre les *synsets*. Dû aux caractéristiques de cette relation sémantique, elle est utilisée pour organiser la signification des noms dans une structure hiérarchique.
- La méronymie (une partie) : et son inverse l'holonymie (le tout) sont des relations sémantiques *partie-tout*.

Chacune des relations est représentée par des pointeurs entre les formes de mots ou les *synsets*. Il existe plus de 116, 000 liens pour identifier les relations entre les concepts précédents, Miller (1995).

Pedersen *et al.* (2004) propose une librairie comprenant six mesures basées sur WordNet. Elles sont classifiées de la façon suivante :

14. Machine à vecteurs de support ou *support vector machine*, en anglais.

- Mesures basées sur la longueur du chemin entre deux concepts. La mesure de Leacock & Chodorow (1998) trouve le chemin le plus court entre deux concepts et le normalise par la longueur maximale du chemin dans la hiérarchie *is-a* [15] dans laquelle les deux concepts se produisent. La mesure de Wu & Palmer (1994) trouve la longueur du chemin entre le nœud racine du subsumé le moins commun *the least common subsumer* de deux concepts, qui est le concept le plus spécifique qu'ils partagent comme racine. Une autre mesure de *chemin* est égale à l'inverse de la longueur du chemin le plus court entre deux concepts, (Pedersen *et al.*, 2004).
- Mesures basées sur le contenu d'information. La mesure de Resnik (1995) suppose que la similarité entre deux concepts dépend du contenu de l'information qui les englobe dans la taxonomie. La mesure de Lin (1998) et celle de Jiang & Conrath (1997) sont considérées comme une extension de la mesure de Resnik (1995).

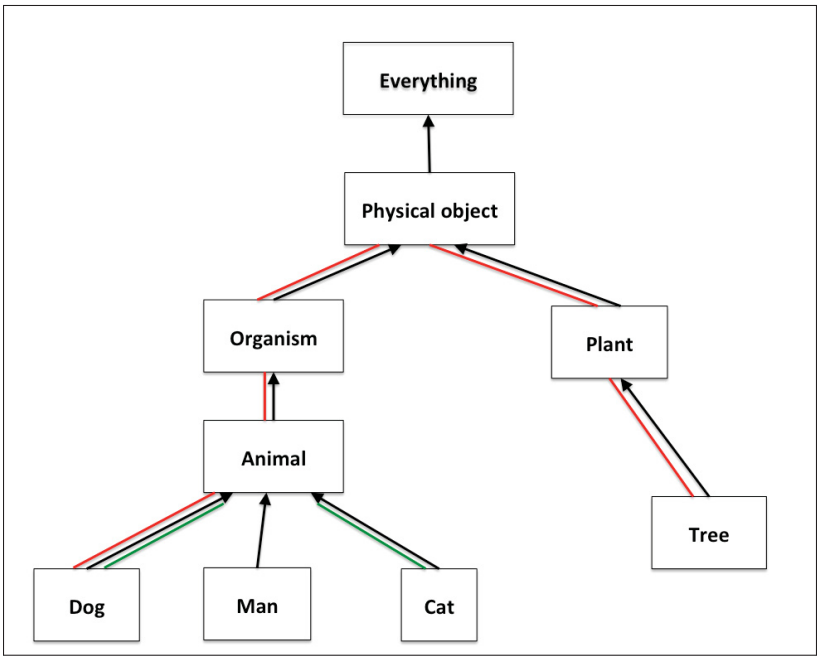


Figure 1.6 Structure d'un graphe de concepts pour le calcul de similarité.

15. En programmation orientée object, *is-a* est une relation entre une sous-classe et une classe.

À la figure 1.6, nous voyons l'exemple d'une structure de concepts sous la forme d'un graphe, ressemblant à ce qu'on peut retrouver dans *WordNet*. Une mesure basée sur la longueur du chemin entre deux concepts utiliserait le nombre d'arêtes entre deux concepts pour exprimer leur similarité. Plus la longueur du chemin est petite, plus les deux concepts sont similaires. Par exemple, la similarité entre les concepts *cat* et *dog* est élevée car ils sont très près dans le réseau ; il y a deux arêtes à parcourir (voir le chemin en vert). Par contre la similarité entre les concepts *cat* et *tree* est basse car il faut parcourir cinq arêtes (voir le chemin en rouge).

Quelle que soit la mesure utilisée, l'objectif est essentiellement le même. Pour calculer la similarité entre deux phrases, chaque phrase est divisée en jetons / mots pour ensuite calculer la similarité entre ces unités. La similarité finale entre deux phrases comprend une sommation de toutes les valeurs de similarité de mots. Par conséquent, nous trouvons quelques efforts pour fusionner les deux groupes, qui combinent la similarité lexicale et taxinomique pour le calcul de la similarité entre les deux phrases. Mihalcea *et al.* (2006) présentent une mesure de similarité pour textes courts qui combine l'information lexicale et sémantique pour le calcul global de la similarité entre deux segments de texte. La mesure de Mihalcea *et al.* (2006) est exprimée sous la forme suivante :

$$sim(T_1, T_2) = \frac{1}{2}\left(\frac{\sum_{w\in\{T_1\}}(maxSim(w, T_2) * idf(w))}{\sum_{w\in\{T_1\}} idf(w)} + \frac{\sum_{w\in\{T_2\}}(maxSim(w, T_1) * idf(w))}{\sum_{w\in\{T_2\}} idf(w)}\right) \tag{1.10}$$

Où $maxSim(w, T)$ sélectionne le mot w qui a la similarité maximale avec les mots dans le segment de texte T. La fonction $maxSim(w, T)$ peut utiliser l'une des mesures de similarité de mots sur WordNet, ce qui assure l'utilisation d'une certaine information sémantique. Le côté lexical de la formule est pris en considération par la valeur idf, (*inverse document frequency*) du mot w. Nous pouvons constater que cette formule est composée de deux éléments et que ces derniers sont séparés par une sommation. L'élément situé du côté gauche du symbole $+$ se charge de faire la sommation de toutes les valeurs maximales des mots faisant partie du segment

de texte T_1, qui se trouvent dans le segment de texte T_2. Cette sommation est normalisée par la sommation des valeurs idf de tous les mots w dans le segment de texte T_1. L'autre élément agit de la même façon, mais il cherchera les mots w du segment de texte T_2 qui se trouvent dans le segment de texte T_1. La moyenne de deux valeurs donne la valeur finale de la similarité.Avec cette moyenne des deux termes, la mesure devient symétrique.

1.3.3 Mesures à base syntaxico-sémantique

Mihalcea *et al.* (2006) suggèrent que les mesures de similarité qui sont conçues avec des approches de type *bag of words* ignorent la structure de la phrase (les relations entre le sujet et le prédicat). Indirectement les auteurs suggèrent, dans leur conclusion, qu'une mesure de similarité sémantique devrait prendre en compte les configurations de la langue comme les relations sémantiques ou syntaxiques.

Pour trouver automatiquement les arbres syntaxiques et les relations sémantiques, nous avons besoin d'un autre type d'outil. Par exemple *FrameNet* (Ruppenhofer *et al.*, 2006), est une ressource lexicale pour l'anglais basé sur la sémantique des cadres [16]. « La sémantique des cadres, qui suit la théorie des cas de Fillmore (1967), est une théorie de la signification qui souligne la relation étroite entre le langage et l'expérience »(Roth, 2014).

L'objectif de FrameNet, disponible en ligne, est de documenter la gamme de combinaisons sémantiques et syntaxiques pour chaque mot dans chacun de ses sens. Ce processus se réalise à l'aide d'ordinateurs. FrameNet contient environ 10 000 unités lexicales et plus de 135 000 phrases annotées. Une unité lexicale est l'appariement d'un mot avec une signification. Généralement, chaque sens d'un mot polysémique appartient à une cadre sémantique différent. Dans ce contexte, un cadre sémantique est une sorte de structure qui décrit une situation particulière, un objet ou même un événement avec ses participants et d'autres éléments qui donne plus d'information sur le même événement (Ruppenhofer *et al.*, 2006). Nous présentons, ici, un exemple tiré de Ruppenhofer *et al.* (2006) :

16. *Frame semantics.* en anglais

(1) [$_{Cook}$ *Matilde*] ***fried*** [$_{Food}$ *the catfish*] [$_{Heating\_instrument}$ *in a heavy iron skillet*]

Dans la phrase présentée en (1), l'unité lexicale est *fried* et les éléments du cadre sont *Cook* pour le sujet, *Food* pour l'objet et *Heating_instrument* pour l'adjoint-locatif. Pour un verbe comme *fried*, nous aurons accès à ses dépendants syntaxiques ; ces éléments permettraient aux outils de pouvoir générer des arbres syntaxiques ou d'identifier des rôles sémantiques dans les phrases. Les cadres ne se limitent pas aux verbes ; sur FrameNet, nous en trouverons pour quelques noms et certains adjectifs.

Un autre outil utilisé pour ce type d'analyse est PropBank, qui contient les annotations des arguments et des adjoints des verbes. Chaque argument du verbe est numéroté à partir de zéro jusqu'à *n*, selon le nombre d'arguments du verbe présents dans une phrase, (Palmer *et al.*, 2005). L'argument zéro, *Arg*0, est généralement le sujet du verbe et *Arg*1 correspond à l'objet du verbe. Si le verbe présente plus d'arguments, il seront numérotés à partir de *Arg*2. La motivation de cette numérotation est due à la difficulté de définir un ensemble de rôles thématiques couvrant tous les types de prédicats, (Palmer *et al.*, 2005). Nous montrons un exemple tiré de Palmer *et al.* (2005), p. 78.

(2) [$_{Arg0}$*John*] *opened* [$_{Arg1}$*the door*] [$_{Arg2}$ *with his foot*]

Dans la phrase (2) le verbe *open* présente trois arguments, *Arg0, Arg1, Arg2*. Le premier correspond au sujet, le deuxième à l'objet et finalement *Arg2* correspond à un instrument.

Deux caractéristiques principales distinguent FrameNet et PropBank. D'abord, FrameNet contient des annotations pour les verbes, les noms et les adjectifs. Alors que FrameNet contient des annotations pour les verbes reliant des annotations fondées sur la sémantique, car elle sont basées sur la sémantique des cadres. PropBank comprend des annotations fondées sur des critères syntaxiques. Le principal problème avec ce type d'outil réside dans le nombre de mots courants, c'est-à-dire les unités lexicales comprises dans la base de données. Il peut arriver que quelques

mots dans la phrases à traiter ne soient pas disponibles. Notons que ces outils sont d'avantage développés pour la langue anglaise.

Nous trouvons aussi d'autres concepts linguistiques qui sont exploités pour la conception de mesures de similarité de textes. Roth (2014) définit une mesure de similarité qui compare les têtes de la structure d'arguments de la phrase (Head Argument, HA), lorsqu'elles partagent la même étiquette argumentale. Pour extraire la HA, il utilise un étiqueteur de rôles sémantiques. Cette mesure se définit au moyen de la formule suivante :

$$sim_{AHeads}(p_1, p_2) \quad = \quad \frac{\sum_{\{a_1,a_2|label(a_1)==label(a_2)\}} sim_{WN}(head(a_1), head(a_2))}{|\{a_1, a_2|label(a_1) == label(a_2)\}|} \quad (1.11)$$

La similarité entre les prédicats p_1 et p_2 est calculée si les AH a_1 et a_2 ont la même étiquette d'argument. Autrement dit, pour chaque tête d'argument, Roth (2014) calcule la similarité pour les deux arguments étiquetés comme *Arg0* et la similarité pour les arguments étiquetés comme *Arg1* [17]. *SimWN* est la similarité de Lin (1998) calculée dans WordNet. Dans son travail, Roth (2014) utilise plusieurs mesures de similarité pour construire des graphes sur lesquels il applique des algorithmes de *clustering* pour aligner les structures d'arguments du prédicat. Roth (2014) constate que la mesure de similarité qui a contribué le plus à un bon alignement est celle de tête de prédicats, puisque cette mesure considère partiellement la structure de la phrase. La mesure de tête de prédicat de Roth (2014) utilise à la fois des informations syntaxique et sémantique ce qui fait de cette mesure une volonté accrue de prendre en compte les caractéristiques du langage dans le calcul de similarité de phrases.

Dans une approche différente, Blanco & Moldovan (2015) proposent une mesure de similarité de texte basée sur la représentation logique d'une phrase, ce qui permet de prendre en compte la structure de la phrase. Leurs résultats représentent une avancé dans la structure de la phrase. Pour ce faire, ils calculent d'abord la forme logique de phrases qui s'approchent de la défini-

17. L'étiquette *Arg0* est généralement associée au sujet d'un verbe ; *Arg1* est associé au complément du verbe.

tion d'événement selon Davidson (2001) décrite à la section 1.5.3. Ils utilisent la logique de premier ordre pour la représentation d'une phrase. La seule ressource externe qu'ils utilisent est WordNet. Pour le calcul de similarité entre deux phrases S_1 et S_2, ils suivent les étapes suivantes :

- Les phrases S_1 et S_2 sont transformées en forme logique. Pour ce faire, un *pipeline* qui contient des outils du TALN ; des annotateurs de classes lexicales, d'entités nommées et de rôles sémantiques sont utilisés. L'ensemble de ces outils construit de prédicats logiques pour les noms, les verbes, les adjectifs et les adverbes. Il est de important souligner que cette approche utilise des ressources empruntées à FrameNet, PropBank, NomBank et les résultats de données provenant des conférences SemEval, (Blanco & Moldovan, 2015).
- Les formes logiques de S_1 et S_2 soumises à un démonstrateur de théorèmes pour réaliser trois transformations logiques. Lors de ces transformations, certains des prédicats seront éliminés lorsqu'une preuve ne pourra pas être trouvée. À partir de ces transformations, certaines des caractéristiques seront conservées. À partir de ce résultat, les prédicats sont utilisés pour calculer la valeur de similarité avec une mesure de similarité WordNet disponibles dans la paquet de Pedersen *et al.* (2004).
- La valeur de similarité finale est réalisée avec un algorithme supervisé de *Machine Learning* qui combine les valeurs de similarité des prédicats et les caractéristiques des preuves logiques.

Pour finir, Blanco & Moldovan (2015) mentionnent que la similarité de texte est symétrique.

Ferreira *et al.* (2016) présentent une stratégie pour mesurer la similarité de phrases en considérant les composants lexicaux, syntaxiques et sémantiques de la phrase. La similarité de chaque composant est calculée indépendamment. Pour le composant lexical, Ferreira *et al.* (2016) divisent la phrase en *tokens* et utilisent l'une des mesures *WordNet* et la distance de *Levenshtein* [18]. La distance de *Levenshtein* sera utilisée lorsque la mesure *WordNet* entre deux éléments

18. La distance de *Levenshtein* est aussi connue sous le nom de *distance d'édition*, il s'agit de considérer combien d'opérations de suppression et addition sont nécessaires pour qu'un mot $m_1$soit converti un autre mot m_2. Plus la distance est grande, plus les mots sont différents.

lexicaux sera inférieure à 0.1. Pour la partie syntaxique, Ferreira *et al.* (2016) utilisent un arbre de dépendances pour identifier les relations grammaticales comme le sujet, l'objet et les modificateurs adverbiaux. Par la suite, cette information est utilisée pour construire un graphe sous la forme d'une représentation Resource Description Framework (désormais RDF). Les nœuds contiennent les éléments obtenus dans l'étape lexicale, et les arêtes dénotent les relations trouvées précédemment. La similarité syntaxique est calculée en utilisant la relation de la couche syntaxique calculée par la correspondance des nœuds des triplets RDF. D'abord, une correspondance entre les triplets est réalisée. Pour calculer la similarité entre les nœuds, les auteurs utilisent les étiquettes qui représentent le nom des nœuds ; cette similarité est calculée avec une similarité de mots. La similarité est la moyenne arithmétique de ces similarités. En ce qui concerne la partie sémantique, Ferreira *et al.* (2016) utilisent *FrameNet* et *Semafor* pour déterminer l'annotation des rôles sémantiques et l'identification du sens. Dans cette couche, le processus de calcul de similarité est similaire à celui de l'étape de similarité syntaxique. L'information sémantique sera utilisée pour construire un graphe. La similarité finale d'une phrase est la combinaison de toutes les couches : lexicale, syntaxique et sémantique.

$$similarity(S_1, S_2) = \frac{(lex_n \times lex_s) + (syn_n \times syn_s) + (sem_n \times sem_s)}{lex_n + syn_n + sem_n} \tag{1.12}$$

Où lex_n est la somme du nombre de mots des deux phrases. syn_n et sem_n sont la somme du numéro de triplets des graphes syntaxiques et sémantiques. lex_s, syn_s, et sem_s sont les valeurs des similarités obtenues dans les couches lexicales, syntaxiques et sémantiques. Ferreira *et al.* (2016) définissent explicitement cette mesure comme étant symétrique, voir (Ferreira *et al.*, 2016, p. 17).

Le tableau 1.1 présente un résumé des mesures décrites précédemment.

Tableau 1.1 Mesures de similarité textuelle.

Nom de mesure	**Base**	**Classification**
Mesure cosinus	Lexical	Symétrique
Mesure Dice	Lexical	Symétrique
Relative Frequency Model	Lexical	Asymétrique
Inclusion Proportion Model	Lexical	Asymétrique
Heavy Frequency Model	Lexical	Asymétrique
Mesure Mihalcea *et al.* (2006)	Taxonomique	Symétrique
Mesure Roth (2014)	Syntaxico-sémantique	Symétrique
Mesure Blanco & Moldovan (2015)	Syntaxico-sémantique	Symétrique
Mesure Ferreira *et al.* (2016)	Syntaxico-sémantique	Symétrique

1.4 Scénario 1 : les dissertations d'étudiants.

1.4.1 La langue et son rôle dans l'apprentissage

Des travaux récents ont démontré l'implication des phénomènes cognitifs et linguistiques dans les processus d'apprentissage (Dascalu *et al.*, 2015; Chen *et al.*, 2014; Jain *et al.*, 2014; Scheihing *et al.*, 2016; Schleppegrell, 2007). Particulièrement, Schleppegrell (2007) fait un survol des défis linguistiques que les étudiants doivent affronter dans les cours de mathématiques. Schleppegrell (2007) soutient que chaque domaine a ses propres chemins pour construire la connaissance et que les étudiants ont besoin d'être en mesure d'utiliser la langue pour participer à ce parcours de connaissance. Pour sa part, Halliday (1978, p. 195) discute les défis linguistiques en apprentissage, particulièrement en mathématiques, et signale que des opérations mathématiques communes comme *compter* ou *mesurer* utilisent des expressions de la langue de tous les jours. Halliday ajoute que les concepts mathématiques appris à l'école exigent une nouvelle utilisation de la langue qui servira à de nouvelles fonctions. Ceci ne veut pas nécessairement dire que les étudiants vont apprendre de nouveaux mots, (plutôt une nouvelle signification de ce même mot), une nouvelle forme d'argumentation et même une nouvelle façon de combiner des éléments déjà existants. Halliday (1978) définit un "registre mathématique" de la façon suivante :

> A set of meanings that is appropriate to a particular function of language, together with the words and structures which express these meanings.We can refer to a “mathematics register”, in the sense of the meanings that belong to the language of mathematics (the mathematical use of natural language, that is : not mathematics itself), and that a language must express if it is being used for mathematical purposes.

Pour ce qui est de l’utilisation de différents registres de langue, soit familier ou soigné, Beaudoin-Bégin (2015, p. 50) signale que ce n’est pas le sujet qui détermine le registre à utiliser ; cette tâche est plutôt tributaire de la situation de communication. Dans un contexte académique, une situation de communication doit répondre aux contextes du domaine où l’étudiant se retrouve. En prenant la définition de registre mathématique de Halliday (1978), nous pourrions supposer que chaque domaine aurait son propre registre, car chaque domaine aurait ses propres besoins de communication. Pourtant, nous considérons que le langage utilisé pour un domaine académique ciblé devrait être analysé dans son propre esprit afin de capturer la cohérence et le sens d’un texte.

Schleppegrell (2007) suggère qu’apprendre le langage d’un nouveau domaine fait partie de l’apprentissage de celui-ci, puisque la langue et l’apprentissage ne peuvent pas être séparés. Étant donné que la première et unique compétence est le registre familier et qu’elle leur a permis de construire leur connaissance du monde, l’école permet d’accéder à un registre soutenu dans contexte scientifique. Bien sûr, il faudra être attentifs aux défis linguistiques qui accompagnent l’appropriation des nouveaux concepts lors de l’apprentissage (Schleppegrell, 2007).

Les travaux qui ont fait l’analyse des défis linguistiques dans le processus d’apprentissage portent principalement sur des aspects sémiotiques, ou même des patrons grammaticaux, de l’utilisation de conjonctions, de vocabulaire technique, de phrases nominales condensées, et de relations logiques (Schleppegrell, 2007).

1.4.2 L'analyse linguistique de textes académiques

Gardner & Davies (2013) présentent un lexique sur le vocabulaire académique en anglais. Celui-ci a été créé à partir d'un corpus de 120 millions de mots. La maîtrise de ce lexique de la part des étudiants pourrait servir à discriminer lors des tests d'admission, tels TOEFL ou MCAT, etc. D'ailleurs, des experts mentionnent l'importance de la maîtrise d'un lexique académique, plus spécifiquement d'un lexique de base comme celui présenté par Gardner & Davies (2013). Ce lexique devient dont essentiel pour une compréhension académique.

Durrant (2014) présente une analyse pour déterminer dans quelle mesure le lexique reste constant dans différents groupes d'étudiants universitaires. L'objectif était de voir les éléments partagés dans les différentes disciplines, Durrant (2014) opte pour une approche de groupement de mots basée sur des statistiques et la fréquence de mots. Durrant (2014) réalise son analyse sur le corpus, *British Academic Written English* [19] qui est constitué d'écrits des étudiants provenant de quatre universités britanniques. Le corpus contient des textes élaborés autant par des locuteurs de langue maternelle anglaise que de langue seconde. Lors de son analyse, Durrant (2014) ne fait aucune distinction par rapport à la langue maternelle des étudiants. Leurs résultats suggèrent qu'il n'existe pas une liste générique de mots qui pourrait appartenir à toutes les disciplines ; le partage de mots était inférieur à 49 % entre les domaines les plus compatibles. Par conséquent, Durrant (2014) suggère la création d'un lexique spécialisé pour chacune des disciplines.

Martinez & Schmitt (2012) présentent l'analyse d'une liste de 505 expressions. Ces dernières sont de taille variable ; le critère pour leur sélection est tout n-gramme ayant une fréquence supérieure à 787 fois sur un corpus de 100 millions de mots. C'est ainsi que nous pourrons trouver dans cette liste des expressions comme : *on the other hand* ou *take for granted*. Le but de cette liste est d'aider l'acquisition des expressions les plus fréquentes dans un contexte d'apprentissage de l'anglais comme langue seconde.

19. Corpus disponible sur : www.coventry.ac.uk/bawe/

Dans un effort pour comprendre la fonction des collocations (*lexical bundles*), Cortes (2004) analyse les écrits d'étudiants en histoire et en biologie de trois différents niveaux académiques. Les écrits des étudiants sont comparés avec des écrits publiés par des professionnels dans les mêmes domaines. Dans son travail, Cortes (2004) catégorise une liste de 109 collocations selon le contexte dans lequel elles apparaissent, par exemple, des marqueurs de temps, de lieu, de quantité, etc. Les collocations choisies répondaient au critère d'une fréquence supérieure à 20 pour un corpus de plus d'un million de mots. Ces collocations étaient restreintes à une taille de quatre mots, ou quadri-grams, par exemple, *as a result of, at the same time, etc.* Son analyse se situe plus au niveau stylistique, puisque ces collocations font appel à une façon spécifique d'exprimer et d'enchaîner les idées. Dans ce sens, Cortes (2004) conclut que les étudiants n'utilisent pas assez les mêmes collocations que les professionnels et que leur utilisation diffère de celle utilisées par ces derniers. Nous voyons une motivation pour analyser plutôt les termes qui sont utilisé par les étudiants dans leurs dissertations. Notons aussi que cette analyse devrait répondre aux particularités du domaine ainsi qu'à celles des étudiants.

1.4.3 Évaluation automatique de dissertations d'étudiants

L'analyse de l'apprentissage (learning analytics, LA, en anglais) se concentre sur la collecte, la mesure et l'analyse des données sur les étudiants et leurs contextes (De Jong, 2015). Au cours de la dernière décennie, la LA a émergé comme un moyen d'analyser quantitativement le processus l'apprentissage. La plupart du temps, la LA se concentre sur les fréquences de participations, les contributions, etc. (De Jong, 2015). Compte tenu des progrès récents en TALN et des techniques d'extraction de texte (Text Mining), il est désormais possible d'intégrer de nouveaux modèles au sein de la LA afin d'étudier le développement de nouveaux concepts par les élèves au sein des approches visant le renforcement des connaissances [20]. Puisque la LA concerne l'analyse de données des étudiants, dissertations ou participations dans les forums, nous présentons quelque travaux en linguistique appliquée portant sur les écrits des étudiants.

20. Knowledge building.

En ce qui concerne l'évaluation automatique des dissertations d'étudiants, Warschauer & Ware (2006) décrivent trois outils commerciaux. Ces outils ont été conçus pour l'évaluation des écrits des étudiants en l'anglais langue seconde et permettent d'identifier des erreurs communes pour les locuteurs natifs d'autres langues comme l'espagnol, le français ou le japonais. Ces outils ont comme objectif d'accompagner les étudiants tout au long de la rédaction des essais et d'en donner une évaluation rapide. Sans ces outils, l'évaluation d'un brouillon prend beaucoup de temps à reviser pour un enseignant. De plus, il faut considérer que l'étudiant peut livrer plusieurs brouillons avant de livrer sa version finale (Warschauer & Ware, 2006) .

L'un des premiers outils discutés par Warschauer & Ware (2006) est *Intellimetric* ^{TM}, qui fait partie d'un produit commercial appelé My Acces !. Il permet d'analyser environ 300 caractéristiques sémantiques, syntaxiques et discursives pour une dissertation en les comparant avec un échantillon de dissertations déjà évaluées par des êtres humains (Elliot, 2003). My acces ! donne un score (dans une échelle de 1 à 4 ou 1 à 6) individuel pour le focus, l'organisation, le contenu, le développement, l'utilisation de la langue et le style (Warschauer & Ware, 2006).

Criterion^{SM} est aussi un autre système d'évaluation automatique de dissertations d'étudiants qui contient les moteurs E-Rater et Critique ainsi que des techniques de TALN ; l'ensemble de ce système fait partie des services de la compagnie *Educational Testing Services, (ETS)*. E-Rater compare les dissertations soumises par les étudiants avec d'autres dissertations qui ont été notées par des experts (Burstein *et al.*, 2004). Plus spécifiquement, pour un document, E-rater analyse et évalue les taux d'erreurs en grammaire, le numéro requis des éléments de discours, et la complexité lexicale utilisée (Warschauer & Ware, 2006). Tous ces critères d'évaluation permettent à Criterion de faire des diagnostics linguistiques et de donner des commentaires aux étudiants au niveau de la grammaire, de l'utilisation des mots, ainsi que du style et de l'organisation du document (Burstein *et al.*, 2004).

Intelligent Essay Assessor^{TM} (IEA) présenté par Landauer *et al.* (2003), utilise *Latent Semantic Analysis* [21] pour l'analyse sémantique des résumés de documents soumis en ligne par des étudiants en les comparant avec les documents originaux. L'utilisation de LSA par IEA permet à cet outil d'être utilisé dans d'autres langues. Dans son étude, Warschauer & Ware (2006) mettent en évidence une lacune méthodologique, notamment le manque de groupes de contrôle, lors de la réalisation des études qui visent à montrer l'utilisation bénéfique des systèmes d'évaluation automatiques de dissertations. Ceci dit, il n'est pas possible de prendre une position objective à cet égard. Finalement, les critères d'évaluation remettent en question la façon d'enseigner et d'apprendre, et les logiciels d'évaluation automatique pourront augmenter cette tendance, menant les étudiants à une approche stéréotype de la composition, (Warschauer & Ware, 2006).

1.4.4 Le TALN dans le contexte de Learning Analytics.

Scheihing *et al.* (2016) analyse les interactions textuelles étudiants-étudiants et étudiants-enseignants sur une plateforme en apprentissage collaboratif supporté par ordinateur (ACSO) [22]. Leur but est d'évaluer les corrélations entre les attitudes discursives et d'autres variables qui sont liées aux activités d'apprentissage pour supporter le monitorat des enseignants et le développement des activités d'apprentissage. Pour ce faire, Scheihing *et al.* (2016) ont classifié les interactions textuelles des étudiants et des enseignants. Ces interactions pouvant être classifiées en trois des catégories d'intention de communications de Jakobson (1972). Elles sont :

- *Phatic* : dénote les efforts pour commencer et maintenir une conversation.
- *Referential* : fait référence au contexte du sujet de la conversation.
- *Emotive* : pouvant dénoter, de façon véritable ou feinte, les émotions et les attitudes du locuteur.

21. Notez que le terme sémantique ici est un abus, car on ne peut pas réduire la sémantique à l'analyse de la synonymie à laquelle LSA prétend. De plus, et pour des raisons déjà exposées, LSA permet uniquement l'analyse du contexte de mots, qui parfois permet de capturer la synonymie .

22. Computer-Supported Collaborative Learning (CSCL) en anglais.

Pour Scheihing *et al.* (2016), la classification des interactions des interventions des étudiants dans ces catégories dénoterait le progrès et l'engagement des étudiants ainsi que la communication avec les enseignants. Techniquement, la classification est automatiquement faite avec une approche basée sur SVM et LDA. Les meilleurs résultats sont obtenus avec l'approche SVM. SVM et LDA sont des approches fondées sur le concept de fréquence des mots et de leurs probabilités.

Dascalu *et al.* (2015) présentent *ReaderBench* qui est un système d'évaluation automatique des collaborations basé sur la cohésion et le *dialogisme*. Les collaborations, matière d'étude de ce travail, appartiennent à un environnement ACSO. *ReaderBench* est capable d'analyser les entretiens et les discussions entre les étudiants. Originalement, *ReaderBench* réalise l'analyse de la voix ; à partir de cette information, il construit les chaînes lexicales couvrant toute la conversation, pour ensuite les fusionner dans des chaînes sémantiques avec un algorithme agglomératif de classification hiérarchique. Dascalu *et al.* (2015) basent l'interprétation de ces analyses sur la théorie de cohérence de texte proposé par Halliday & Hassan (1976). Cette théorie postule que la cohésion d'un texte est basée sur les caractéristiques qui mettent en évidence les relations entre des éléments constitutifs (mots, phrases ou expressions). La cohésion du texte peut également être décrite comme l'ensemble des relations lexicales, grammaticales et sémantiques qui relient les unités textuelles ensemble. Halliday & Hassan (1976) décrivent formellement la cohérence en termes de références, de substitutions, d'ellipses, de conduction et de cohésion lexicale. Cette dernière est peut-être le concept le plus adéquat pour l'analyse que Dascalu *et al.* (2015) réalisent sur les contributions des étudiants. En effet, selon Halliday & Hassan (1976) la cohésion lexicale est atteinte en analysant le vocabulaire ; les étudiants doivent faire une sélection des mots pour cristalliser la cohésion dans leur textes. Si la cohérence lexicale est atteinte par la bonne utilisation des relations sémantiques, un processus de similarité basé sur WordNet pourrait servir à évaluer cette cohérence grâce à la présence de relations sémantiques qui peuvent être exploitées à cet effet.

1.5 Scénario 2 : les textes journalistiques

1.5.1 Les origines

En Rome antique, nous trouvons deux types de publications : la *Acta Diurna* aussi appelée *Acta populii* ou *Acta publica* agissait comme gazette des événements politiques et sociaux. Elle a fait son apparition avant 59 BCE. Après le 27 BCE, la *Acta diurna* devient une publication quotidienne ; elle peut être considérée, de nos jours plus ou moins, comme un journal. Nous trouvons aussi l'*Acta senatus* qui présentait les minutes des procédures du Sénat. Elle fait son apparition en 50 BCE et était uniquement disponible aux sénateurs. À l'arrivée d'Auguste, en 27 av. J-C. la *Acta senatus* fut mise en accès restreint. Sous le pouvoir de Tiberius, 1^{er} siècle CE, la *Acta senatus* est conservée dans les archives impériales et les bibliothèques publiques (of Encyclopædia Britannica, 2016).

Ces premières publications peuvent-elles être vraiment considérées comme des journaux selon les termes actuels ? Quels sont alors les critères pour définir une publication comme un journal. À cet égard, Richardson (2006) signale que le journalisme existe pour permettre aux citoyens de mieux comprendre leur vie et leur position dans le monde. Ces premiers publications ont eu un rôle plus marquant dans le développement de la vie de l'empire comme tel, mais dû à leur accès difficile et leur reproduction, il est très peu probable que de telles publications auraient eu une influence sur la compréhension du monde des habitants de l'empire.

Richardson (2006) présente quatre critères pour qu'un média imprimé soit considéré comme un journal :

a. qu'il soit publié,
b. qu'il soit périodique,
c. qu'il aborde des thématiques différentes,
d. qu'il soit universel.

Revenant sur la trace historique des publications, l'invention de l'imprimante par Johannes Gutenberg de Mainz, au XV siècle, a permis la circulation plus rapide des versions imprimées

des manuscrits. Gutenberg, qui se consacrait la plupart du temps à la reproduction des bibles, a aussi reproduit dans une série de petites publications *« Türkenkalender »*, des avertissements à la chrétienté contre les Turcs. Cela a marqué le début de la presse politique. Ces publications étaient divisées en chapitres correspondant à chaque mois. La plupart des publications en Europe pendant le Moyen Âge portaient principalement sur des sujets en politique (Weber, 2006).

Selon Weber (2006), la presse moderne, comme nous la connaissons de nos jours, est née à l'automne de 1605 dans la ville de Strasbourg en France. C'est grâce à Johann Carolus, qui demande la permission de reproduire un nouveau type de produit dans sa presse, ou plutôt d'avoir le monopole pour le faire. Johann Carolus était un imprimeur et écrivain ; il cumule ces deux habilités pour reproduire des lettres de nouvelles dans son hebdomadaire, le *Relation*. Carolus publiait des textes politiques sous la prémisse « *without any additions and not otherwise than in the manner in which it has been written and received* » (Weber 2006, p. 393). Par la suite, plusieurs journaux ont suivi la même formule. Cela assurait le respect des idées originales d'un écrivain, ou plutôt d'un journaliste ; en même temps, le journal présentait sa défense contre toute possible accusation de manuscrits faux ou offensifs, (Weber, 2006).

Un peu plus d'un siècle plus tard, le *Relation aus rem Parnasso*, un journal établi en 1787 à Hambourg publiait les premiers articles de discussions, de commentaires ou de jugements (Weber, 2006). À l'époque, il y avait des journaux imprimés et écrits à la main, dont le contenu restait le même : rapports politiques, évènements diplomatiques et militaires provenant de toutes les parties du monde connu. Plus rarement, il était possible de trouver des nouvelles abordant des sujets sur les conditions de météo inhabituelles, les catastrophes naturelles, les inondations, les tremblements de terre, les éruptions volcaniques, les incendies, les crimes ou les miracles Weber (2006). De plus, Weber (2006) mentionne aussi que c'est pendant le XVII siècle, que les journaux prennent leur propre ligne éditorial en sélectionnant les manuscrits concernant à un parti politique en particulier.

Speed (1893) remarque un changement rapide dans le contenu des journaux aux États-Unis. Dans son article *Do newspapers now give the news ?*, Speed situe ce changement entre 1881 et 1893 dans les journaux les plus populaires de New York à époque, *The Tribune, The World, The Times and The Sun*. Par exemple, en 1891, la présence de nouvelles abordant la religion et les critiques d'art occupait entre une ou deux colonnes ; en 1893, ce même type de nouvelles n'était plus présent . Dans ce cas, nous constatons une perte dans la couverture de l'information par rapport à une thématique. Le cas contraire a été observé aussi par Speed avec des sujets reliés au sport, qui a vu une augmentation de trois colonnes à dix dans la même période.

Comme que nous le voyons, la présence des thématiques et leur couverture se voient entre autres conditionnées par divers critères, soit le choix de l'éditeur ou la préférence du public, etc. Nous constatons aussi que, dès le début du journalisme, en 1605, le « biais » était déjà présent. Par exemple, des positions politiques était déjà prises dans les écrits du *Relation* à Hambourg. Le phénomène du biais est naturellement important dans le cadre de cette thèse.

1.5.2 Le biais

Le contenu d'une nouvelle dépend de divers critères qui ont fait l'objet d'études de polarisation des médias, c'est-à-dire le biais. D'Alessio & Allen (2000) classifient le biais en trois groupes :

- *Gatekeeping* ou *contrôle d'information*. Ce type de biais se produit quand les journalistes ou les éditeurs choisissent, dans un groupe de nouvelles, celle qui sera présentée au grand public. En conséquence, ils omettent les nouvelles qui ne seront pas abordées dans leurs journaux (D'Alessio & Allen, 2000).
- Couverture. Ce type de biais se retrouve quand plus d'espace est accordé aux événements particuliers concernant une personne ou un groupe. Ce type de biais se mesure « physiquement »par la quantité d'informations produites pour chaque thématique. Dans la presse imprimée, nous parlons de nombre de colonnes ou même du nombre de pouces carrées qui contient de l'information pour une thématique (D'Alessio & Allen, 2000).

- *Déclaration*. Il se produit quand les journalistes et éditeurs intègrent leurs opinions dans le texte de nouvelle. Le ton de cette opinion pourrait être favorable ou défavorable à propos d'une personne ou d'un groupe en particulier (D'Alessio & Allen, 2000).

Dans le cadre de cette thèse, nous nous intéressons au second type de biais, celui de la couverture. Une façon de déterminer la couverture d'une nouvelle consiste à identifier la nouvelle information qui apparaît dans les médias. La nouveauté, dans ce cas, se fait à partir de la comparaison du contenu informatif d'une nouvelle avec une autre, ou même avec un ensemble de nouvelles. La découverte de l'information non abordée par la première source d'information est une indication d'un manque de couverture de la part d'un journal. En TALN dans le domaine de *text mining*, nous trouvons le sujet de détection de nouvelle information. La conférence *Text REtrieval Conference (TREC)* proposait, entre 2002 et 2004, une tâche visant la détection de nouvelle information.

1.5.3 La structure d'une nouvelle : les événements

Le texte journalistique est caractérisé par une structure narrative. Cette dernière comprend une séquence dans laquelle les événements se produisent chronologiquement et dans le même cadre. Selon Richardson (2006), une narration a une structure d'actions et elle comporte les éléments suivants :

- introduction des caractères et mise en situation,
- augmentation des actions,
- introduction de complications,
- climax où la compilation est résolue,
- résolution finale.

D'une façon plus simplifiée, une nouvelle pourrait respecter une structure suivante :

- Compilation : l'événement actuellement reporté,
- Établissement du cadre de l'événement,

- Résolution.

Cette structure est visualisée dans une structure pyramidale inversée, où l'information plus importante est présentée dans les premiers paragraphes du texte. Dans le premier paragraphe, on trouve les réponses aux questions *qui a fait quoi à qui ? pourquoi ? où ?* et *quand ?* (Bell, 1991). Ces dernières questions nous amènent à l'exploration des concepts en linguistiques pour aborder le contenu des nouvelles.

Chaque macro-événement est cependant décrit par un ensemble de phrases qui, d'un point de vue linguistique, décrit autant de petits événements ou micro-événements. Examinons de plus près ce concept puisqu'il s'avère capital à la définition d'une mesure adéquate.

Essentiellement, le contenu d'une nouvelle est un événement ou un ensemble d'événements. La définition d'événement varie selon le champ d'études. Par exemple, en statistiques, un événement peut être considéré comme un pic dans un ensemble de données observées. Pour nous aligner avec le point de vue journalistique, nous parlerons des événements comme les choses qui se passent dans le monde réel et leur interprétation linguistique. Selon Rosen (1999), la linguistique offre trois possibilités pour représenter un événement :

- Le lexique : Par exemple, comme Pannini et Platon l'avaient décrit, les catégories lexicales pourraient être reliées aux événements dans la mesure où les noms dénotent des choses et les verbes les actions. De nos jours, la théorie de la structure des arguments suppose que le verbe, c'est-à-dire l'action, contrôle ce qui se passe dans la phrase. Par conséquent, les participants de l'événement sont déterminés par les arguments du verbe.
- La sémantique : La signification de la phrase est étroitement liée aux caractéristiques de l'événement.
- La syntaxe : Le début et la fin d'un événement sont connectés aux fonctions syntaxiques de cas et d'accord.

Le concept d'événement a été largement étudié en philosophie et en linguistique. Il existe plusieurs variates de la définition selon l'interprétation de l'auteur. Le philosophe Donald Davidson présente, dans Davidson (2001), une forme logique pour des phrases simples. Sa pro-

position inclut les verbes d'action et les modificateurs adverbiaux. Pour une phrase comme (3), nous aurons une forme logique comme 1.13 :

(3) Jonh buttered the toast with a knife in the kitchen

$$\exists e[butter(Jonh, the\ toast, e) \wedge with\_a\_knife(e) \wedge in\_the\_kitchen(3)] \qquad (1.13)$$

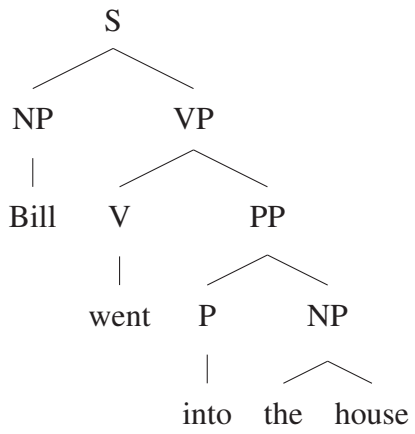
Dans cette représentation, l'élément *e* est une variable qui représente l'événement et qui peut être ajoutée à tout modificateur de l'élément principal qui contient le verbe *butter (Jonh, the toast, e)* ... tel comme on le voit dans les éléments $with\_a\_knif(e)$ et $in\_the\_kitchen(3)$ de 1.13. Nous allons retenir que Davidson (2001) fait correspondre la structure argumentale du verbe au corps de l'événement. Cette dernière caractéristique sera reprise par plusieurs chercheurs en sémantique et en syntaxe pour la représentation des événements.

Une des limitation de la forme logique de Davidson est que nous ne sommes pas en mesure de déterminer le nombre d'arguments à partir de la forme logique d'une phrase ; cela dépend entièrement de la syntaxe (Ratté, 1995). La composition de la phrase dépend donc de la syntaxe, puisqu'elle détermine la quantité d'arguments pour un verbe en particulier. Par contre, la création d'une forme logique d'une phrase apparaît une fois que la phrase a été entièrement analysée. La forme logique correspond donc à l'interprétation du sens d'une phrase ; par contre, la syntaxe permet la création de règles de production de phrases (des grammaires, ou simplement des patrons), qui guident à leur tour, la création du sens.

Jackendoff (1992) présente des structures conceptuelles formées à partir de éléments sémantiques. Il s'agit d'un inventaire de catégories sémantiques qui se veulent universels. Ces composants forment une structure arborescent contrainte selon des règles de combinaison appliquées sur des éléments simples. Parmi les catégories sémantiques universelles présentées par Jackendoff, nous trouvons la représentation des événements, *EVENT*. Ainsi, pour la phrase (4) nous obtenons la structure syntaxique en (5).

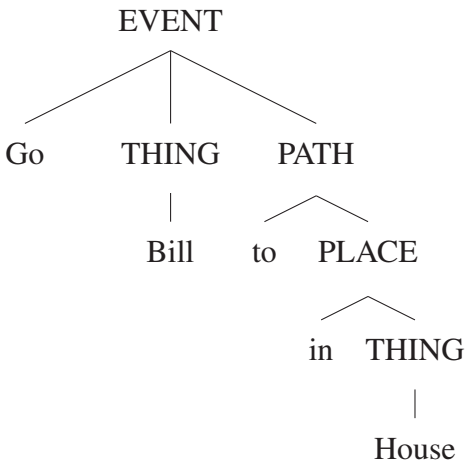
(4) Bill went into the house

(5) [$_{S}$ [$_{NP}$ Bill] [$_{VP}$ [$_{V}$ went] [$_{PP}$ [$_{P}$ into] [$_{NP}$ the house]]]]



Sous la forme proposé par Jackendoff (1992), nous obtenons la structure conceptuelle présentée en (6) :

(6) [$_{EVENT}$ Go ([$_{THING}$ Bill], [$_{PATH}$ to ([$_{PLACE}$ in ([$_{THING}$ House])])])]



Spécifiquement, la catégorie sémantique *EVENT* présente un événement sous la forme d'une structure d'arbre. Cette représentation permet aussi de voir les relations qui existent entre les différents éléments. Dans cette structure, nous voyons que le nœud mère définit le type de

constituant. Le nœud le plus à gauche, qui se retrouve à être aussi la tête du constituant, représente la fonction (Saeed, 2013), c'est-à-dire le verbe ; les autres deux nœuds sœurs (*thing, path*) sont ses arguments. De cette façon cette structure représente l'événement en considérant les arguments du verbe.

Les structures conceptuelles de Jackendoff (1992) ont une syntaxe propre : les catégories sémantiques sont construites par des éléments plus petits à partir de règles de combinaisons (Saeed, 2013). Cela s'approche, grosso modo, à la structure argumentale du verbe.

Traditionnellement, la linguistique divise la phrase en deux parties : le sujet et le prédicat, (Palmer, 1994). Le sujet se compose d'un syntagme nominal et le prédicat, d'un syntagme verbal. Cette même analyse est clairement exprimée dans *Structure syntaxiques*, Chomsky (2002) où la première règle est :

$S \rightarrow NP + VP$

Une alternative à cette façon de diviser la phrase est de concevoir la phrase comme un prédicat et un ou plusieurs arguments. Théoriquement, le prédicat exprime la relation entre les arguments (Palmer, 1994).

$ARGUMENT_1$ + PRÉDICAT + $ARGUMENT_2$

Où $ARGUMENT_1$ est un syntagme nominal et PREDICAT est un syntagme verbal qui contient seulement l'élément verbal. $ARGUMENT_2$ représente la partie nominale du syntagme verbal du prédicat. L'$ARGUMENT_1$, est considéré comme *externe* au prédicat. L'$ARGUMENT_2$ est considéré comme *interne* au prédicat (Palmer, 1994; Saeed, 2013).

Les deux arguments ont une relation sémantique différente avec le prédicat. Tout changement dans l'ordre des arguments provoque un changement dans la relation sémantique, et donc dans le sens de la phrase.

La règle syntaxique pour représenter la deuxième partie de la phrase devrait être comme cela :

S $\rightarrow$ NP + VP

VP $\rightarrow$ V + NP

La structure argumentale est déterminée par la syntaxe du verbe, c'est-à-dire qu'un verbe dit *intransitif* aura seulement un argument externe. Un verbe *transitif* demandera donc deux arguments, l'un externe et l'autre interne. Un verbe *ditransitif* prendrait trois arguments, un externe, et deux internes (Carnie, 2015). Généralement, la structure argumental est en correspondance avec des relations grammaticales. Aussi, l'argument externe est associé au sujet du verbe, et les arguments internes, aux compléments du verbe. Dans ce contexte, les relations grammaticales trouvent aussi leur correspondance avec les rôles thématiques de la théorie de cas de Fillmore (1967). Par exemple, Saeed (2013) observe, lors de l'articulation d'une phrase, que nous avons une tendance à placer l' *Agent* dans la position sujet, puis le *recipient, ou le bénéfactif*, le thème ou le patient apparaissent dans les positions argumentales suivantes.

Tableau 1.2 Structure d'arguments verbaux.

Classification	**Nombre d'arguments**	**Example**
Intransitif	1 argument externe	rire, arriver, tomber, courir
Transitif	1 argument externe ; 1 argument interne	aimer, manger, frapper
Ditransitif	1 argument externe ; 2 arguments internes	donner, demander, offrir

La table 1.2 montre les arguments dont un verbe a besoin. Ici il convient de ne pas confondre les adjoints avec les arguments. Les arguments correspondent à l'objet direct et indirect d'un verbe. Les adjoints sont optionnels ; ils sont, structurellement, moins attachés au verbe, (Saeed, 2013). Les adjoints généralement se trouvent dans un syntagme prépositionnel et ils apportent un peu plus de détails sur le contexte de l'action, par exemple l'endroit, ou la manière, (Saeed, 2013).

Les approches abordées dans cette section ont mis en évidence deux choses : que les événements ont une structure et que cette structure retrouve des correspondances dans la structure

argumentale du verbe. En ce sens, Di Sciullo (2013) mentionne que la structure argumentale du prédicat dénote des événements.

Toutefois, il existe des approches en syntaxe encore plus détaillées pour exposer la structure d'un événement. Par example (Pustejovsky, 1991; Tenny & Pustejovsky, 2000, 2001).

1.5.4 Représentation des événements en TALN et traitement automatique des nouvelles

Julinda *et al.* (2014) utilisent une approche de regroupement [23] pour créer un référentiel d'événements. Leur première hypothèse est que deux phrases appartiennent au même événement si celles-ci sont dans le même groupe et contiennent des marqueurs de temps qui réfèrent au même repère temporelle. Par ailleurs, pour regrouper deux phrases dans un même groupe, la similarité entre deux phrases est déterminée avec un outil d'alignement de mots qui combine une stratégie de similarité lexicale de mots et une similarité sémantique de mots dans une taxinomie. Pour être en mesure de traiter les phrases comme des faits, Julinda *et al.* (2014) filtrent les phrases portant seulement des marqueurs de temps, toutes les autres étaient ignorées. Ce filtrage engendre la perte de toute autre phrase appartenant à un événement, mais qui ne contient pas un marqueur de temps.

En utilisant le même concept de marqueurs de temps, Kessler *et al.* (2012) détectent les dates qui apparaissent dans une phrase avec un verbe factuel. Les auteurs s'intéressent en particulière à l'extraction de dates saillantes plutôt qu'à l'extraction des événements qui pourraient être associés à celles-ci. Kessler *et al.* (2012) ont noté qu'un événement important porte une date, et si celle-ci se répète à plusieurs reprises, l'événement pourrait avoir une importance majeure d'un point de vue journalistique. En outre, Kessler *et al.* (2012) mentionnent que les mesures de similarité de texte pourraient aider à trouver des phrases qui sont plus représentatives des événements. Comme Julinda *et al.* (2014), Kessler *et al.* (2012) conditionnent un événement par la présence d'un marqueur temporel. Nous croyons que cela entraîne une perte d'information au niveau particulier d'un Mi-E et de façon plus globale pour le Ma-E.

23. Clustering.

D'autres travaux ont inclus plusieurs configurations linguistiques pour extraire les événements. Par exemple Van Hage *et al.* (2011) présentent le modèle d'événement simple (SEM) pour le Web. Ce modèle représente un effort pour modéliser des événements dans tous les domaines. Van Hage *et al.* (2011) mentionnent la nécessité d'un modèle qui permet d'identifier *qui fait quoi et quand* et le rôle que chaque acteur a joué dans l'événement, ainsi que son temps de validité. Ce modèle contient trois type de structure argumentale du prédicat de la phrase qui correspond à la théorie des cas de Fillmore (1967). SEM permet la représentation des différents rôles pour un même acteur dans l'événement, mais a besoin d'un traitement linguistique poussé pour être en mesure d'identifier les rôles des acteurs des événements. De plus, Van Hage *et al.* (2011) ne traitent pas les relations qui existaient entre des sous-événements, dans notre cas, des Mi-E.

Comme plusieurs auteurs le soulignent, le traitement d'événements implique deux concepts majeurs : la similarité de texte et la structure argumentale des prédicats. L'utilisation d'un modèle basé sur la structure d'argument des prédicats aide à identifier les événements. De plus, la présence d'adjoint, dans les phrases apporte plus d'informations sur les événements. La présence des adjoints pour un même Mi-E n'est pas assurée dans les textes à comparer, il est donc important d'entreprendre une stratégie de comparaison du contenu des sources pour trouver ces différences.

Nous terminons ce chapitre avec les remarques suivantes concernant le processus de couverture, l'évaluation de la couverture et les mesures.

– À propos du processus de couverture :

Nous avons présenté la couverture d'information comme un processus générique comportant trois composantes : l'objet de référence, l'observateur et la projection. Afin d'illustrer l'application de ce processus, nous avons identifié deux scénarios distincts. Le premier concerne la production de textes par des étudiants, le second, la production de textes journalistiques. Dans le premier cas, le filtre de l'observateur, ici l'étudiant, représente la thématique choisie par

lui ou par l'enseignant. Dans le deuxième cas, ce filtre correspond aux critères du journaliste-observateur tels que la ligne éditoriale du journal ou ses opinions personnelles. Dans les deux cas, le filtre de l'observateur détermine la façon dont l'observateur génère les projections qui deviendront ainsi les objets sur lesquels le processus de couverture s'applique. Nous avons aussi noté qu'il existe une asymétrie entre l'objet de référence et ces projections, puisque l'observateur pourrait éluder, délibérément ou pas, certaines des caractéristiques de l'objet de référence. Dans le cadre du scénario deux, nous considérons qu'il existe une deuxième asymétrie entre les projections elles-mêmes, puisque la première couverture devient la seule couverture de référence sur laquelle nous pourrons appliquer une évaluation de la couverture.

En admettant ce schéma, nous avons constaté qu'il existe ainsi une première asymétrie (I) entre l'objet de référence et/ou les projection(s) résultantes. Cette asymétrie est bien mise en relief dans le cadre du premier scénario, puisque la projection de chaque étudiant (sa dissertation) couvre en partie l'objet de référence constitué par les textes de spécialité. Dans le cadre du scénario deux, nous avons noté que plusieurs projections sont possibles pour le même objet de référence, projections qui peuvent être le produit de plusieurs journalistes-observateurs ou du même journaliste-observateur. De ce contexte précis, nous reconnaissons qu'il existe une deuxième asymétrie (II) entre les projections elles-mêmes puisque la première devient en quelque sorte la projection de référence en vertu de laquelle la couverture des projections suivantes seront évaluées. C'est cette seconde asymétrie que nous avons choisie d'étudier dans ce contexte précis. Par ailleurs, il convient de noter que cette seconde asymétrie pourrait également apparaître dans le cadre de la production de textes par des étudiants dans le cas où la thématique choisie ainsi que l'objet de référence (les textes de spécialité) seraient imposés à tous les étudiants. Dans ce cas de figure, il serait éventuellement possible de définir une projection de référence (la dissertation du meilleur étudiant, par exemple) et d'évaluer la couverture des projections des autres étudiants. Nous avons plutôt opté pour étudier d'abord l'asymétrie entre chaque projection individuelle selon un objet de référence particulier à chaque étudiant. La comparaison asymétrique de type II entre les projections des étudiants n'est pas utilisée ici.

– À propos de l'évaluation (quoi mesurer) :

L'évaluation de la couverture d'information a originalement été effectuée grâce à une mesure de la quantité d'information (D'Alessio & Allen, 2000), soit en comptant le nombre de mots utilisés, soit en mesurant l'espace physique utilisé. Cette approche nous semble trop limitée, car elle se résume à associer la couverture à des unités primitives qui ne tiennent pas compte du contenu proprement dit. Étant donné la nature des textes journalistiques, les informations liées à la notion d'événement semblent fondamentales. Dans le cas des dissertations, nous avons constaté que les étudiants n'utilisent pas les expressions linguistiques comme le font les auteurs professionnels de textes de spécialité (Cortes, 2004). De plus, et comme il semble que le vocabulaire partagé entre différents domaines académiques est de taille modeste (Durrant, 2014), il convient plutôt de centrer l'analyse sur la terminologie spécifique à la thématique couverte par chaque étudiant et du domaine dans lequel cette thématique s'inscrit. Conséquemment, pour tenir compte du contenu dans le cadre de nos deux scénarios, nous concluons que l'évaluation de la couverture doit se fonder sur deux unités distinctes, soit la terminologie dans le cas des dissertations d'étudiants, et les éléments composant l'événement dans le cas des textes journalistiques.

– À propos des mesures :

Nous avons d'abord constaté que le concept d'asymétrie est fondamental en sciences cognitives pour définir plusieurs processus cognitifs tels la comparaison des objets (Tversky, 1977; Tversky & Gati, 1978). Ce concept apparaît également en linguistique (Leyton, 1992) où il sert à décrire les relations structurelles du langage (Reinhart, 1976; Chomsky, 1993; Kayne, 1994; Chomsky, 2005; Di Sciullo, 2016), notamment les relations prédicat-arguments (Di Sciullo, 2013, 2005).

Pour évaluer la couverture d'information, nous adoptons une mesure fondée sur la similarité des textes qui devra respecter l'asymétrie lors de la comparaison.

De nombreuses mesures ont été proposées dans la littérature, la plus populaire étant la mesure du cosinus de similarité. Shivakumar & Garcia-Molina (1995) mentionnent cependant que le caractère symétrique de cette mesure génère trop souvent des valeurs très faibles et ce, même si un document forme le sous-ensemble d'un autre. Nous avons d'ailleurs noté que toutes les mesures symétriques partagent cette même caractéristique, ce qui les rend inadaptées à l'évaluation de la couverture d'information. Notons de plus que toute mesure (souvent purement lexicale) conçue selon une approche géométrique telle *bag-of-words* ne considère pas la structure hiérarchique de la phrase.

En ce qui concerne les mesures taxinomiques et sémantico-syntaxiques qui ont pu être proposées, leur désavantage réside plutôt du côté des ressources et des outils linguistiques utilisées (WordNet, FrameNet, analyseurs, etc). En effet, ces ressources ne couvrent pas tous les mots de l'anglais et cette couverture devient carrément déficiente lorsque d'autres langues sont prises en compte. Le problème devient plus flagrant lorsqu'on désire employer des analyseurs syntaxiques. Pour certains textes plus complexes, les analyseurs n'offrent pas la stabilité requise à un traitement rapide et ce, même en anglais.

Pour ce qui est des mesures asymétriques qui ont pu être proposées, elles présentent le désavantage de ne pas considérer l'ordre des mots dans la phrase. Cela se comprend puisque ces mesures ont surtout été utilisées pour la détection du plagiat (Shivakumar & Garcia-Molina, 1995; Brin *et al.*, 1995; Bao *et al.*, 2003) ; dans ce contexte d'applicaton précis, les critères de similarité sémantiques sont peu importants puisque c'est la détection d'une copie, c'est-à-dire un partage de la même surface lexicale, qui est déterminante ici. Dans notre cas, la prise en compte de l'ordre de mots dans le cas du scénario deux devient essentielle puisque cela permet, dans une certaine mesure, d'identifier les relations sémantiques entre les acteurs d'un événement.

CHAPITRE 2

PROBLÉMATIQUE ET OBJECTIFS

Dans ce chapitre, nous présentons d'abord la problématique, ensuite la justification de notre démarche pour limiter la portée de notre recherche. Finalement, nous présentons les objectifs.

2.1 Problématique générale

Dans la revue de la littérature, nous avons présenté la couverture d'information comme un processus à trois éléments, l'objet de référence, l'observateur et les projections (voir la figure 1.1, reprise ici en 2.1). Nous avons ainsi considéré que les projections sont asymétriques relativement à l'objet de référence, car l'observateur pourrait filtrer certaines caractéristiques lors du processus de projection. Certaines informations appartenant à l'objet de référence n'apparaîtront donc pas dans les projections.

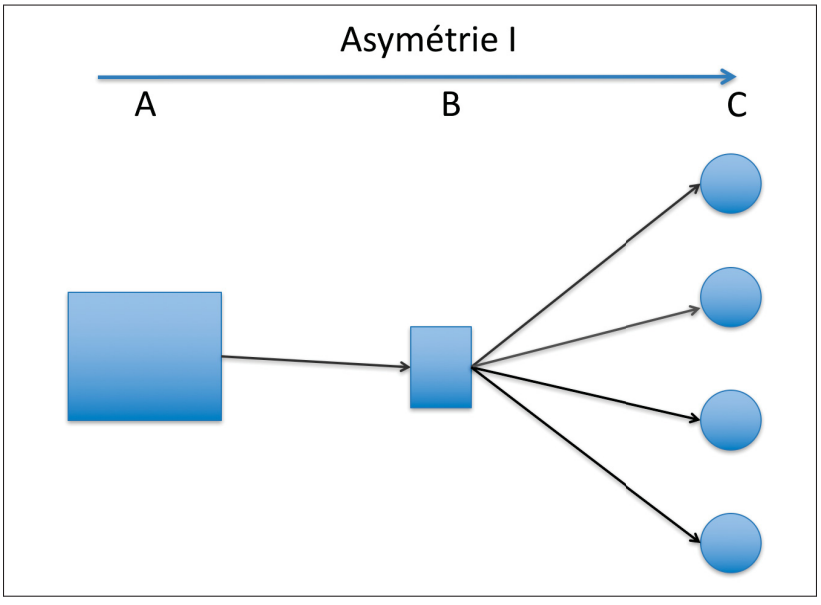


Figure 2.1 Diagramme de couverture (reprise de la fig.1.1).

Nous avons présenté deux scénarios possibles en lien avec la réalisation de ce diagramme générique. Le premier renvoie au cas où l'objet de référence est constitué de plusieurs documents

qui, une fois filtrés par l'observateur, sont transposés en une seule projection (voir la figure 1.2, reprise ici en 2.2). Nous avons rattaché ce scénario au contexte de la production de dissertations par des étudiants. Dans le cadre de ce scénario, lorsque nous parlons du référent, nous nous rapportons à l'objet de référence lui-même (constitué par les textes de spécialité). L'asymétrie qui nous intéresse particulièrement ici est celle qui existe entre ce référent et chacun des textes (projections) des étudiants et la couverture qui, naturellement, se reflète dans le texte.

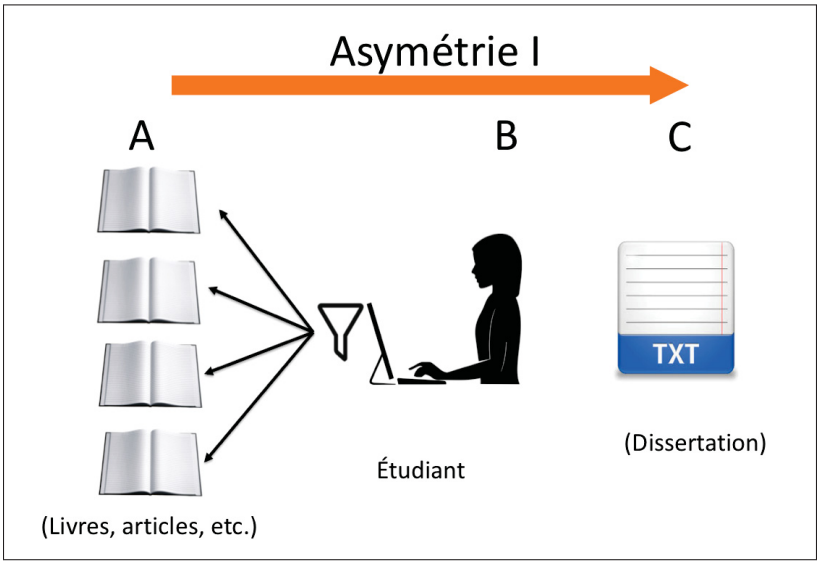


Figure 2.2 Diagramme de couverture scénario 1 (reprise de la fig. 1.2).

Le deuxième scénario renvoie au cas où un événement dans le monde constitue l'objet de référence pouvant donner lieu à plusieurs projections possibles au travers le filtre de plusieurs observateurs (voir la figure 1.3, reprise ici en 2.3). Nous avons rattaché ce scénario au contexte de la production de textes journalistiques. Dans le cadre de ce scénario, lorsque nous parlons du référent, nous nous rapportons à cette première projection, car c'est elle qui permet, dans un premier temps, de connaître l'objet de référence (l'événement réel dans le monde). L'asymétrie qui nous intéresse ici est celle qui existe :

1. entre la première projection (choisie chronologiquement ou selon d'autres critères) et les projections subséquentes

2. la couverture qui, naturellement, se reflète dans le texte

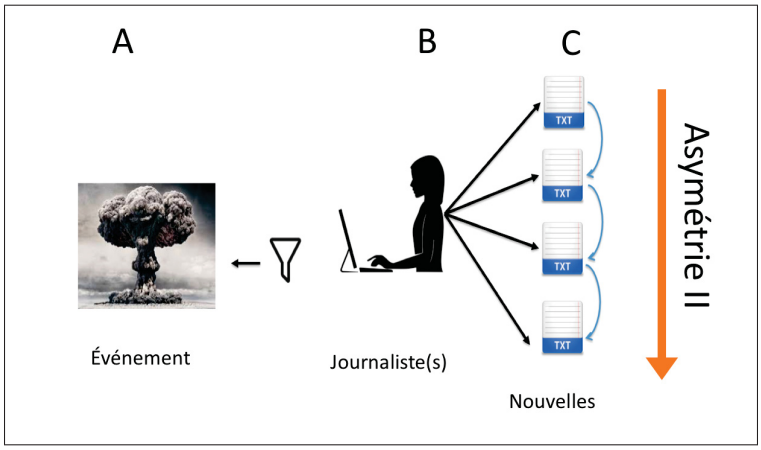


Figure 2.3 Diagramme de couverture scénario 2 (reprise de la fig. 1.3).

2.2 Objectifs

Notre objectif consiste à proposer un modèle de détection de la couverture d'information basé sur une approche asymétrique dans la comparaison de textes. Ce modèle doit :

- Évaluer la couverture d'information dans le scénario 1, pour faire resortir les rôles des objets impliqués dans la comparaison, plus spécifiquement, pour :
  a. Comprendre le rôle du référent et du sujet de comparaison.
  b. Identifier les documents des références qui ont eu le plus d'influence dans la production des écrits des étudiants.
- Évaluer la couverture d'information dans le scénario 2, pour :
  a. Identifier les projections qui apportent de la nouvelle information au référent.
  b. Construire une nouvelle plus complète.

CHAPITRE 3

MÉTHODOLOGIE

Ce chapitre présente la méthodologie que nous avons adoptée dans cette recherche. En nous appuyant sur Creswell (2013), nous amorçons ce chapitre en justifiant l'orientation de notre travail, à la section 3.1. Les sections 3.2 et 3.3 présentent la méthodologie que nous adoptons pour la couverture des productions de textes d'étudiants et les textes journalistiques respectivement.

3.1 Justification du choix méthodologique

Huff (2009, p. 18) résume l'importance de la philosophie en recherche :

> « It shapes how we formulate our problem and research questions to study and how we seek information to answer the questions [1]. »

Quatre postulats philosophiques sous-tendent toute recherche :

- Les postulats *axiologiques* abordent le sujet, ses valeurs et leur rôle dans la recherche. Dans ce cas, le chercheur reconnaît que la recherche est chargée de valeurs et que les biais sont présents (Creswell, 2013).
- Les postulats *épistémologiques* explorent la nature de ce qui compte comme connaissances, et comment les affirmations de la connaissance peuvent être justifiées (Creswell, 2013).
- Les postulats *ontologiques* rejoignent la nature de la réalité et ses caractéristiques (Creswell, 2013).
- Les postulats *méthodologiques* touchent les procédures qui sont utilisées pour conduire la recherche (Creswell, 2013).

Creswell (2013) conseille de mentionner les postulats philosophiques et les *cadres d'interprétations* utilisés. Même si nous ne conduisons pas une recherche en sciences sociales, nous

1. N.T. Elle (*la philosophie*) façonne la manière dont nous formulons la problématique et les questions de recherche à étudier ; elle détermine aussi notre façon de chercher des informations afin des répondre les questions.

considérons important d'exposer notre position dans la conduite de cette recherche. En effet, ces postulats président aux choix méthodologiques et à notre interprétation des résultats.

Creswell (2013) mentionne que les postulats philosophiques sont les premières idées qui nous viennent à l'esprit au moment d'amorcer une recherche. Toutefois, la manière dont ces postulats se rapportent à l'ensemble de la recherche demeure un mystère. Selon nous, le lien entre les postulats et la recherche ne tient pas au mystère, mais il est omniprésent et essentiel. À chacune des étapes, les quatre postulats agissent, mais certains ont une influence plus grande que les autres. L'influence d'un postulat dans une étape du projet fluctue selon la tâche.

De plus, nous présenterons les cadres interprétatifs adoptés, qui déterminent la façon dont ces postulats seront énoncés. Bien que les cadres interprétatifs en recherche qualitative soient clairement définis (voir Creswell, 2013, p. 36 - 37), nous nous contenterons de commenter le post-positivisme, auquel nous adhérons. Dans la figure 3.1, nous présentons les étapes de notre étude en lien avec notre choix méthodologique et nous expliquons ces postulats.

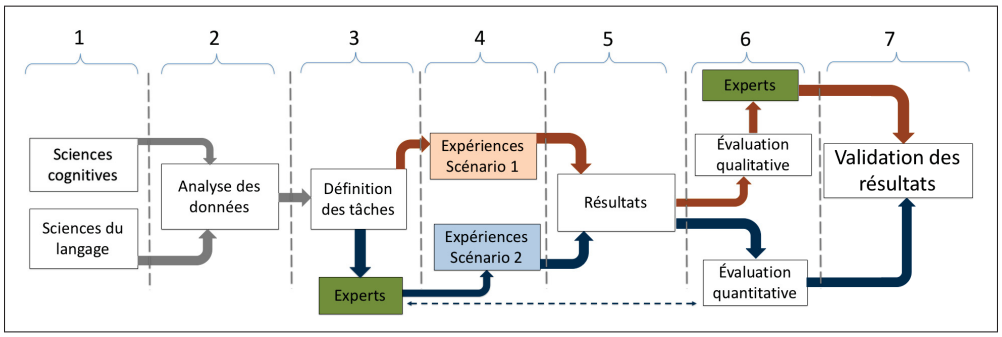


Figure 3.1 Étape présidant à notre choix méthodologique.

La figure 3.1 donne un aperçu général des étapes suivies dans le cadre de cette recherche. À l'étape 1, nous prenons en compte les théories en sciences cognitives et en linguistique. Ici, nos inquiétudes répondent plus aux postulats ontologiques, car ces théories constatent une réalité sur la présence de l'asymétrie dans la langue et dans le processus de comparaison chez l'être humain.

En ce qui concerne l'étape 2, les postulats ontologiques sont prédominants, car nous analysons les caractéristiques des données traités à l'étape 1. Ici notre intérêt porte sur l'analyse de la structure de phrases, et comment celle-ci peut avoir un impact dans un processus de comparaison.

Pour l'étape 3, les postulats méthodologiques sont prédominants, notre intérêt étant de concevoir une tâche répondant aux impératifs en lien avec notre problématique.

Voici la description de la tâche dans le scénario 1 et celui du scénario 2 :

(1) *Pour une seule projection (dissertation d'un étudiant) qui agit comme étant l'objet de comparaison, trouver tous les référents (documents des références de l'étudiant) qui ont le plus d'influence sur la projection.*

(2) *Pour la première projection d'un Ma-E ayant le rôle de référent, trouver tous les Mi-Es (phrases) apportant une information nouvelle dans un groupe de projections qui représenteront, chacune à leur tour, l'objet de comparaison*.

Dans le scénario 2, nous ferons appel à l'intervention d'un groupe d'experts qui sera prise en compte dans l'étape 6.

À l'étape 4, le scénario 1 et le scénario 2 présentent des configurations différentes, répondant en effet à des tâches distants. Ici ce sont les postulats méthodologiques auxquels nous prêtons plus d'importance puisque chaque scénario propose une façon d'élaborer des expériences selon les besoins spécifiques à chaque scénario.

À l'étape 5, nous validerons les résultats à partir des hypothèses par les deux scénarios. Ici le postulat épistémologique est proéminent, car les données serviront à construire la représentation recherchée. Chacun de nos scénarios obtiendra de l'évidence empirique après des expériences, mais pour être en mesure de créer la connaissance nous les traitons séparément.

Aux étapes 6 et 7, nous procéderons à l'évaluation des interprétations obtenues dans le scénario 1 (la production de textes d'étudiants). Ces interprétations seront soumises à une évaluation qualitative par les experts (les enseignants responsables du cours). Nos intérêts lors de cette étape concernent le postulat axiomatique puisque ce sont les valeurs du chercheur qui sont en jeu dans l'interprétation et dans l'évaluation de textes. Dans ce contexte le post-positivisme implique que la connaissance antérieure et les valeurs du chercheur aient une influence sur la recherche (Creswell, 2013). Notre interprétation est engagée dans un processus de va-et-vient avec l'évaluation. Il s'agit donc d'une négociation de l'interprétation entre nous et les enseignants. Nous reconnaissons la présence d'un biais sur notre interprétation ainsi que dans sa validation. Cependant, nous nous engageons par cette négociation à réduire, dans la mesure du possible, la présence du biais.

Le scénario 2, les résultats sont soumis à une évaluation quantitative. Leur interprétation démontrera le degré de précision de notre proposition. Ici, il n'y a pas de processus de négociation de l'interprétation des résultats entre nous et des experts. Toute interprétation voit nécessairement son point de repère dans l'intervention des experts, mais dans le scénario 2, l'interprétation se fait en amont (étape 3 dans le diagramme de la fig. 3.1) alors que dans le scénario 1, elle se fait en aval.

Le reste de ce chapitre est divisé en deux grandes sections correspondant à chacun de nos deux scénarios. D'abord, à la section 3.2, nous abordons le scénario 1 qui correspond à la production de textes d'étudiants. Dans cette section, nous utilisons un cas dans le dommaine *l'analytique des apprentissages* (LA) où nous analysons la couverture des concepts des références dans les dissertations des étudiants. Tel que mentionné par Tversky (1977), le choix du référent dépend de la proéminence des caractéristiques des objets. Dans ce cas-ci, nous interprétons chacune des références comme le référent. Plus particulièrement, nous considérons la terminologie utilisée dans les documents professionnels comme les caractéristiques les plus proéminentes auxquelles Tversky fait référence. Idéalement, l'écrit d'un étudiant devrait démontrer la compréhension d'un texte des références en maîtrisant la terminologie et, de ce fait, en réalisant une bonne couverture de cette dernière. Si l'on formule un énoncé de comparaison,

on dirait que *le texte d'un étudiant ressemble aux textes des documents cités dans les références* et non vice versa. Nous montrons ainsi que le sujet de comparaison est le texte de l'étudiant et le référent est tout texte provenant des références.

À la section 3.3, nous analysons la couverture d'information des textes de nouvelles dans un contexte particulier de biais. Le cas des nouvelles est particulier puisque les critères pour définir le référent sont intrinsèques aux lecteurs. Nous considérerons comme le référent la première nouvelle sur une thématique ayant apparu dans le temps. La deuxième nouvelle apparue est considérée comme le sujet de comparaison et ainsi de suite pour toute autre nouvelle. Dans ce cas, le référent est une entité de taille variable ; à mesure que de nouvelles informations sont repérées, celles-ci sont incluses dans le corps de référent qui est graduellement augmenté.

3.2 Scénario 1 : couverture d'information dans les dissertations des étudiants

La section 3.2.1 présente la construction et le prétraitement du corpus. La section suivante (3.2.2) présente les mesures de similarité choisies, comprenant les mesures utilisées comme référentiel et la mesure de couverture que nous proposons. Les valeurs issues de ces mesures sont ensuite utilisées pour aligner les dissertations et les références (section 3.2.3). Finalement, la méthodologie d'évaluation des résultats est présentée à la section 3.3.6.

3.2.1 Données et prétraitement

Cette partie de notre recherche a été réalisée en collaboration avec l'université Stoas - Université Vilentum des Sciences appliquées . Ce groupe de chercheurs (sous la direction du professeur Frank De Jong) a développé une expertise en Construction des Connaissances [2]. Ils souhaitaient utiliser une mesure automatique pour refléter comment les étudiants utilisent les références de leur programme spécialisé en pédagogie universitaire. L'utilisation des dissertations écrites par les étudiants a fait l'objet d'un certificat d'éthique de l'université. Nous avons eu accès aux dissertations annonymisées.

2. Ceci est le terme français du mot anglais Knowledge Building.

Nous avons recueilli quatre dissertations écrites {1, 2, 3, 4} par des étudiants dans le cadre du cours de *construction des connaissances* offert dans le cadre du programme de maîtrise en apprentissage et innovation. Les dissertations ont été produites par trois étudiants que nous nommerons A, B et C. L'étudiant A est l'auteur des dissertations 1 et 2. Les étudiants B et C ont produit les dissertations 3 et 4 respectivement. Chaque étudiant pouvait choisir librement le sujet de sa dissertation. L'étudiant A traite de la *motivation*, l'étudiant B, de la *construction des connaissances*, et l'étudiant C, de la *curiosité*. Dans le cadre de notre recherche, ces quatre dissertations constituent nos sujets de comparaison.

Pour former le corpus du référent, nous utilisons les documents qui faisaient partie du syllabus et aussi les références utilisées par les étudiants dans leurs dissertations. Puisque nous considérons nécessaire la distinction de l'ensemble des références, nous présentons le diagramme de la distribution des références dans la fig. 3.2. Ce corpus comporte :

1. Les *références générales* (RG) qui ont été suggérées par les enseignants et qui portent sur une thématique générique du contenu du cours. Dans la figure 3.2, cet ensemble de documents est représenté par le cercle bleu. Toutes les RG ne sont pas nécessairement citées dans les dissertations.
2. Les *références individuelles* (RI) présentent les différents sujets traités dans les dissertations (choisies par les étudiants). L'ensemble de ces références est représenté dans la fig.3.2 par le cercle jaune et vert. Les RI sont nécessairement citées dans les dissertations des étudiants. Le diagramme de la fig. 3.2 présente cette intersection entre RC et RI.
3. Les *références considérées* (RC), l'intersection de RC et RI celles-ci ont été considérées lors d'une évaluation de la part des étudiants pour connaître leur degré d'utilisation. Elles sont aussi identifiées dans le diagramme de la fig. 3.2.

Le reste des documents des RI qui ne sont pas des RG constituent les *références spécialisées* (RS) des dissertations. Cet ensemble de références est le reste du cercle jaune qui n'est pas en contact avec le cercle bleu du diagramme de la fig.3.2.Toutes les références, RG et RI

sont rédigés en anglais et leur taille varie entre 6 et 244 pages. Elles se composent d'articles scientifiques, de chapitres de livres et de livres entiers.

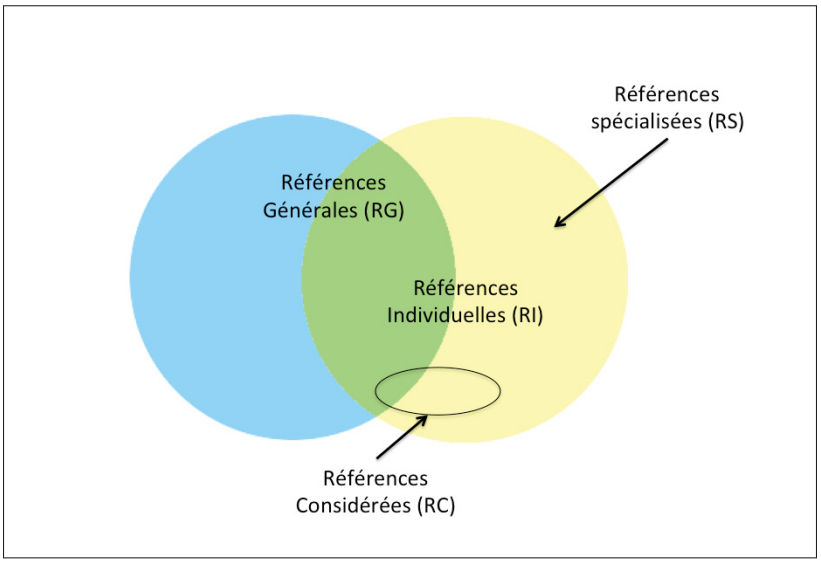


Figure 3.2 Diagramme de la distribution des documents des références et leur nomenclature utilisée.

Puisque les étudiants ont rédigé leur dissertation en néerlandais, et que les textes du référent sont rédigés en anglais, nous avons choisi d'utiliser Google Translate pour effectuer la traduction du néerlandais à l'anglais. Bien entendu, l'outil n'est pas aussi efficace qu'un traducteur professionnel humain. Puisque notre étude concerne la couverture de la terminologie et que celle-ci s'effectue grâce à une mesure qui n'utilise pas les informations concernant le style d'écriture, le recours à Google Translate se justifie (Steinberger, 2012).

Pour traiter les dissertations et les références équitablement, nous avons divisé les textes en paragraphes de 10 phrases [3], ce qui nous a permis de traiter des textes de longueurs similaires. Nous avons également extrait une liste de mots-vides, qui est disponible sur python version 2.7.

3. Nous avons utilisé la librairie pracnlptools pour séparer le texte en phrases. Nous avons décidé d'utiliser le concept de phrase à la place des lignes de texte, car une ligne peut autant contenir plus d'une phrase qu'une phrase incomplète.

3.2.2 Mesures de similarité lexicale et de couverture

Nous avons utilisé deux stratégies pour calculer la similarité lexicale. D'abord, nous utilisons la mesure du cosinus comme *référentiel* (voir formule 1.2 reprise aussi en 3.1). De plus, nous avons aussi utilisé une autre mesure de similarité lexicale basée sur le coefficient de Dice (formule 1.9 reprise en 3.2).

$$Sim\_cosinus(\overrightarrow{x}, \overrightarrow{y}) = \frac{\overrightarrow{x} \cdot \overrightarrow{y}}{\| \overrightarrow{x} \| \| \overrightarrow{y} \|} \tag{3.1}$$

$$Sim\_Dice(A,B) = \frac{2 \mid A \cap B \mid}{\mid A \mid + \mid B \mid} \tag{3.2}$$

Comme Tversky le souligne, la similarité implique une relation asymétrique, où le sujet de comparaison est semblable au référent, et non vice versa. Rappelons que le référent a plus de caractéristiques que le sujet de la comparaison. Pour nous, le sujet de la comparaison est la dissertation, composée d'un ou plusieurs paragraphes, alors que le référent est tout texte des RG ou des RS. Nous supposons que le texte d'une référence est plus riche dans l'utilisation de termes spécialisés (caractéristiques). Considérant cela, nous évaluons la couverture de la terminologie et des concepts des RG ou des RS.

Dans le cadre de cette recherche, nous présentons un coefficient de couverture asymétrique. Nous nous sommes inspirés du travail de Mihalcea *et al.* (2006). Nous avons utilisé le premier terme de la formule que les auteurs ont proposée afin de maintenir la direction asymétrique de la comparaison. La formule finale est présentée en 3.3 :

$$couverture(R,S) = \frac{\sum_{w \in \{R\}} maxSim(w,S) * idf(w)}{\sum_{w \in \{R\}} idf(w)} \tag{3.3}$$

Où R est le référent et S est le sujet d'une comparaison. La fonction $maxsim(w,S)$ calcule la similarité du mot w avec tous les mots contenus dans S. De ce fait, cette fonction sélectionne

le mot w ayant la valeur maximale de similarité avec les mots de S. Le mot w est utilisé pour construire un diagramme illustrant la relation entre le sujet de comparaison et le référent (voir 4.1.3). De plus, la fonction $maxsim(w,S)$, dans la formule 3.3, peut utiliser toute mesure de similaritié disponible sur WordNet ; nous utilisons la mesure de similarité de Lin (1998). La partie $idf(w)$ de la formule, correspond à la valeur *inverse document frequency* du mot w. Nous appellerons cette mesure *Asymmetric Coverage Hybrid Measure* ou ACHM.

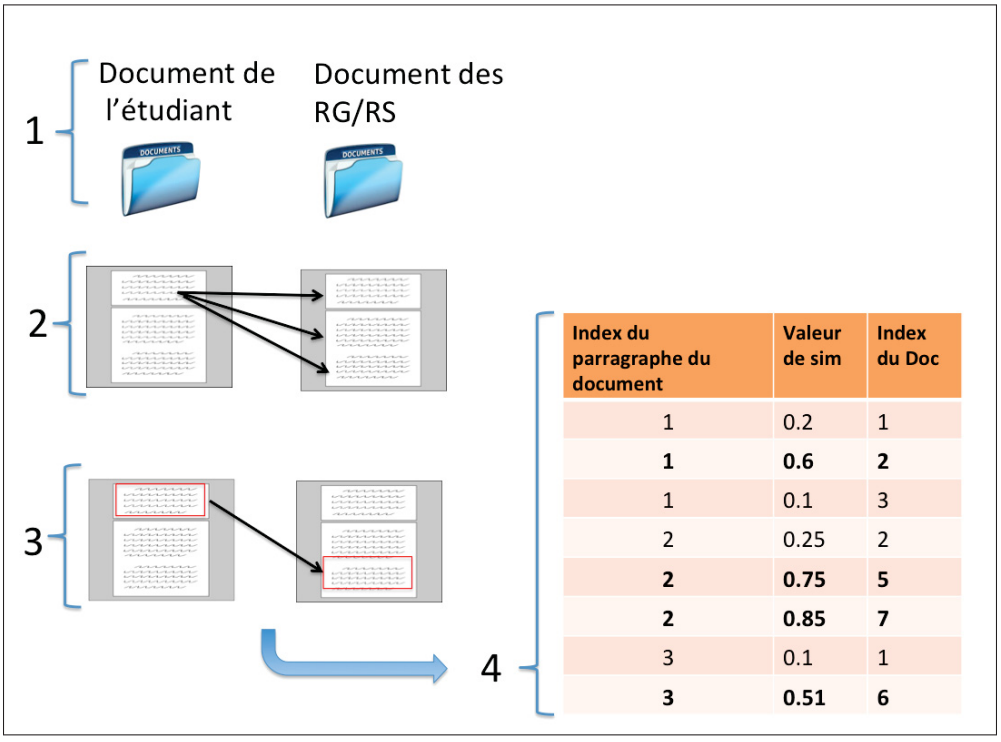
3.2.3 Alignement des dissertations par rapport aux RG et aux RS

Notre méthode est inspirée de la méthode d'alignement du texte proposée dans (Beamer & Girju, 2009) ; les auteurs alignent des diapositives à des documents scientifiques en utilisant la similarité cosinus entre les segments de texte. Par la suite, ils fixent un seuil sur la valeur de similarité pour déterminer l'alignement. Notre approche s'inspire de Beamer & Girju (2009) puisque nous alignons des paragraphes avec des documents plus longs. Nous utilisons les mesures lexicales mentionnées à la section 3.2.2 et la mesure ACHM.

Étant donné notre intérêt d'évaluer l'influence du texte des références sur les textes des étudiants, nous avons aligné les paragraphes des dissertations avec chaque document des RG et RS (aussi décomposés en paragraphes courts). Chacun de ces paragraphes a été construit en utilisant 10 phrases. La figure 3.3 présente le diagramme de la méthode que nous avons implémentée pour réaliser l'alignement des dissertations avec les RG et les RS. Nous avons calculé la similarité entre ces segments de texte afin de préserver l'équilibre relatif entre les deux sources. Cela signifie que pour chaque paragraphe dans une dissertation, nous calculons sa similarité avec tous les paragraphes d'un document provenant des RG ou RS (voir figure 3.3-2). Ensuite, nous sélectionnons le paragraphe du document des RG ou des RS qui a la valeur de similarité maximale avec ce paragraphe de la dissertation en question (voir figure 3.3-3). Le paragraphe de la dissertation compte ainsi sur un paragraphe candidat (provenant d'un document des RG ou RS) à aligner. Nous construisons un tableau avec tous les documents candidats des RG ou du RS (voir 3.3-4).

Pour déterminer si un paragraphe de la dissertation a été aligné avec un document des RG ou des RS, nous avons fixé un seuil. Si la valeur de similarité dépasse ce seuil, nous considérons que le paragraphe de la dissertation est aligné avec ce document des RG ou RS. Par conséquent, le paragraphe d'une dissertation pourrait, bien entendu, être aligné avec plus d'un document des RG ou RS.

Par exemple, à la figure 3.3-4, si nous fixons un seuil à 0.5 ; le paragraphe 1 de cette dissertation (première colonne du tableau) serait aligné avec le document 2 (troisième colonne) des RG ou des RS puisque ces deux éléments partagent une valeur de similarité de 0.6 (deuxième colonne). Tous les autres documents seront ignorés, car ils ont une valeur inférieure au seuil. Si nous regardons le tableau, nous remarquons que le paragraphe 2 a une valeur supérieure au seuil pour les documents 5 et 7, donc le paragraphe 2 sera aligné avec ces deux documents des RG ou des RI.



Index du parragraphe du document	Valeur de sim	Index du Doc
1	0.2	1
1	**0.6**	**2**
1	0.1	3
2	0.25	2
2	**0.75**	**5**
2	**0.85**	**7**
3	0.1	1
3	**0.51**	**6**

Figure 3.3 Diagramme présentant le processus de calcul de couverture et l'alignement des paragraphes d'une dissertation en lien avec documents des RG ou RS.

3.2.4 Évaluation

Étant donnée l'origine des données, nous ne pouvons compter sur la présence d'un *ground truth* pour évaluer la performance de notre approche. Nous avons donc décidé de nous inscrire dans une position de recherche qualitative à la place d'une évaluation quantitative. Nous avons choisi de conduire un Processus de Vérification par les Membres (PVM) [4] afin d'obtenir des évaluations qualitatives de nos résultats. Le PVM est une forme de validation d'études de cas. Il permet aux participants de vérifier et possiblement d'influencer les descriptions de cas ou les interprétations (Bygstad & Munkvold, 2007). Dans les œvrages consultées , il existe des avantages et des désavantages à l'utilisation des PVM. Nous analysons ces postures et nous présenterons, par la suite, la nôtre.

- Bygstad & Munkvold (2007) mentionne que les PVM peuvent renforcer la validité des études de cas.
- Le PVM permet aussi la génération de plus de données, qui plus tard pourront être utilisées dans l'étude (Bygstad & Munkvold, 2007).
- Le PVM peut être aussi vu comme une sorte de négociation de l'interprétation des résultats (Bygstad & Munkvold, 2007).
- Il existe une possible influence d'une rationalisation *post-hoc* [5].
- Le degré de langage utilisé par les participants et le chercheur pourrait aussi donner des interprétations erronées (Bygstad & Munkvold, 2007).
- Une barrière importante est posée par les différences culturelles, le background et la terminologie. Dans le cas des études en systèmes d'information, ce problème est moins présent que dans les sciences sociales (Bygstad & Munkvold, 2007). La pertinence des PVMs dans les projets en génie ou plutôt en études de système d'information a été d'ailleurs mis en évidence par Dubé & Paré (2003).
- Morse (1998) signale comme problème principal que les résultats ont été synthétisés, décontextualisés et abstraits des participants individuels. Cela ne permettrait pas à un parti-

4. Ceci est le terme en français pour le mot anglais *Member Check*.
5. Il s'agit d'un sophisme qui prend la cause comme un antécédent à un événement.

cipant en particulier de faire un lien entre ses propres expériences et les résultas dans le PVM.

Nous avons décidé d'utiliser le PVM comme instrument d'évaluation parce qu'il semble être adapté pour son application aux systèmes d'information, (Dubé & Paré, 2003). De plus, les auteurs suggèrent que le format d'un PVM devrait être conduit sous la forme d'une présentation formelle avec les participants pour corroborer l'évidence de l'interprétation des résultats. Ceci permet aussi une négociation de l'interprétation des résultats entre le chercheur et les participants du PVM.

Nous avons donc préparé une séance avec deux des enseignants du cours. Cette séance a consisté en une présentation des résultats sous la forme de trois types de visualisations qui condensaient les résultats. La présentation contenait une introduction aux contextes de la production des données et des concepts techniques nécessaires pour comprendre la méthodologie utilisée. À chaque fois qu'une visualisation était présentée aux enseignants, nous consacrions une pause pour discuter et analyser nos interprétations. De la même façon, lors de cette pause nous avons confronté les enseignants avec d'autres interprétations possibles des résultats. Toute la séance a été enregistrée en commun accord avec les participants.

3.3 Scénario 2 : couverture d'information de textes journalistiques

Pour cette partie de notre recherche, nous avons utilisé une partie du corpus de la conférence *Text REtrieval conference* (TREC). L'une des tâches de cette conférence était dédiée à la détection de nouveauté (Novelty TREC) proposée de 2002 à 2004, année où la conférence a abandonné cette tâche, possiblement à cause des faibles résultats obtenus par la majorité des participants (voir Harman, 2002; Soboroff & Harman, 2003; Soboroff, 2004). À la section 3.3.1, nous abordons la description de la tâche Novelty TREC et la composition du corpus. Nous avons mené de notre côté un étiquetage d'une partie du corpus original. Nous discutons cette démarche à la section 3.3.2.

Finalement, à la section 3.3.3, nous présentons les patrons linguistiques qui tentent de capturer la structure d'un Mi-E. Ces patrons seront utilisés dans notre mesure de couverture d'information décrite à la section 3.3.4. La section 3.3.5 présente notre expérimentation et notre méthode d'évaluation des résultats.

3.3.1 Remarques sur le corpus TREC

Originalement, le *novelty track corpus* est constitué de 50 thématiques différentes dont 25 sont des nouvelles de type narratif, et 25 sont des nouvelles de type opinion. La collection de données provient de trois sources de nouvelles différentes couvrant les mêmes périodes :

- Le *New York Times Service*, de juin 1998 à septembre 2000.
- Le *Associated Press (AP)* de juin 1998 à septembre 2000.
- Le *Xinhua News Service* de janvier 1996 à septembre 2000.

Cette juxtaposition de sources dans le temps implique une plus grande redondance, donc moins d'information nouvelle à identifier. Ceci a été reporté comme une condition augmentant le réalisme de la tâche (Soboroff & Harman, 2005).

Pour réaliser les expériences sur la couverture d'information de textes journalistiques, nous avons d'abord réalisé l'étiquetage d'une partie du corpus *TREC*. Nous avons pris cette décision pour deux raisons : premièrement les critères sous lesquels le corpus a été annoté ont été déjà mis en question par d'autres travaux (Schiffman, 2002; Tsai & Chen, 2002; Collins-Thompson *et al.*, 2002). De plus, les basses performances des participants lors des compétitions en 2002 (Harman, 2002), 2003 (Soboroff & Harman, 2003) et surtout en 2004 Soboroff (2004) nous laissent croire que des problèmes dans la définition de la tâche ont eu un impact sur les résultats. La dernière compétition *Novelty TREC* fut conduite en 2004 (Soboroff, 2004).

La définition de la tâche générale était la suivante[6] : *"given a topic and an ordered set of documents segmented into sentences, return sentences that are both relevant to the topic and novel given what has already been seen" (Soboroff, 2004)*. Ainsi, les participants devaient faire

6. Instructions disponibles sur : http ://trec.nist.gov/data/t12_novelty/novelty03.guidelines.html

face à deux problèmes différents : d'abord l'identification des phrases pertinentes et, parmi ces dernières identifier les phrases apportant de la nouvelle information. À partir de cette tâche générale, quatre tâches découlent de la compétition *novelty track* :

1. *"Given the set of documents for the topic, identify all relevant and novel sentences,* (Soboroff, 2004).
2. *"Given the relevant sentences in all documents, identify all novel sentences,* (Soboroff, 2004).
3. *"Given the relevant and novel sentences in the first 5 documents only, find the relevant and novel sentences in the remaining documents. Note that since some documents are irrelevant, there may be any relevant or novel sentences in the first 5 documents for some topics,"* (Soboroff, 2004)
4. *"Given the relevant sentences from all documents, and the novel sentences from the first 5 documents, find the novel sentences in the remaining documents*" (Soboroff, 2004)

Les descriptions des tâches 2 et 4 auraient pu bien correspondre à la définition de notre tâche puisqu'elles considèrent un ensemble de phrases pour déterminer la pertinence et la nouveauté. Cet ensemble de phrases peut bien correspondre à la formation d'un référent ; le problème de compatibilité réside dans la composition du référent. Dans le cas de la compétition *novelty track* celui-ci est conçu en termes de pertinence tandis que nous prenons le contenu de la première nouvelle, d'une thématique pour construire le référent. Nous reformulons les tâches 2 et 3 de la compétition *novelty TREC* en utilisant la terminologie [7] employée en (3).

(3) *Pour la première projection d'un Ma-E ayant le rôle de référent, trouver tous les Mi-Es (phrases) apportant de la nouvelle information dans un groupe de projections qui représenteront, chacune à leur tour, l'objet de comparaison.*

Nous constituons le référent en termes chronologiques puisque nous considérons la première projection d'un évènement (qui se trouve à être une nouvelle) comme le seul objet existant

7. Il s'agit de la tâche que nous avons définie précédemment pour le scénario 2 en (2).

dans un premier temps ; c'est le seul objet comprenant les informations d'un événement. Par la suite, les autres projections de ce même évènement auront le rôle d'objet de comparaison que nous utiliserons pour déterminer ce qui est nouveau.

Nous croyons que pour la compétition *novelty* TREC, il y a eu deux principaux problèmes qui ont augmenté la complexité de la tâche. Le premier problème est relié au concept de pertinence puisqu'il doit exister un consensus sur ce qui est pertinent pour la création d'un *ground truth* qui permettrait d'évaluer des résultats. De plus, ce consensus devrait être connu par les concepteurs de l'algorithme qui résoudra la tâche.

En laissant de côté la tâche algorithmique, à laquelle les compétiteurs ont dû faire face, nous voyons une déficience lors de la formation du *ground truth*. Selon les compte-rendus de la compétition *novelty track* (Harman, 2002; Soboroff & Harman, 2003; Soboroff, 2004), les annotateurs (groupe d'experts) devaient déterminer ce qui est pertinent, puis ce qui est nouveau.

Notre première observation par rapport à la façon d'étiqueter le corpus est que la pertinence est une caractéristique qui dépend des critères propres aux utilisateurs ; elle peut être très variable d'un utilisateur à l'autre. Ces critères sont déterminants pour réaliser l'étiquetage du corpus.

Notre deuxième observation est que si les critères déterminant la pertinence d'un document ne sont pas définis ou standardisés pour les annotateurs, nous ne pouvons pas faire confiance à l'accord (bas) qui pourrait résulter de tant de malentendus sur ces critères. Par conséquent, l'évaluation des résultats des algorithmes est difficile, car ces derniers sont conçus selon des critères différents des attentes de la tâche qu'ils sont supposés résoudre.

D'ailleurs, dès la première année de la compétition, quelques participants ont supposé que la difficulté de la tâche était, peut-être, due à la subjectivité du processus d'annotation et à la pertinence des phrases (voir Schiffman, 2002; Tsai & Chen, 2002; Collins-Thompson *et al.*, 2002).

En ce qui concerne l'étiquetage du corpus, pour créer le *ground truth* qui servirait à l'évaluation des résultats des participants de la conférence *TREC novelty track* 2004, la tâche fut réalisée de la façon suivante :

- Des annotateurs NIST [8] ont d'abord réalisé la tâche.
- À partir d'un ensemble de documents, l'annotateur sélectionne les phrases pertinentes, puis parmi celles-ci, sélectionne les phrases pertinentes qu'il considère comme nouvelles.
- Chaque thématique fut traitée séparément par deux annotateurs différents, le premier annotateur étant responsable de déterminer le sujet de la thématique et chercher les documents qui la conforment ; le deuxième annotateur, responsable de réaliser l'étiquetage.

Le deuxième problème relié à la réalisation de la tâche de la compétition TREC est la quantité de documents recueillis pour chaque thématique que les annotateurs doivent étiqueter. Par défaut une thématique contient 25 nouvelles, mais certaines atteignent jusqu'à 69 nouvelles. Toute cette information a dû être analysée par un seul annotateur afin de détecter ce qui est pertinent. Nous croyons qu'une telle quantité d'information représente une surcharge cognitive pour l'annotateur et que la fatigue aurait aussi pu jouer un rôle (Baddeley & Hitch, 1974; Miller, 1956; Shiffrin & Nosofsky, 1994; Ma *et al.*, 2014). Voir Soboroff (2004), pour les informations sur la quantité de phrases pour chaque thématique.

Le corpus TREC fournit une liste de phrases nouvelles et de phrases pertinentes pour chaque thématique, ce qui permet de réaliser une évaluation. Cette liste a été modifiée durant les trois années d'existence de la tâche. Nous avons exploré cette liste et nous avons remarqué que, par exemple, pour la thématique N80, il y a seulement quarante-huit phrases qui ont été étiquetées comme étant nouvelles. Cette thématique aborde les tremblements de terre qui ont eu lieu entre 1998 et 1999 en Turquie. Il y en a 25 nouvelles couvrant une période de deux ans avec des séismes différents, pour un total de 447 phrases. Les annotateurs ont annoté 104 phrases pertinentes et seulement 51 phrases nouvelles. Ce chiffre nous semblait incohérent par rapport au nombre total de phrases disponibles pour la thématique puisque nous ne considérons pas la pertinence pour déterminer la nouveauté des phrases. Après une analyse du processus

8. Des annotateurs NIST sont des experts du *National Institute of Standars and Technology*, qui est un organise qui prépare des données pour leur utilisation dans les domaines universitaires ou industriel ou l'industrie.

d'étiquetage du corpus, nous avons conclu que l'étiquetage original ne s'accordait pas à nos besoins.

3.3.2 Étiquetage du corpus

Pour réaliser l'étiquetage du corpus, nous avons demandé la participation de quatre experts que nous présentons maintenant. Le profil des experts est le suivant :

- L'un est un étudiant au baccalauréat en linguistique.
- Les trois autres experts sont des étudiants au doctorat possédant des connaissances en TALN.

Tous les experts ont été formés pour réaliser la tâche. Le temps de réalisation de celle-ci est évalueé à environ une semaine, c'est-à-dire cinq jours ouvrables, plus une fin de semaine. La transcription de la tâche est présentée un peu plus loin.

De l'ensemble de vingt-cinq thématiques présentes dans le corpus TREC, nous en avons choisi cinq qui correspondent aux index N55, N63, N74, N79 et N80, toutes de type narratif. Le tableau 3.1 présente un aperçu des thématiques choisies. Pour chaque thématique, nous avons choisi aléatoirement cinq nouvelles entre 1998 et 1999. La première nouvelle, c'est à dire, la plus récente, est considérée comme le référent ; le reste des nouvelles, c'est à dire les quatre nouvelles restantes sont considérées comme les sujets de comparaison.

Comme nous le mentionnons précédemment, la tâche originale exigeait une grande charge de travail cognitif (Baddeley & Hitch, 1974; Miller, 1956; Shiffrin & Nosofsky, 1994; Ma *et al.*, 2014) puisque l'étiqueteur devrait se rappeler l'information contenue dans le référent pour être en mesure de classifier les nouvelles phrases du sujet de comparaison comme des phrases nouvelles. Nous avons donc décidé de réduire la charge cognitive des annotateurs afin que la tâche soit moins exigeante et qu'ils puissent idientifier les phrases contentant de l'information nouvelle.

Pour l'étiquetage de ce nouveau corpus, nous avons procédé de la façon suivante :

Tableau 3.1 Description des thématiques choisies dans le corpus TREC

Index dans le corpus	**Index dans l'expérience**	**Nombre de phrases**	**Description**
N55	1	62	Tests nucléaires au Pakistan
N63	2	25	Embargo cubain imposé par les États-Unis
N74	3	25	Mort de la princesse Diana
N79	4	119	Vie et mort de Charles Schulz
N80	5	23	Tremblement de terre en Turquie

1. Pour chaque thématique, nous avons demandé aux participants de lire la première nouvelle qu'ils devaient, dès lors, considérer comme le référent. L'objectif était de se souvenir le mieux possible de l'information originale afin de pouvoir identifier toute nouvelle information dans les textes subséquents. Il faut souligner que, en tout temps, les participants pouvaient consulter le référent.
2. Par la suite, les experts ont dû lire le texte de la deuxième nouvelle et identifier les phrases qui apportent de l'information nouvelle.
3. À chaque fois que l'expert trouvait une phrase avec une nouvelle information, il devait d'abord l'annoter comme nouvelle, puis l'ajouter au référent. De cette façon, le référent agit aussi comme un sac cueilleur d'information nouvelle en augmentant sa taille.
4. L'expert devait aussi lire le reste des nouvelles (trois, quatre et cinq) avec les même directives (identifier la nouvelle information et l'ajouter dans le referent).

Pour faciliter l'identification des éléments qui apportent de la nouvelle information dans une phrase, nous avons d'abord identifié les constituants des phrases à l'aide d'un analyseur syn-

taxique de surface. Nous avons utilisé l'analyseur du *Cognitive Computation Group* de l'Université d'Illinois [9]. Pour la phrase (4), nous obtenons un résultat présenté en (5).

(4) The princess and Al Fayed arrived in Paris on Saturday afternoon.

(5) [NP *The princess and Al Fayed*] [VP *arrived*] [PP *in*] [NP *Paris*] [PP *on*] [NP *Saturday afternoon*].

Les constituants ont été indexés par rapport à la phrase où ils apparaissent. Par exemple, (6) montre la version indexée des constituants de la phrase (5).

(6) [NP_1 *The princess and Al Fayed*] [VP_2 *arrived*] [PP_3 *in*] [NP_4 *Paris*] [PP_5 *on*] [NP_6 *Saturday afternoon*].

Pour l'identification de la nouvelle information, nous avons demandé aux experts de créer des groupes différents avec les constituants de la phrase. La taille de ces groupes peut comprendre deux constituants ou plus. L'ordre des groupes dans la phrase originale devait être respecté. La création de ces groupes visait à tenir compte du changement d'ordre des constituants ; tout changement d'ordre serait interprété comme un nouveau Mi-E. Par exemple pour les phrases :

(7) [NP_1 *John*] [VP_2 *hits*] [NP_3 *Mary*] .

(8) [NP_1 *Mary*] [VP_2 *hits*] [NP_3 *John*].

Même si les phrases (7) et (8) partagent les mêmes éléments lexicaux, l'ordre est différent, ce qui entraîne l'interprétation d'un autre Mi-E. Les experts devaient être en mesure d'identifier ce changement de sens. De cette façon, nous établissons que la nouveauté correspond plutôt à

9. http ://cogcomp.cs.illinois.edu/page/demo_view/ShallowParse

l'ordre des éléments lexicaux et non seulement à leur présence dans un Mi-E. Nous incluons, en annexe, le document en anglais qui contient les directives pour l'étiquetage du corpus (voir l'annexe VI).

Nous avons aussi prévu les différentes dénominations pour une seule entité nommée. Par exemple, dans la thématique 1, nous trouvons les dénominations suivantes : *The United States, President Bill Clinton, President Clinton.* Ces dernières réfèrent à une même entité. Nous avons demandé aux annotateurs d'interpréter ces variantes lexicales comme la même entité nommée. Nous comprenons que cette décision allait nécesairement avoir un impact sur le calcul de similarité lexicale entre deux segments de texte, mais nous voulions préserver une interprétation la plus similaire au jugement humain.

Prenons l'exemple suivant pour montrer ce que nous attendions des annotateurs :

(9) [NP_1 *British Princess Diana*] [VP_2 *was seriously injured*] and [NP_3 *her friend Egyptian millionaire Dodi el-Fayed*] [VP_4 *was killed*] when [*NP_5 their car*] [VP_6 *crashed*] [PP_7 *in a Paris road tunnel*] [PP_8 *on Saturday night*].

(10) [NP_1 *Britain's Princess Diana*] [VP_2 *died*] [PP_3 *in hospital*] [PP_4 *after a car crash*].

Nous supposons que la phrase (9) fait partie du référent et que la phrase (10) fait partie du sujet de comparaison. La thématique de ces deux phrases est l'accident de la princesse Diana à Paris en 1998. La phrase (9) contient beaucoup de détails par rapport au Ma-E qu'elle relate. Logiquement, la phrase (10) est apparue quelques heures après l'accident de la princesse. Nous nous attendons à ce que les annotateurs identifient la phrase (10) comme apportant de la nouvelle information (non couverte par le référent) et qu'ils identifient les combinaisons d'éléments qui leur font déduire la nouveauté, dans ce cas les éléments VP_2 et PP_3.

3.3.3 Création de patrons linguistiques

Pour capturer la couverture d'information des événements dans les textes journalistiques, nous avons décidé d'utiliser des patrons linguistiques afin de modéliser, d'une certaine façon, la structure des Mi-E. Nous avons déterminé qu'un Mi-E tente de répondre aux questions : Qui a fait quoi ? À qui ? Quand ? Et où ? La réponse à ces questions nous permet d'identifier le sujet et l'objet du prédicat ainsi que quelques adjoints. Nous avons besoin d'être en mesure d'identifier ces éléments par des patrons. Ce que nous définissons comme Mi-E pourrait aussi retrouver une certaine correspondence avec la définition d'événement en linguistique théorique (Davidson, 2001; Jackendoff, 1992; Di Sciullo, 2013). Il est clair que les patrons devaient tenter de capturer les relations grammaticales, c'est-à-dire : le sujet, le prédicat, le complément direct, le complément indirect, les marqueurs spatio-temporels.

Nous avons utilisé un étiqueteur de catégories lexicales et un lemmatiseur [10]. Nous nous servons des catégories lexicales pour la création des patrons ; nous utilisons les lemmes pour capturer les variations morphologiques des mots (pluriels des noms et adjectifs, conjugaisons des verbes).

Une fois le corpus étiqueté, nous commençons par éliminer certaines catégories lexicales pour ne considerer que les noms, les verbes et les prépositions. Par exemple, nous n'utilisons pas les adverbes ; Tenny (2000) mentionne d'ailleurs que les adverbes sont périphériques à la structure argument-prédicat et qu'ils ne représentent pas une classe homogène permettant de les interpréter. Dans cette thèse, nous nous intéressons au fait qu'un événement a eu lieu, pas à la façon dont il s'est déroulé.

La conception des patrons se base sur une séquence de mots qui appartient à une même catégorie lexicale. Il s'agit donc d'une série de patrons linéaires ; la combinaison de ces patrons vise à capturer certaines relations grammaticales (voir tableau 3.2).

10. La librairie en question est PractoolNLP de python, http ://pynlpl.readthedocs.io/en/latest/.
11. En l'anglais, c'est la préposition *by*. Ce patron est facilement adaptable au français puisque c'est la préposition *par* qui introduit ce groupe.

Tableau 3.2 Patrons linguistiques en *R* et *S*

Patron en *R*	**Patron en *S***	**Interprétation visée**
N–V	N–V	Sujet : nom suivi par un verbe
V–N	V–N	Objet : verbe suivi par un nom
V–P–N	N–V	Voix passive : un verbe suivi une préposition [11], et cette dernière, suivie d'un nom.
P–N	P–N	Adjoints : instrumentaux, locatives et objets indirects.
N–V	V–N	Structure subjet - Verbe - Objet

Nous sommes conscients que ces patrons ne correspondent pas à une analyse profonde de la structure de la phrase, mais d'une façon générale ils s'adaptent à nos besoins. Nous utilisons ces patrons par les raisons suivantes :

- Les annotateurs de catégories lexicales sont performants et ils sont disponibles en plusieurs langues. De plus, ils ont une meilleure performance que les analyseurs sémantiques.
- D'un point de vue théorique, ces patrons pourraient être facilement appliqués dans un contexte multilingue avec très peu de modifications. Par exemple, changer la préposition « by » par sa correspondante en français pour la représentation de la voix passive.
- Ils peuvent capturer grossièrement les relations sujet–verbe, verbe–objet et certains adjoints.

Lors de la formation des patrons, nous procédons à la fusion des mots voisins qui ont la même catégorie lexicale, par exemple, un N suivit d'un autre N. Ceci est le cas pour les noms et les verbes. Regardons la phrase suivante :

(11) Donald Trump is desesperately building a wall in the Mexican border.

Après l'étiquetage des cartégories lexicales, nous avons :

(12) Donald$_{NN}$ Trump$_{NN}$ is$_{VBZ}$ desperately$_{RB}$ building$_{VBG}$ a$_{DT}$ wall$_{NN}$ in$_{PP}$ the$_{DT}$ Mexican$_{JJ}$ border$_{NN}$.

Après la lemmatisation, nous obtenons :

(13) Donald$_{NN}$ Trump$_{NN}$ be$_{V}$ desperately$_{RB}$ build$_{V}$ a$_{DT}$ wall$_{NN}$ in$_{PP}$ the$_{DT}$ Mexican$_{JJ}$ border$_{NN}$.

À partir de (13), nous fusionnons « Donald$_{NN}$ » et « Trump$_{NN}$ » pour créer une seule chaîne de caractères « Donald/Trump$_{NN}$ ». De la même façon, nous fusionnons « be$_{V}$ », « build$_{V}$ » en « be/build$_{VV}$ ». Nous nous inspirons ici de Turney (2012) en fusionnant différents mots d'une même catégorie lexicale en une seule chaîne de caractères. De cette façon, nos patrons tentent de capturer les catégories syntagmatiques de la phrase. Pour (13), nous obtiendrons les patrons suivants :

- Donald/Trump$_{NN}$ - be/build$_{VV}$: NN – VV
- be/build$_{VV}$ - wall$_{NN}$: VV – NN
- in$_{PP}$ - border$_{NN}$ PP – NN

3.3.4 Mesure de couverture adaptée

Nous présentons ici notre mesure de couverture d'information ; dans ce cas-ci, nous voulons vérifier que les éléments formant un Mi-E soient couverts dans les deux phrases à comparer (le référent et le sujet de comparaison). Nous incluons les patrons linguistiques décrits à la section précédente dans le calcul de notre mesure.

$$couverture(R,S) = \frac{\sum_{p\in\{R\}} MaxiSim(p,S) \times \alpha_p}{\sum_{p\in\{R\}} \alpha_p} \tag{3.4}$$

Où R est le référent, S et le sujet de la comparaison. p est un doublet de patrons décrits dans le tableau 3.2. Le paramètre α_p (première colonne du tableau 3.3) permet de pondérer, avec des valeurs différentes, chacun des doubles de patrons p_i. $MaxSim()$ est une fonction qui calcule la similarité lexicale maximale entre deux chaînes de caractères des patrons. Pour ce faire,

Tableau 3.3 Doublets de patrons linguistiques et le paramètre α

	Doublet de patrons	
α_m	**Patron en** R	**Patron en** S
α_1	N–V	N–V
α_2	V–N	V–N
α_3	V–P–N	N–V
α_4	P–N	P–N
α_5	N–V	V–N

elle identifie un premier patron en *R* (deuxième colonne du tableau 3.3), ensuite elle cherche le deuxième patron correspondant en *S* (troisième colonne du tableau 3.3) afin de créer le doublet. Cette fonction peut utiliser toute stratégie de similarité de texte ; nous avons utilisé le coefficient de Dice comme mesure de similarité lexicale. Prenons par exemple les phrases (14) et (15).

(14) British Princess Diana was seriously injured in a car crash.

(15) Britain's Princess Diana died in hospital after a car crass.

Après l'étiquetage des catégories lexicales nous avons :

(16) British$_{NN}$ Princess$_{NN}$ Diana$_{NN}$ was$_{VBD}$ seriously$_{RB}$ injured$_{VBN}$ in$_{IN}$ a$_{DT}$ car$_{NN}$ crash$_{NN}$.

(17) Britain$_{NN}$'s$_{POS}$ Princess$_{NN}$ Diana$_{NN}$ died$_{VBD}$ in$_{IN}$ hospital$_{NN}$ after$_{IN}$ a$_{NN}$ car$_{NN}$ crass$_{NN}$.

Après la lemmatisation, nous avons :

(18) British$_{NN}$ Princess$_{NN}$ Diana$_{NN}$ be$_{V}$ seriously$_{RB}$ injure$_{V}$ in$_{IN}$ a$_{DT}$ car$_{NN}$ crash$_{NN}$.

(19) Britain$_{NN}$'s$_{POS}$ Princess$_{NN}$ Diana$_{NN}$ die$_{V}$ in$_{IN}$ hospital$_{NN}$ after$_{IN}$ a$_{NN}$ car$_{NN}$ crass$_{NN}$.

Nous faisons la fusion des mots avec la même classe lexicale, puis nous choissisons seulement les mots que nous avons déterminés pour créer les patrons.

(20) British/Princess/Diana$_{NN}$ be/injure$_{V}V$ in$_{IN}$ car/crash$_{NN}$.

(21) Britain/Princess/Diana$_{NN}$ die$_{VV}$ in$_{IN}$ hospital$_{NN}$ after$_{IN}$ car/crass$_{NN}$.

Pour la phrase (20) nous avons les patrons suivants :

- British/Princess/Diana$_{NN}$ be/injure$_{VV}$
- in$_{IN}$ car/crash$_{NN}$

Pour la phrase (21) nous avons les patrons suivants :

- Britain/Princess/Diana$_{NN}$ die$_{VV}$
- in$_{IN}$ hospital$_{NN}$
- after$_{IN}$ car/crass$_{NN}$

Pour montrer le comportement de la fonction *Lexical_Sim*(), prenons le premier patron de la phrase (20) (British/Princess/Diana$_{NN}$ be/injure$_{V}V$) et le premier patron de la phrase (21) (Britain/Princess/Diana$_{NN}$ die$_{VV}$), qui correspondent aux patrons de type N–V.

La fonction calculera la similarité lexicale pour chaque chaîne de caractères qui partage la même étiquette, c'est-à-dire British/Princess/Diana$_{NN}$ de la phrase (20), et Britain/Princess/Diana$_{NN}$ de la phrase (21) pour l'étiquette NN (partie nominale). Le même processus sera appliqué pour l'étiquette VV(partie verbale). La valeur finale de la similarité entre les patrons N–V des phrases (20) et (21) comprend la moyenne de la similarité de la partie nominale et de la partie verbale.

3.3.5 Expérimentation

L'objectif de cette expérience est de mesurer la couverture d'information de la première projection (Ma-E) d'un événement dans les projections subséquentes. Notre stratégie est la suivante :

- Pour chaque thématique, nous comptons cinq nouvelles. Nous établissons comme référent la nouvelle la plus récente dans le corpus (en termes chronologiques). Les quatre autres nouvelles (Ma-Es) apparaissent aussi en ordre chronologique ; chacune est considérée, à son tour, comme un nouveau sujet de comparaison (voir la figure 3.4, pour identifier ces éléments).
- Pour toutes les phrases du référent, nous vérifions si chaque phrase de l'objet de comparaison arrive à les couvrir. C'est à cette étape que nous calculons la couverture d'information entre le référent et le sujet de comparaison avec notre mesure de couverture.
- Il convient de noter que la taille du référent augmente d'une phrase à chaque fois qu'une nouvelle information est trouvée. Ceci arrive quand une phrase du sujet de comparaison n'arrive pas à couvrir celle(s) du référent. Pour déterminer si une phrase du référent est couverte par une autre phrase du sujet de comparaison, nous établissons un seuil pour la valeur de couverture. Une phrase qui ne dépasse pas ce seuil devrait donc être ajoutée dans le référent, car elle est composée d'information nouvelle. Dans le cas contraire, nous continuons à vérifier la couverture d'information du référent avec une autre phrase du sujet de comparaison.

La figure 3.4 visualise notre méthode pour calculer la couverture d'information dans les textes journalistiques.

3.3.6 Evaluation

Pour évaluer la performance de notre proposition, nous devons d'abord évaluer l'accord entre les annotateurs. Au début, nous avons choisi d'utiliser *Kappa Fleiss* et *ICC* [12] comme coefficient pour exprimer l'accord entre plus de deux annotateurs, mais nous avons été rapidement

12. Interclass correlation coefficient.

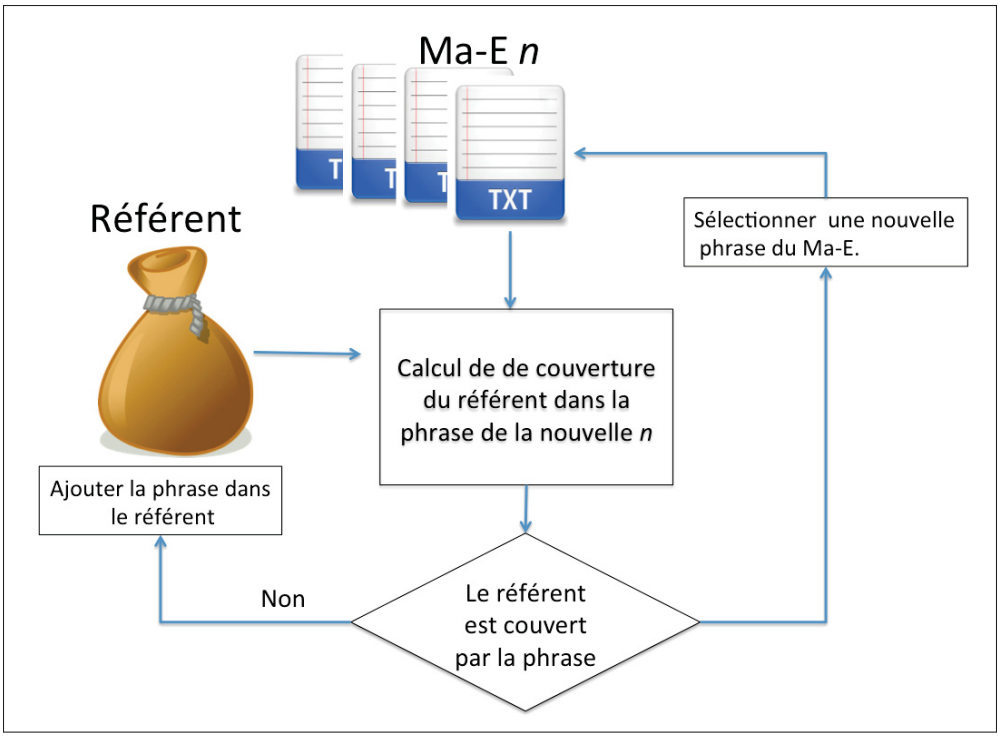


Figure 3.4 Diagramme du calcul de couverture de textes journalistiques

confronté à un problème ; les valeurs de ces coefficients sont très bas même si l'accord entre les annotateurs est élevé. Nous faisons face à deux des paradoxes du coefficient *Kappa*, révélés par Feinstein & Cicchetti (1990).

Le coefficient *Kappa* se calcule de la façon suivante :

$$\kappa = \frac{p_o - p_e}{1 - p_e} \tag{3.5}$$

Où p_o est le pourcentage total d'accord entre les annotateurs. L'élément p_e est la proportion d'accord attendue au hasard. Les valeurs p_o et p_e sont calculées à partir d'une table de confusion (3.4) :

Où v_1 et v_2 sont les totaux marginaux verticaux ; h_1 et h_2 sont les totaux marginaux horizontaux. $N = v_1 + v_2$ ou $N = h_1 + h_2$. a représente la somme des objets étiquetés comme positifs

Tableau 3.4 Table de confusion pour le calcul de p_o et p_e.

		Annotateur A		
		Oui	Non	Totaux
Annotateur B	Oui	a	b	$h_1 = a+b$
	Non	c	d	$h_2 = c+d$
	Totaux	$v_1 = a+c$	$v_2 = b+d$	N

par les annotateurs A et B, c'est-à-dire l'accord positif. L'élément b est la somme des objets classifiés comme négatifs par l'annotateur A mais ayant été classifiés comme étant positifs par l'annotateur B. L'élément c est la somme d'objets classifiés comme étant positifs par l'annotateur A, mais classifiés comme étant négatifs par l'annotateur B. L'élément d est la somme des objets étiquetés comme étant négatifs par les annotateurs A et B, c'est-à-dire l'accord négatif.

$$p_0 = \frac{a+d}{N} \tag{3.6}$$

$$p_e = \frac{v_1 h_1 + v_2 h_2}{N^2} \tag{3.7}$$

Feinstein & Cicchetti (1990) identifient deux paradoxes lors du calcul du coefficient κ. Un premier paradoxe apparaît si la valeur de p_e et p_0 sont élevées. Examinons le tableau 3.5 en guise d'exemple.

Tableau 3.5 Exemple du premier paradoxe de *Kappa*.

		Anotateur A			
		Oui	Non	Totaux	
Annotateur B	Oui	80	10	90	h_1
	Non	5	5	10	h_2
	Totaux	85	15	100	
		v_1	v_2		

La valeur de $p_o = (80+5)/100 = 0.85$ La valeur de $p_e = (85 \times 90 + 15 \times 10)/100^2 = 0.78$. La valeur finale de $\kappa = (0.85 - 0.78)/1 - 0.78 = 0.3181$. Nous constatons que même si l'accord entre les annotateurs est élevé (voir tableau 3.5) la valeur de κ est très basse. Ce paradoxe peut apparaître lors d'un déséquilibre entre les totaux marginaux, soit verticaux, soit horizontaux (Feinstein & Cicchetti, 1990).

Tableau 3.6 Exemple d'une matrice de confusion sans paradoxe du coefficient κ.

		Anotateur A			
		Oui	Non	Totaux	
Annotateur B	Oui	40	9	49	h_1
	Non	6	45	51	h_2
	Totaux	46	54	100	
		v_1	v_2		

Nous pouvons remarquer que, dans le tableau 3.6, les totaux marginaux sont équilibrés et donc le paradoxe 1 ne se présente pas. La valeur $p_0 = 0.85$. la valeur $p_e = 0.5$. La valeur $\kappa = 0.70$ dans ce cas.

Cependant, un deuxième paradoxe apparaît lorsque, κ est élevée avec un déséquilibre asymétrique plutôt qu'un équilibre symétrique dans les marginaux totaux (Feinstein & Cicchetti, 1990). Un déséquilibre asymétrique dans les totaux marginaux est présenté dans le tableau 3.7 où nous observons des valeurs asymétriquement décompensés pour les totaux : $v_1 = 30$ et $h_1 = 60$; pour $v_2 = 70$ et $h_2 = 40$.

Tableau 3.7 Exemple du deuxième paradoxe de *Kappa*.

		Annotateur A			
		Oui	Non	Totaux	
Annotateur B	Oui	25	35	60	h_1
	Non	5	35	40	h_2
	Totaux	30	70	100	
		v_1	v_2		

Pour l'exemple du tableau 3.7, la valeur de $p_e = (30 \times 60 + 70 \times 40)/100^2 = 0.46$ Dans ce cas, les totaux marginaux ($v1, h1$ et $v2, h2$) ont un balance asymétrique. La valeur $p_o = (25 + 35)/100 = 0.6$. La valeur $\kappa = (0.6 - 0.46)/(1 - 0.46) = 0.26$.

Il faut aussi remarquer que le deuxième paradoxe apparaît aussi lorsqu'il y a une symétrie imparfaite plutôt qu'une symétrie parfaite dans l'équilibre des marginaux totaux (Feinstein & Cicchetti, 1990). Une symétrie imparfaite est observée dans les totaux marginaux du tableau 3.8, où $v_1 = 70$ et $h_1 = 60$ sont symétriques, mais imparfaits. En conséquence, nous observons cette même symétrie imparfaite pour $v_2 = 30$ et $h_2 = 40$.

Tableau 3.8 Exemple du deuxième paradoxe de *Kappa*.

		Annotateur A			
		Oui	Non	Totaux	
Annotateur B	Oui	45	15	60	h_1
	Non	25	15	40	h_2
	Totaux	70	30	100	
		v_1	v_2		

Le tableau 3.8 présente un cas où les totaux marginaux ont un déséquilibre plutôt symétrique. Pour ce cas, la valeur de $p_e = (70 \times 60 + 30 \times 40)/100^2 = 0.54$. Cet exemple est plutôt dit « symétrique » par balance des valeurs $v1, h1$ et $v2, h2$. La valeur de $p_o = (45 + 15)/100 = 0.6$. La valeur de $\kappa = (0.6 - 0.54)/(1 - 0.54) = 0.13$. Dans ces deux derniers exemples, nous remarquons que l'accord total (p_o) entre les annotateurs est le même, mais p_e est différent, ce qui altère grandement la valeur du coefficient.

Pour résoudre ces paradoxes, les mêmes auteurs dans un article distinct, (Cicchetti & Feinstein, 1990), proposent d'utiliser deux coefficients séparés pour exprimer l'accord, soit pour les catégories positive ou négative entre deux annotateurs. Nous utilisons les *coefficients d'accord* suivants comme des mesures pour exprimer séparément l'accord positif et l'accord négatif entre les annotateurs.

$$P_{pos} = \frac{2a}{v_1 + h_1} \tag{3.8}$$

$$P_{neg} = \frac{2d}{N - (a - d)} \tag{3.9}$$

Le calcul des coefficients pour les classes positive et négative s'adapte parfaitement à notre cas, car nous sommes intéressés à comprendre ce qui permet aux annotateurs d'identifier ce qui est nouveau et ce qui ne l'est pas. Les coefficients sont calculés entre chaque paire d'annotateurs. Les valeurs de ces coefficients accompagnent la valeur du coefficient κ pour une meilleure interprétation. Nous suivons ainsi la recommandation de Cicchetti & Feinstein (1990).

3.4 Synthèse des choix méthodologiques

Dans ce chapitre, nous avons exposé les postulats philosophiques et leur implication sur notre méthodologie. Nous avons expliqué que nous adhérons au post-positivisme comme cadre interprétatif. Nous avons adopté cette position, car le post-positivisme admet qu'une connaissance *a priori* puisse être testée par les résultats des expériences. Ce cadre interprétatif nous permet en particulier d'utiliser des théories en linguistique et en cognition pour élaborer nos propositions et les évaluer.

En ce qui concerne nos expériences, chacune aborde les scénarios que nous avons identifiés. Pour le scénario 1, qui traite la couverture d'information dans les textes d'étudiants, nous utilisons un corpus de quatre dissertations et l'ensemble de documents dans les références générales (RG) et les références spécialisées (RS). Nous nous servons d'une mesure de couverture d'information pour analyser ces textes. La mesure prend en compte la terminologie entre un référent (document des RG ou des RS) et la dissertation d'un étudiant. La valeur de cette couverture sera utilisée pour faire un alignement entre les paragraphes d'une dissertation et les documents des RG et des RS. Cet alignement pourrait refléter l'influence des RG et des RI dans les dissertations.

L'évaluation des résultats pour le premier scénario se fait avec un PVM. Les sessions sont réalisées sous la forme d'une présentation avec les enseignants du cours où les dissertations ont été construites. Cette rencontre nous permet de donner d'abord notre interprétation aux enseignants et de susciter la discussion et la négociation des interprétations des résultats.

En ce qui concerne le scénario 2 qui traite de la couverture d'information dans les textes journalistiques, nous utilisons une partie du corpus *novelty TREC*, sur lequel nous produisons un nouvel étiquetage. Pour des raisons liées à la quantité d'information qu'un être humain peut traiter, nous avons choisi cinq thématiques du corpus original avec cinq nouvelles pour chaque thématique.

L'étiquetage est mené par quatre experts possédant tous une bonne connaissance en TALN. Les annotateurs devaient considérer les groupes syntagmatiques et leur combinaison pour déterminer si une phrase contenait de l'information nouvelle. Cet étiquetage sert à l'évaluation de notre proposition.

Pour évaluer la couverture d'information dans les textes journalistiques, nous proposons une mesure de couverture, qui utilise des patrons linéaires qui tentent de capturer les fonctions grammaticales de sujet, objet, et d'adjoint donc le lieu et le temp. Cette mesure est utilisée dans une expérience avec le corpus étiqueté par nos experts. Pour chacune des cinq thématiques qui composait le corpus, la première nouvelle dans un ordre chronologique est considérée comme le référent. Les quatre autres nouvelles sont traitées à leur tour comme le sujet de comparaison. Le référent est de taille variable puisqu'il sera alimenté avec les phrases nouvelles du sujet qui seront trouvées lors de la vérification de la couverture de l'information.

CHAPITRE 4

RÉSULTATS

Dans ce chapitre, nous présentons les résultats de nos expériences. De la même façon que dans les chapitres précédents, nous consacrons une section au scénario 1, la couverture d'information dans les dissertations d'étudiants, et une autre section au scénario 2, la couverture d'information dans les textes journalistiques.

Dans la section 4.1, nous exposons les résultats du scénario 1. Nous avons utilisé trois types de visualisation pour illustrer les résultats, dont confère une interprétation différente des résultats. Cette section inclut aussi l'évaluation que nous avons conduite sous la forme d'un *Processus de Vérification par les Membres* (PVM). Les participants du PVM sont les enseignants du cours pour lequel les étudiants ont produit leurs dissertations. Le corpus utilisé dans le scénario 1 est constitué de deux parties, les dissertations des étudiants et les références qu'ils ont utilisées pour réaliser leur dissertation.

Dans la section 4.2, nous présentons les résultats du scénario 2. Pour celui-ci, nous avons étiqueté une partie du corpus *Novelty TREC*. Nous présentons d'abord une évaluation de l'accord entre les annotateurs, pour ensuite présenter l'évaluation quantitative des résultats issus du processus de couverture d'information.

Une brève conclusion termine ce chapitre.

4.1 Scénario 1 : la couverture d'information dans les dissertations d'étudiants

Comme nous l'avons mentionné dans l'introduction de ce chapitre, le corpus que nous utilisons dans ce scénario est constitué de deux parties, les dissertations des étudiants et les références. Cette dernière partie est divisée en deux ; nous avons repris le diagramme de la distribution des références (à la fig 4.1, originalement fig. 3.2) pour décrire ses composantes. La première partie contient les documents des *références générales* (RG) qui ont été suggérés par les enseignants et qui couvrent une thématique générique du contenu du cours. Elles sont représentées par le

cercle bleu dans le diagramme à la fig. 4.1. Les RG ne sont pas toutes nécessairement citées dans les dissertations, puisque les étudiants ont la liberté de choisir leurs propres références. La deuxième partie est composée des documents des *références individuelles* (RI) qui ont été choisies par l'étudiant pour élaborer sa dissertation ; elles portent sur une thématique particulière, liée au sujet de la dissertation. Les RI sont représentées par le cercle jaune à la fig. 4.1. Les RI sont nécessairement citées dans les dissertations des étudiants. Puisque les RI les étudiants auraient pu utiliser des RI et des RG dans leur dissertations, il y a une intersection entre les RG et RI. L'ensemble de cette intersection est représenté dans le diagramme de la fig 3.2 par l'intersection des deux cercles, soit la partie en vert. Les documents des RI qui ne sont pas des RG constituent les *références spécialisées* (RS) des dissertations [1] ; elles sont représentées par la partie complètement jaune du cercle dans le diagramme à la fig 4.1. Lors d'une évaluation, les étudiants ont signalé le degré d'utilisation de quelques références pour l'élaboration de leurs dissertations. Nous appelons cet ensemble les *références considérées* (RC). Les RC sont identifiés dans le diagramme à la fig. 4.1 par l'ovale qui se situe entre l'intersection des RG et RI et une partie des RS.

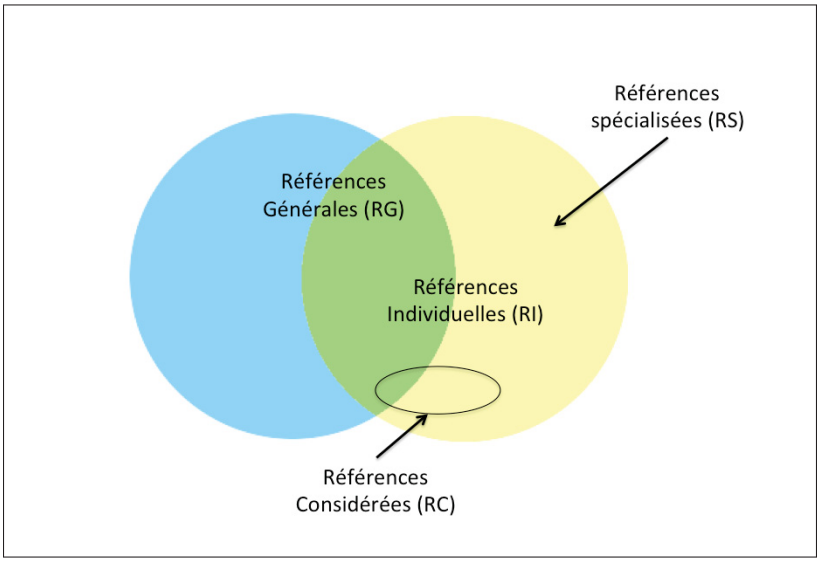


Figure 4.1 Diagramme de la distribution des documents des références et leur nomenclature utilisée.

1. La notation d'ensembles est : $RS = RI \setminus RG$.

Pour évaluer la couverture d'information, nous avons implémenté une stratégie d'alignement paragraphes/documents, en calculant la similarité entre les paragraphes de chaque dissertation et les paragraphes des RG et des RS.

Cette méthode nous permet de comparer trois mesures de similarité : la similarité cosinus, le coefficient Dice et la mesure que nous proposons. Les deux premières sont utilisées comme *référentiel*.

Afin de comparer chaque mesure, nous avons fixé des seuils spécifiques qui permettent d'obtenir le meilleur alignement qu'une mesure peut donner individuellement. Les seuils choisis empiriquement sont présentés au tableau 4.1. Ainsi, un paragraphe dont la valeur de mesure est supérieure ou égale à 0.3, dans le cas de cosinus, sera identifié comme étant aligné à un document des références.

Tableau 4.1 Seuils établis pour chaque mesure.

Mesure	**Seuil**	**Caractéristique**
Similarité cosinus	0.3	Symétrique
Coefficient de Dice	0.43	Symétrique
ACHM	0.43	Asymétrique

Pour interpréter les résultats, nous proposons trois visualisations différentes. Dans la section 4.1.1, nous utilisons des graphes en nuage de points. Cette visualisation permet de montrer, pour chaque dissertation, combien de documents des RG et des RS ont été alignés à chacun des paragraphes.

La section 4.1.2 présente les résultats sous la forme d'histogrammes. Cette visualisation permet de voir le nombre de fois qu'un document des RG ou des RS a été aligné avec l'une des quatre dissertations. Nous présentons cette visualisation pour les deux mesures de similarité que nous considérons comme *référentiel* et pour notre mesure ACHM.

Dans la section 4.1.3, nous présentons l'interaction des paragraphes des dissertations avec les RG et les RS à l'aide d'une visualisation à base de graphes tripartites.

Finalement, la section 4.1.4 présente le PVM que nous avons conduit avec les enseignants du cours.

4.1.1 Nombre de documents couverts par chaque dissertation

Tout d'abord, nous construisons un graphe en nuage de points qui montre, sur l'axe des *X*, l'index de tous les paragraphes d'une dissertation, et sur l'axe des *Y*, la valeur de couverture avec notre coefficient ACHM. Les points représentent les documents des RG ou des RS avec lesquels le paragraphe est aligné. Dans ces graphes, nous avons aussi inclus, pour chaque paragraphe, une étiquette qui contient le nombre de fois où le paragraphe a été aligné. Par exemple, la figure 4.2 montre que les paragraphes 6, 8 et 9 ont été alignés 14 fois avec les RG. Nous avons conduit séparément notre analyse pour les RG et les RS.

– Les RG :

Nous présentons les résultats d'analyse pour l'étudiant A qui a remis deux dissertations différentes. L'alignement de la dissertation 1 avec seulement les RG est illustré à la figure 4.2. Cette dissertation a reçu une note d'échec. Comme nous pouvons le voir, l'étudiant A a peu de RG alignées avec les paragraphes, ce qui signifie que les paragraphes de sa dissertation couvrent mal ces documents.

L'étudiant A, ayant repris le même cours, a présenté une nouvelle dissertation ; la fig. 4.3 présente l'alignement de cette dissertation 2 avec les RG seulement. Cette dissertation a reçu une note de succès de la part des enseignants. Comme nous pouvons le voir, la dissertation 2 présente une distribution plus dense et plus fluide de l'alignement entre les paragraphes de la dissertation avec les documents des RG. Cette caractéristique n'est pas présente dans l'analyse de la dissertation 1 (voir fig. 4.2).

Si nous comparons le nombre d'alignements (étiquettes sur les paragraphes) de la figure 4.2 (dissertation 1) et ceux de la figure 4.3 (dissertation 2), nous constatons que cette dernière présente plus d'alignements que la dissertation 1. En effet, la dissertation 1 présente 107 ali-

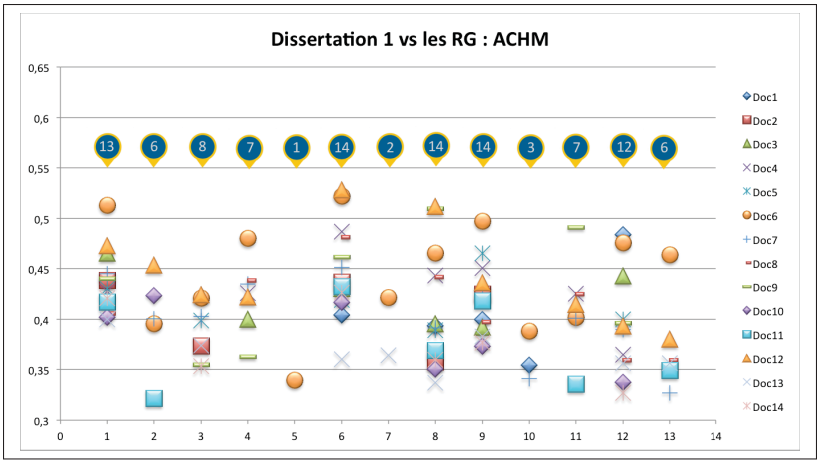


Figure 4.2 Dissertation 1. Alignement des paragraphes avec les documents des RG.

gnements ; la moyenne d'alignements est de 8.2 avec un écart type de 4.7. En ce qui concerne la dissertation 2, celle-ci présente 140 alignements, une moyenne d'alignements de 9.3 avec un écart type de 4.4.

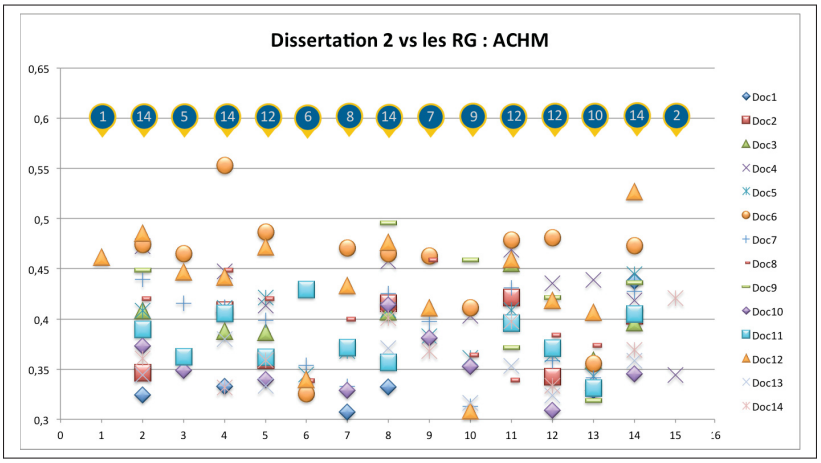


Figure 4.3 Dissertation 2. Alignement des paragraphes avec les documents des RG.

– Les RS :

Les fig. 4.4 et 4.5 montrent l'alignement entre les dissertations 1 et 2 (étudiant A) et les RS. Pour la dissertation 1, nous avons récupéré douze des RS (l'étudiant cite un total de quinze

références dans cette dissertation). Nous voyons dans cette analyse que la dissertation 1 présente une densité d'alignement plus faible que celle de la dissertation 2 (voir les chiffres des étiquettes). La quantité de paragraphes alignés de la dissertation 2 dénote une meilleure couverture des documents des RS.

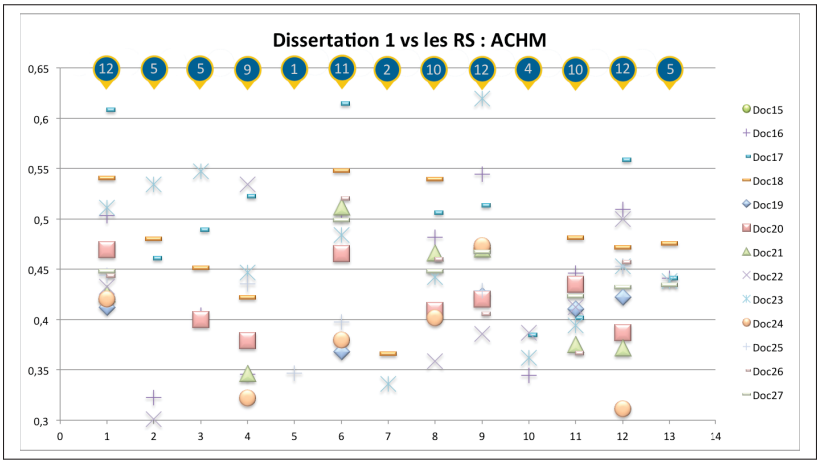


Figure 4.4 Dissertation 1. Alignement des paragraphes avec les documents des RS.

Dans le cas de la dissertation 2, nous avons récupéré quatorze des RS (l'étudiant a cité originalement vingt-neuf références). La figure 4.5 présente l'alignement des paragraphes et les RS. Nous pouvons constater que la dissertation 2 montre un plus grand nombre de paragraphes alignés avec les RS que la dissertation 1. Le nombre d'alignements sur ce graphe (documents alignés) est supérieur par rapport à la dissertation 1(voir les chiffres des étiquettes dans les deux graphes).

Dans le cas de l'alignement des deux dissertations avec les RS, la dissertation 1 présente 98 alignements avec une moyenne d'alignements de 7.5 et un écart type de 4.0. En ce qui concerne la dissertation 2, elle présente 154 alignements avec une moyenne d'alignements de 10.3 et un écarte type de 3.4.

– Les RG et les RS :

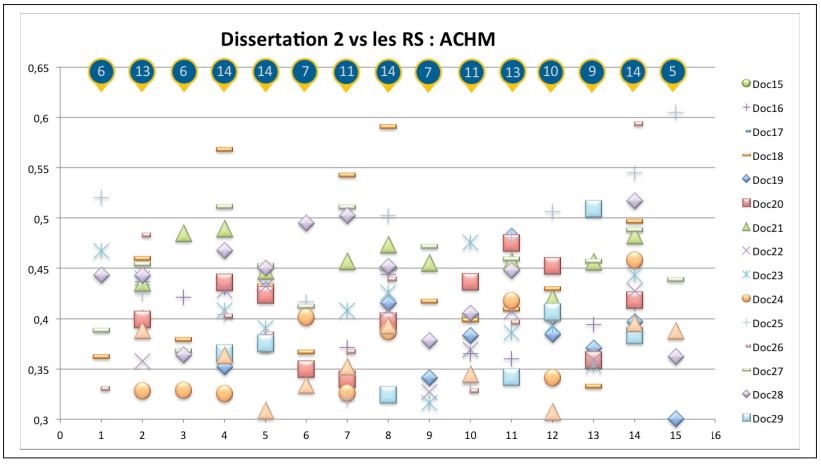


Figure 4.5 Dissertation 2. Alignement des paragraphes avec les documents de RS.

Nous présentons aussi des boîtes à moustaches sur la fig 4.6 pour illustrer la distribution des valeurs de couverture des quatre dissertations. Ce graphe analyse toutes les valeurs de couverture pour les quatre dissertations avec les RG et les RS.

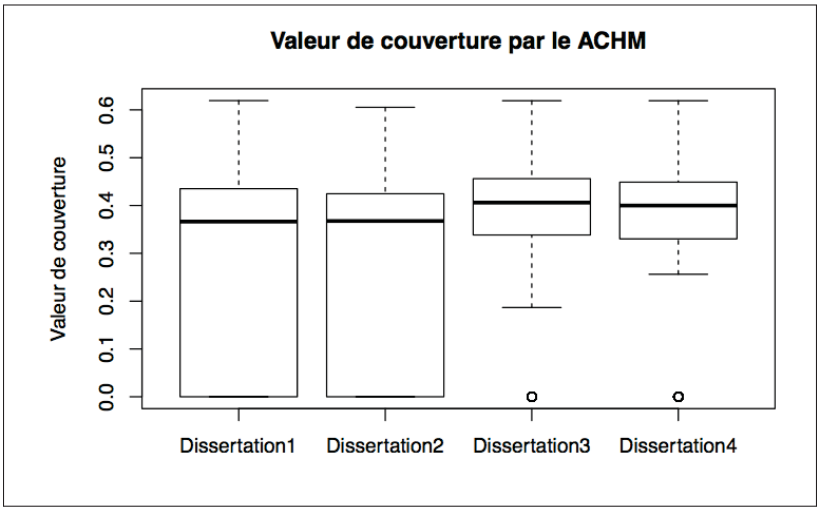


Figure 4.6 Distribution des valeurs de couverture des dissertations avec le ACHM.

Comme nous pouvons l'observer à la figure 4.6, la boîte pour les dissertations 1 et 2 inclut les valeurs de couverture minimales pour les deux dissertations. Nous observons dans ces graphes

que 75 % des valeurs de couverture sont comprises dans un intervalle compris entre 0 et 0.45 plus ou moins. Alors, la couverture des documents de la bibliographie est très dispersée. Pour ces deux dissertations (1 et 2), la présence des valeurs à zéro dénote aussi que plusieurs des documents provenant des RG ou des RS n'ont pas été couverts dans la dissertation.

Les dissertations 3 et 4, quant à elles, comportent des *données aberrantes* (représentées par un point dans le graphe), ce qui dénote le fait que très peu de RG ou de RS n'ont pas été couvertes dans ces dissertations. Remarquons que la dissertation 3 a une moyenne supérieure aux autres dissertations (légèrement par rapport à la dissertation 4). Les valeurs minimales de couverture pour cette dissertation sont de 0.2. La boîte qui représente 50 % de la distribution des données est située entre 0.35 et 0.45. Cette analyse reflète que cette dissertation a une bonne couverture des documents des RG et des RS.

Pour la dissertation 4, la valeur minimale est de 0.25. La boîte représentant 50 % des valeurs de couverture est située entre 0.33 et 0.43. La couverture d'information des RG et des RS pour cette dissertation est encore plus concentrée dans les valeurs élevées par rapport aux trois autres dissertations.

4.1.2 L'influence des RG et des RS sur la production des dissertations

Pour identifier l'influence qu'un document des RG ou des RS a produit sur chacune des dissertations, nous avons eu recours au nombre d'alignements. Nous considérons qu'un document des RG ou des RS, qui est aligné avec la plupart des paragraphes d'une dissertation, a produit une influence marquante sur la production de la dissertation.

Pour donner un aperçu de l'alignement des quatre dissertations avec les RG et les RS, nous présentons nos résultats à l'aide d'histogrammes. Chacun des graphes montre, sur l'axe des abscisses, les quatre dissertations des étudiants. L'axe des ordonnées correspond au nombre de paragraphes auquel un document des RG ou des RS a été aligné. Cette valeur sera appelée Occurrence d'Alignement d'un Document (OAD). La ligne noire, dans chaque graphique, représente la moyenne de tous les OAD, que nous appellerons la Moyenne Générale d'Aligne-

ment (MGA). Le point noir sur chaque dissertation est la Moyenne Individuelle d'Alignement (MIA). Finalement, nous avons séparé cette analyse pour les RG et les RS.

– *Similarité cosinus :*

Les figures 4.7, et 4.8 montrent l'alignement des quatre dissertations avec les RG et les RS respectivement. Ces images montrent une analyse basée sur l'approche de similarité cosinus. Cette visualisation nous permet de déduire quels sont les documents des RG (fig 4.7) ou des RS (fig. 4.8) qui ont le plus influencé chaque dissertation.

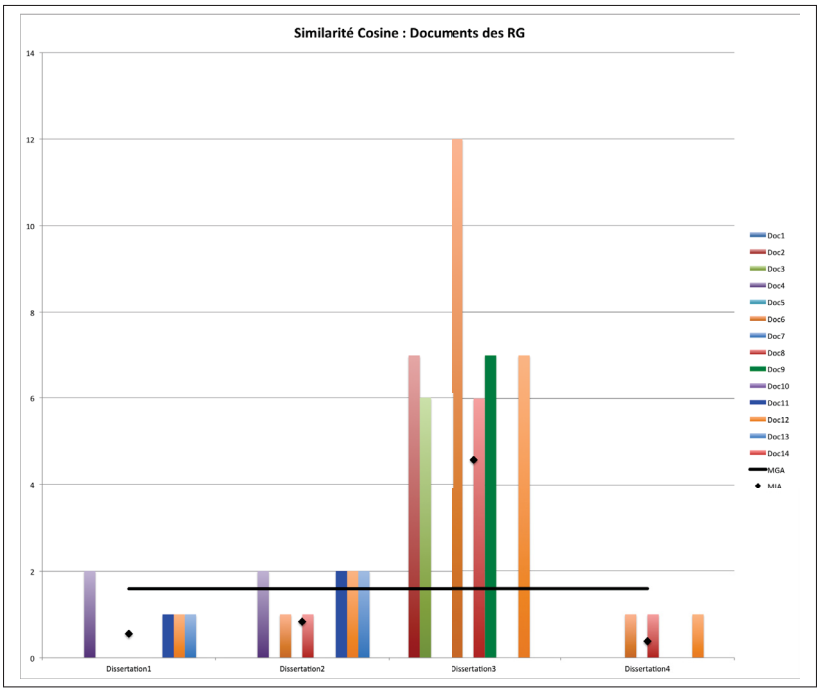


Figure 4.7 Alignements des dissertations avec les RG.
Approche par similarité cosinus

À la figure 4.7, nous pouvons observer que pour la dissertation 1 (première tentative de l'étudiant A), un seul document des RG atteint la MGA. Toujours pour la dissertation 1, la situation se répète à la figure 4.8, où un seul document des RS atteint la MGA. Pour la dissertation 2 (deuxième tentative étudiant A), quatre documents des RG (fig.4.7), et deux documents des RS dépassent la MGA (fig. 4.8). En ce qui concerne la dissertation 3, il y a six documents des RG

(fig. 4.7) et dix documents des RS (fig. 4.8) qui dépassent la MGA. La dissertation 3 est celle qui est le mieux alignée par rapport aux trois autres dissertations. Un problème surgit lorsqu'on examine les résultats couvrant la dissertation 4. Nous savons que la note de cette dissertation a été de réussite lors à son évaluation, mais ces deux graphiques montrent qu'aucun des documents des RG ne dépasse la MGA (fig. 4.7) ; et que seulement un document des RS est aligné (fig. 4.8). Il s'agit d'une situation pire que pour la dissertation 1 si nous considérons les notes assignées à chaque dissertation.

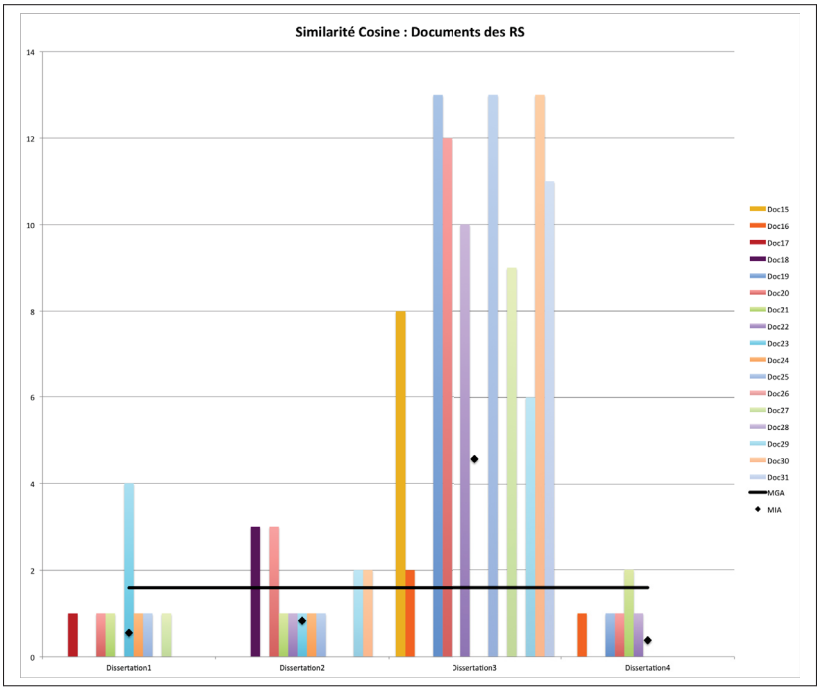


Figure 4.8 Alignements des dissertations avec les RS.
Approche par similarité cosinus

L'approche par similarité cosinus présente donc une difficulté à refléter la situation réelle lors de l'évaluation des dissertations. De plus, cette approche ne permet pas de visualiser l'influence des RG ou des RS dans les dissertations. Si nous considérons aussi la MIA, nous remarquons que les dissertations 1, 2 et 4 ont une moyenne similaire et que la dissertation 3 a une moyenne individuelle très élevée. Si nous utilisions cette information pour projeter ces dissertations dans un espace géométrique, les dissertations 1, 2 et 4 seront près les unes des autres, formant ainsi

un groupe. La dissertation 3 serait éloignée de ce groupe. Dans les espaces géométriques, la distance est une mesure de similarité ; dans ce type d'approche, les objets qui sont près sont censés partager des caractéristiques. Si nous savons que la dissertations 1 a obtenu un échec et le reste des dissertations a obtenu une réussite, nous souhaiterions avoir un groupe séparé pour les échecs et les réussites.

– *Coefficient de Dice :*

Nous avons aussi testé une approche basée sur le coefficient de Dice. La fig 4.9 correspond à l'alignement des dissertations avec les RG. La fig 4.10 correspond à l'alignement des dissertationss avec RS. Dans le cas où nous utilisons le coefficient de Dice pour faire l'alignement, la MGA (ligne noire) est beaucoup plus élevée par rapport à la mesure de similarité cosinus.

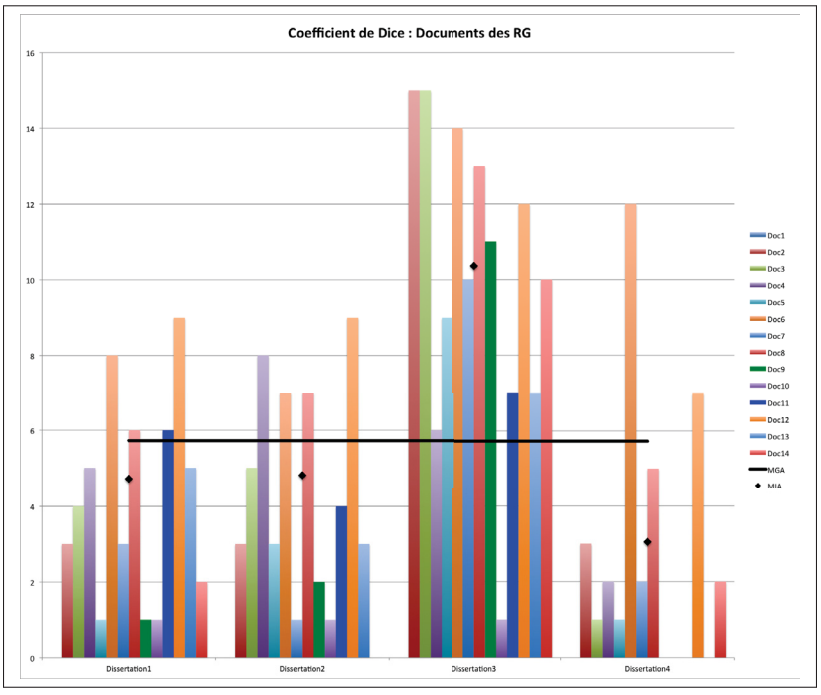


Figure 4.9 Alignements des dissertations avec les RG. Approche coefficient de Dice

À la fig. 4.9, nous pouvons observer que, dans la dissertation 1, il y a deux documents des RG qui dépassent à peine la MGA et deux autres qui le font avec un plus grand nombre d'ali-

gnements. La dissertation 2 comporte quatre documents dont des RG qui dépassent la MGA. La dissertation 3 présente toujours le plus grand nombre de documents des RG alignés (douze documents) qui dépasse la MGA. La dissertation 4 présente deux documents des RG alignés dépassant la MGA. Cette approche montre aussi un alignement très pauvre pour la dissertation 4 qui a reçu une note de réussite. Si on se fie au nombre d'alignements pour déterminer l'influence de chacune des RG pour l'élaboration des dissertations, nous pourrions déduire facilement l'influence individuelle des RG sur les dissertations 1, 2 et 4. Par contre, pour la dissertation 3, il serait un peu plus difficile de déterminer l'influence de chacune des RG puisque presque tous ces documents présentent un grand nombre d'alignements.

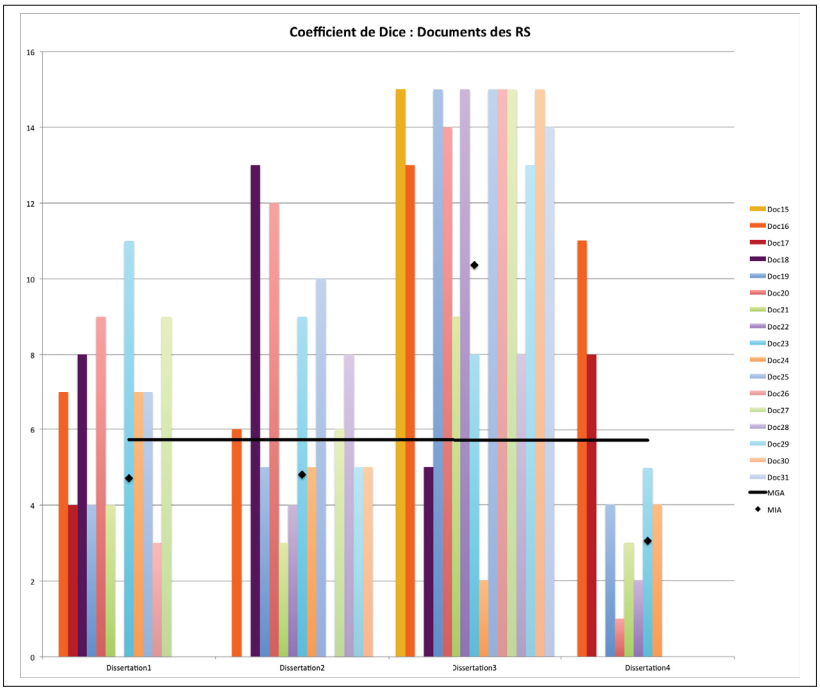


Figure 4.10 Alignements des dissertations avec les RS. Approche coefficient de Dice.

En ce qui concerne les RS, la fig. 4.10 montre que la dissertation 1 détient sept documents qui dépassent la MGA. Pour la dissertation 2, nous observons également que sept documents des RS dépassent la MGA. En ce qui concerne la dissertation 3, nous voyons que la plupart

des RS ont dépassé la MGA sauf deux documents (index 18 et 24). La dissertation 4 présente seulement deux documents des RS qui ont dépassé la MGA.

Le graphe à la fig. 4.10, nous permet d'observer une différence avec le nombre des RS alignés entre les dissertations 1 et 2. La dissertation 1 a également des RS qui dépassent la MGA de la dissertation 2 ; la différence réside dans le nombre de paragraphes alignés. Rappelons que ces dissertations ont été élaborés par l'étudiant A et que c'est seulement la dissertation 2 qui a obtenu une note de réussite. Malgré cela, l'approche basée sur la mesure de Dice ne reflète pas la réalité de la note des dissertations 1 et 2. De même, la dissertation 4 a un nombre très bas de documents dont des RS qui dépasse la MGA.

Les figures 4.9 et 4.10 présentent des moyennes individuelles plus élevées que celles observées en utilisant l'approche de similarité cosinus. La dissertation 3 présente toujours la MIA la plus élevée, et le document 4 correspond à la valeur minimale parmi les quatre documents. En utilisant cette analyse et en faisant une projection des moyennes dans un espace coordonné, les dissertations 1, 2 et 4 sont toutes proches, ce qu'indique une similarité entre eux. Idéalement, nous devrions constituer un groupe pour les dissertations qui ont reçu une bonne note et un autre pour les dissertations qui ont eu un échec. L'analyse avec le coefficient de Dice ne permet pas d'exprimer cette réalité.

– ACHM :

La fig. 4.11 et la fig. 4.12 présentent les résultats de notre mesure ACHM. Dans les deux graphes, nous remarquons que la MGA est plus élevée dans notre approche que dans les deux précédentes.

La fig 4.11 montre que la dissertation 1 dispose de trois documents des RG qui dépassent la MGA. Si nous les comparons avec le reste des dissertations, le nombre d'alignements qui dépasse la MGA est plus faible. En ce qui concerne la dissertation 2 (deuxième tentative de l'étudiant A), le nombre de RG qui dépasse la MGA est nettement plus élevé. Pour la dissertation 2, il y a huit documents qui dépassent la MGA ; parmi eux, les documents des RG qui ont

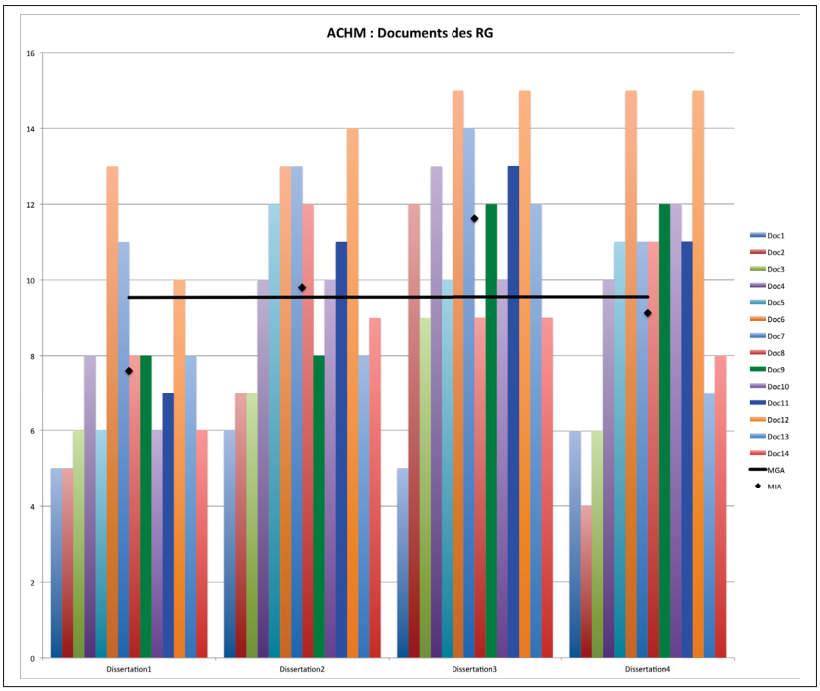


Figure 4.11 Alignements des dissertations avec les RG. Approche ACHM .

plus d'alignements sont les documents 6, 7 et 12 [2]. La dissertation 3 présente dix documents qui dépassent le MGA. Les RG les plus influentes pour la dissertation 3 sont les documents 6 et 12 qui correspondent aux valeurs maximales d'alignement dans le graphique. Cette fois, la dissertation 4 présente neuf documents des RG qui dépassent la MGA. Les documents qui ont le nombre le plus élevé d'alignements sont les documents 6 et 12. Nous pourrions dire que les documents qui ont un nombre élevé d'alignements auraient plus d'influence sur le contenu des dissertations. La dissertation 1 présente un faible nombre d'alignements par rapport aux trois autres dissertations. L'alignement des dissertations versus les documents des RG avec l'ACHM semble bien représenter la réalité des notes attribuées aux dissertations des étudiants.

À la figure 4.11, nous pourrons aussi observer que la MIA des dissertations est différente des deux autres approches (similarité cosinus et coefficient de Dice). Cette fois, la dissertation 1 détient la moyenne la plus faible. La MIA des dissertations 2 et 3 est au-dessus de la MGA.

2. Nous discuterons le contenu des RG et des RS dans la section 4.1.3.

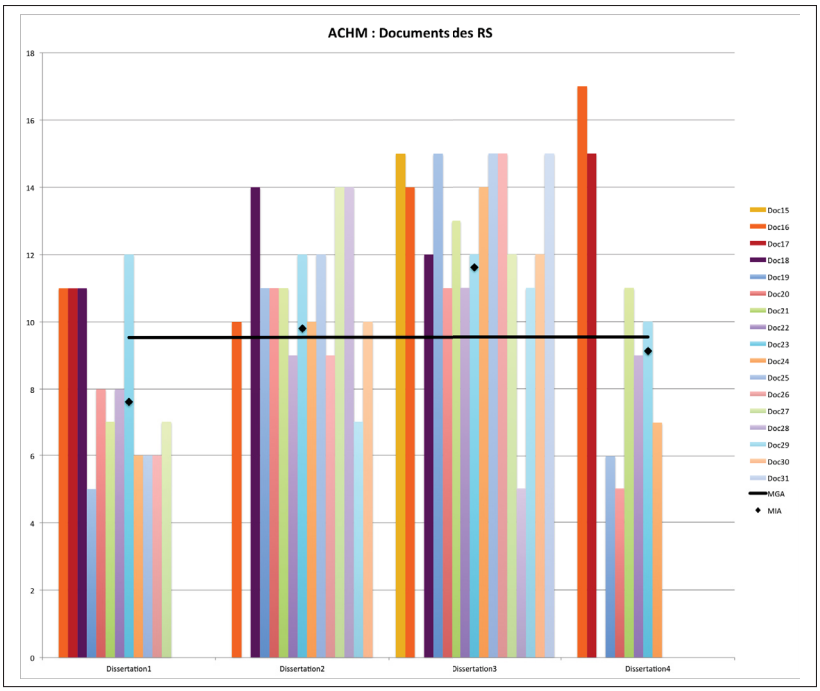


Figure 4.12 Alignements des dissertations avec les RS. Approche ACHM.

La fig. 4.12 montre l'occurrence des l'alignement des paragraphes avec les RS. Nous voyons que la dissertation 1 présente quatre documents des RS qui ont dépassé la MGA. La dissertation 2 dispose de huit documents des RS qui dépassent la MGA. Pour cette dissertation, les documents 24, 25 et 26 comptent le plus grand nombre d'alignements. En ce qui concerne la dissertation 3, quinze documents des RS ont dépassé la MGA. Parmi ceux-ci, les documents 15, 16, 22, 24 et 30 comptent le nombre le plus élevé d'alignements. Pour la dissertation 4, nous remarquons que trois documents dépassent la MGA : parmi eux, le document 16 possède une occurrence de 17, soit la valeur maximale, si on la compare aux trois autres dissertations.

Si nous considérons le nombre de paragraphes d'une dissertation alignés avec un document des RG ou des RS, nous pourrions déduire quels sont les documents des RG ou des RS qui ont eu une certaine influence sur la rédaction d'une dissertation. Un autre aspect à considérer pour constater une telle influence sur la rédaction d'une dissertation est de faire une analyse plus détaillée de certains des alignements entre les paragraphes et les documents des RG et des RS. Dans la section 4.1.3, nous présentons cette analyse pour la dissertation 1 et la dissertation 3.

Pour obtenir une analyse qui reflète la réalité des notes assignées aux dissertations, nous devrons considérer séparément l'alignement des RG et des RS. Les documents des RG fournissent un espace d'analyse uniforme pour toutes les dissertations. Par contre, les RS varient en nombre et en longueur d'un étudiant à l'autre.

En ce qui concerne les MIA, l'approche de l'ACHM permet à la dissertation 3 d'avoir toujours la valeur maximale, alors que la dissertation 1 comporte la valeur minimale des quatre dissertations. La dissertation 4 a une MIA près de la MGA, et la dissertation 2 la surpasse. La valeur de la MIA pour les trois dissertations qui ont eu une bonne note dans le cours est plus proche que dans les deux autres approches. Si nous utilisions cette information pour créer des groupes dans un espace géométrique, il y aurait un groupe contenant les documents 2, 3 et 4, car ils sont très proches. Ce groupe correspondrait au seuil de réussite. La dissertation 1 serait isolée, dénotant ainsi qu'elle ne partage pas les mêmes caractéristiques que les autres documents, menant ainsi à un échec.

Nous avons aussi élaboré des boîtes à moustaches pour montrer l'alignement des dissertations avec les RG et les RS. Ce graphe permet aussi de visualiser la MIA (linge horizontale de la boîte) de chaque dissertation.

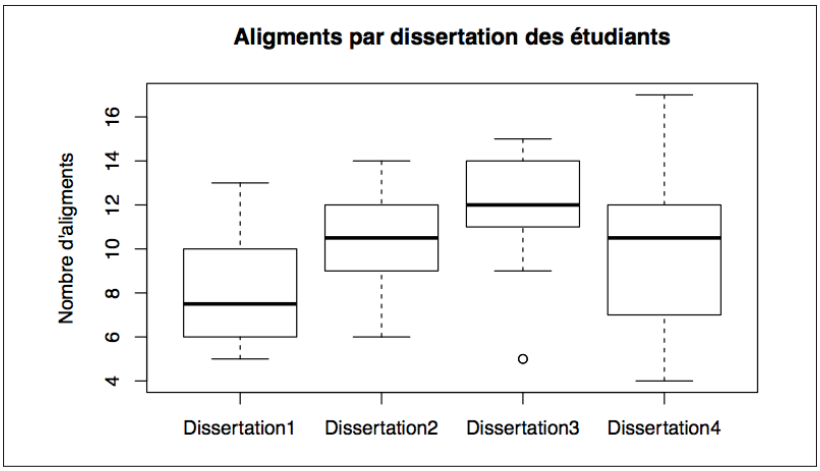


Figure 4.13 Distribution des valeurs de couverture des dissertations avec le ACHM.

La fig 4.13 montre que la dissertation 1 comporte un faible nombre d'alignements avec les RG et les RS ; l'intervalle d'alignement se situe entre cinq et treize. La moitié des RG et des RS a été aligné entre six et dix fois. La dissertation 2 comporte un nombre d'alignements légèrement plus élevé que la dissertation 1 ; le rang se situe entre six et quatorze. On note que 50 % des documents des RG et des RS ont été alignés entre neuf et douze fois. La dissertation 3 montre de deux à cinq *outliers*, que nous pouvons observer ces éléments sur les fig 4.11 (document 1) et 4.12 (document 27). Cette même dissertation présente le nombre d'alignements le plus élevé par rapport aux autres étudiants. La moitié des documents des RG et des RS a été aligné entre onze et quatorze fois. En ce qui concerne la dissertation 4, nous pouvons voir que le nombre d'alignements est beaucoup plus dispersé que pour les autres dissertations, ce qui montre que 50 % des documents des RG et des RS ont été alignés entre sept et onze fois avec cette dissertation. Il est important à voir que la dissertation 4 comporte l'intervalle le plus grand par rapport aux trois autres dissertations. Cela veut dire que cette dissertation contient les deux extrêmes : le paragraphe avec le nombre d'alignents minimal et le paragraphe avec le nombre d'alignements maximal de la collection. Nous observons aussi que pour les dissertations 2, 3 et 4, la MIA (ligne horizontale) de ces documents est supérieure à celle de la dissertation 1, ce qui dénote également une différence en termes de nombre d'alignements.

4.1.3 Réseaux de mots des dissertations

Le nombre de documents alignés et le nombre d'alignements permettent de donner une mesure quantitative du contenu, mais ne permettent pas d'analyser le contenu des dissertations. Pour ce faire, nous avons aussi élaboré une visualisation à l'aide de graphes tripartites pour illustrer, grâce à la terminologie, l'interaction entre les paragraphes des dissertations, et les RG et les RS. Nous analysons le cas de la dissertation 1 et de la dissertation 3. La première dissertation résultant en un échec et la seconde, en un succès. Pour les deux dissertations, nous avons élaboré séparément une visualisation à base de graphes avec les RG et une autre avec les RS. La fig.4.14 et fig. 4.15 montrent les réseaux pour la dissertation 1.

Dans cette représentation, les éléments ont des significations différentes qui sont caractérisées par les couleurs et la taille des nœuds et celle des liens. Les nœuds bleus, sur le côté gauche, représentent les paragraphes de la dissertation. Les nœuds orange, au milieu, représentent les concepts qui ont contribué le plus à la similarité d'un paragraphe de la dissertation avec les RG ou les RS. Ces concepts ont été extraits lors du calcul de similarité grâce à la fonction *maxsim*() de la formule de l'ACHM. Dans le reste de cette thèse, nous allons nous référer à ces concepts comme les *concepts maxsim*. Les nœuds de couleur verte représentent les documents des RG ou les documents des RS.

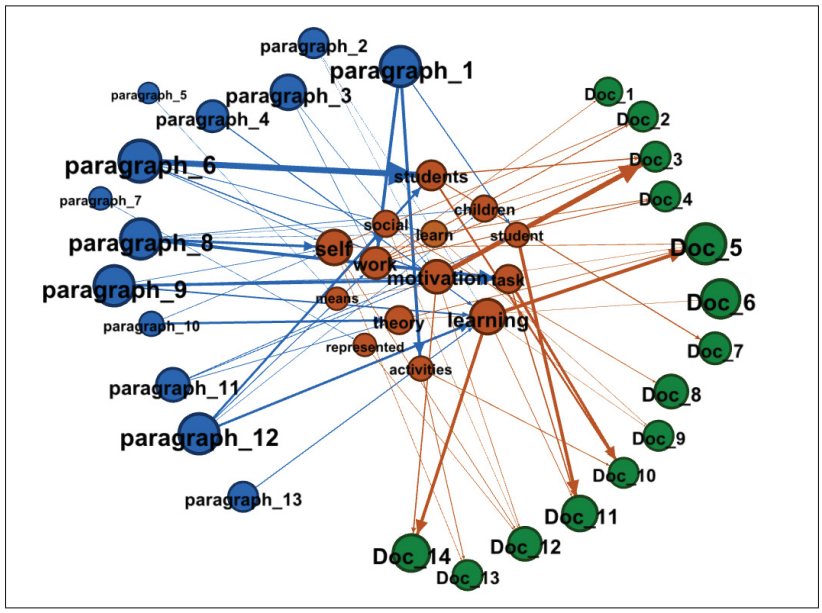


Figure 4.14 Réseau de mots pour la dissertation 1 avec les RG. (Cette dissertation appartient à l'étudiant A.)

La taille des nœuds bleus représente le nombre de *concepts maxsim* extraits de ce paragraphe. L'épaisseur des lignes bleues indique le nombre de fois que le même *concept maxsim* a été utilisé dans le même paragraphe. Cela veut dire que le même *concept maxsim* a contribué à plusieurs reprises à une similarité supérieure avec un paragraphe de la dissertation et un document des RG ou des RS. Nous pouvons dire que l'épaisseur des lignes bleues détermine la prédominance du *concept maxsim* dans le paragraphe. L'ensemble des *concept maxsim* liés à un paragraphe détermine ainsi le sujet de ce paragraphe dans la dissertation.

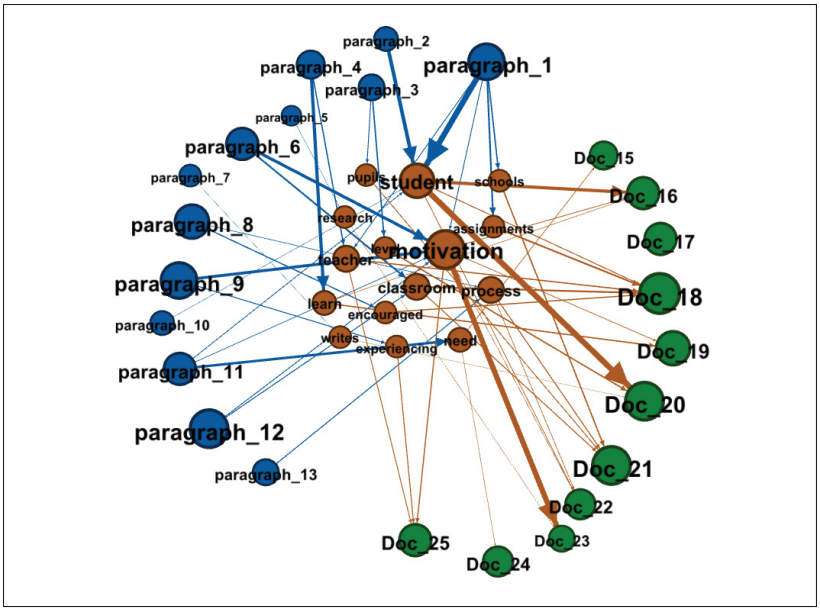


Figure 4.15 Réseau de mots pour le dissertation 1 avec les RS. Ce dissertation appartient à l'étudiant A

La taille des noeuds oranges indique le nombre de fois où ce *concept maxsim* a été sélectionné comme contribuant à la valeur maximale de similarité. L'épaisseur de la ligne orange indique le nombre de fois que chaque *concept maxsim* a été sélectionné comme contribuant à la valeur de similarité maximale avec ce document des RG ou des RS. Nous interprétons ceci comme le concept du document (RG ou RS) qui aurait influencé la dissertation. L'ensemble des *concepts maxsim* qui apparaissent dans le graphe détermine ainsi le sujet de la dissertation.

La taille des nœuds verts reflète le nombre de fois qu'un *concept maxsim* a été repéré de ce document (RG ou RS). Plus la taille d'un nœud vert est grande, plus ce document a eu une influence lors de la rédaction de la dissertation.

Le numéro qui apparaît dans les nœuds blues correspond à l'index du paragraphe de la dissertation. Le numéro qui apparaît dans les nœuds verts correspond aux index des documents des RG et des RS. Nous avons inclus dans l'*annexe IV* les titres de ces documents.

La fig 4.14 et la fig. 4.15 montrent respectivement l'alignement avec les RG et les RS ; nous y voyons que les concepts « *motivation* » et « *student* » apparaissent dans les deux cas. Les

concepts les plus utilisés à la fig. 4.14 sont : *« children », « motivation », « theory », « work », et « learning »*, tandis que dans la figure 4.15, les concepts *« motivation », « student », et « self »* sont les plus utilisés. Le seul concept partagé dans les fig. 4.14 et 4.15 est « *motivation* », qui correspond au sujet de la dissertation de l'étudiant A. Nous nous attendons à obtenir une couverture beaucoup plus riche dans les concepts des RS que des concepts des RG puisque les concepts de ces dernières étant plus génériques et donc moins adaptés à la thématique choisie par les étudiants.

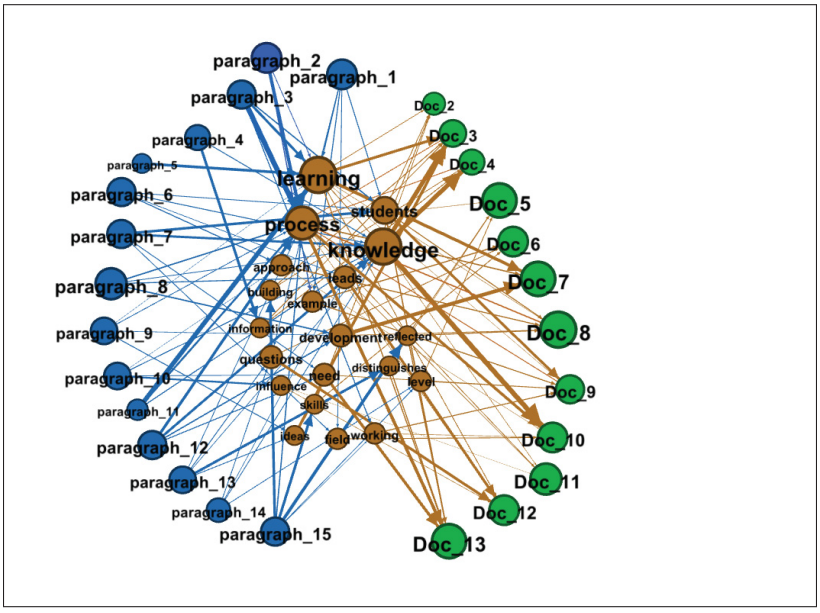


Figure 4.16 Réseau de mots pour la dissertation 3

Les visualisations du réseau de mots pour la dissertation 3 (fig. 4.16 et fig.4.17) montrent que les *concepts maxsim* partagés entre les deux visualisations sont *« learning », « process », et « knowledge »*. L'étudiant 3 a choisi comme sujet de discussion le « *knowledge building* ». Ces trois concepts coïncident en fréquence avec la visualisation de notre analyse du contenu des RG et des RS.

Nous constatons que le concept « *learning* » est principalement extrait du document 3 et 7 des RG (fig. 4.16), tandis qu'il est extrait des documents 18, 23 et 25 des RS (fig.4.17). Le concept « *knowledge* » est principalement extrait des documents 3, 4 et 10 des RG (fig. 4.16)

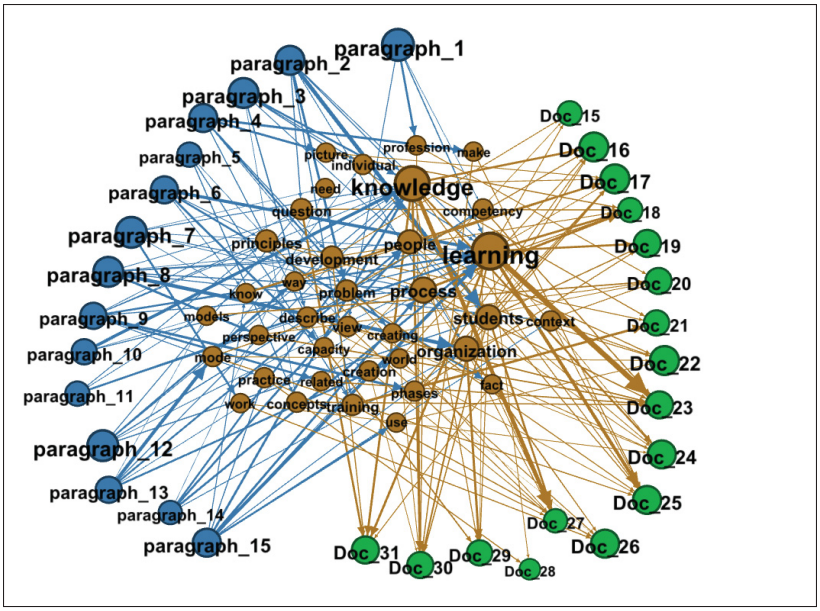


Figure 4.17 Réseau de mots pour la dissertation 3

et des documents 27, 30 et 31 des RS (fig.4.17) . Les figures 4.16 et 4.20 sont statiques et ne permettent pas au lecteur de voir les interactions fluides entre les concepts et les documents des RG et des RS. Cependant, nous avons utilisé le logiciel de visualisation Gephi 3 [3], qui permet la manipulation dynamique et l'interaction avec les éléments du réseau. Par exemple, la fig. 4.18 montre un cas d'interaction dans le logiciel Gephi. Si nous pointons le curseur sur un nœud, ses voisin seront illuminés. Dans cet exemple nous pouvons voir que le concept « *learning* » est utilisé dans les paragraphes 1, 2, 3, 5, 9 et 11, et qu'il est principalement extrait des documents 2, 3, 7, 8, 9, 10, 12 et 13. Les images présentées dans cette section sont des captures d'écran de ce logiciel.

4.1.4 Évaluation

Afin d'obtenir une évaluation qualitative de nos résultats, nous avons conduit un *Processus de Vérification par les Membres (PVM)*. Pour cela, nous avons rencontré les enseignants responsables du cours en deux occasions pour leur présenter l'ensemble des graphes obtenus sous la forme d'une présentation. La première fois que nous avons rencontré les enseignants, nous

3. Voir le lien : https ://gephi.org/

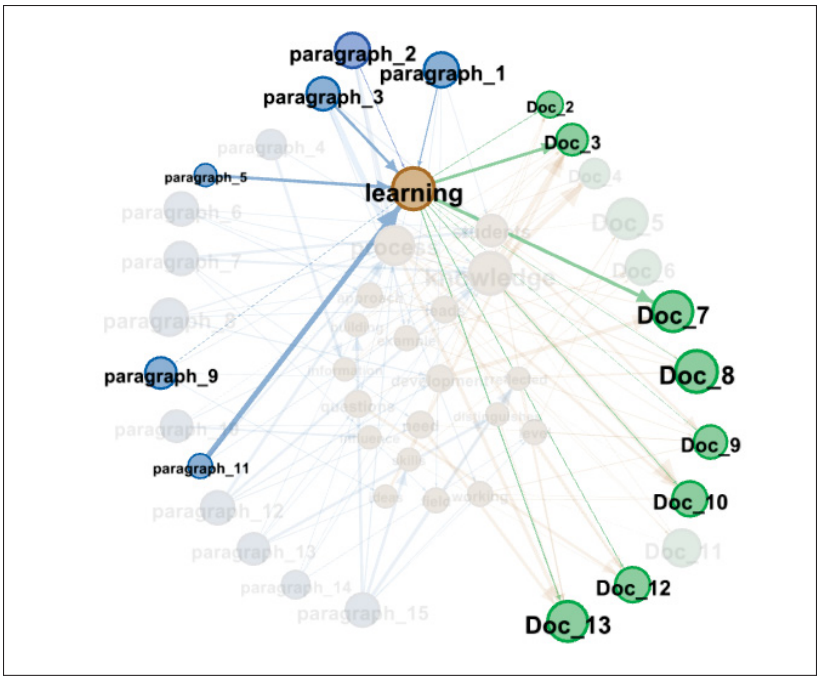


Figure 4.18 Exemple de manipulation d'un réseau de mots dans le logiciel Gephi.

avons montré uniquement l'analyse avec les références générales (RG). La deuxième fois, nous avons inclus les références spécialisées (RS). Nous présentons par la suite les interprétations condensées de ces deux rencontres. Étant donné la nature des PVM, ces interprétations ont été négociées entre le chercheur et les enseignants du cours (Bygstad & Munkvold, 2007).

Concernant les figures de la section 4.1.1 (*Nombre de documents couverts par chaque dissertation*) et les graphiques de la section 4.1.3 (*Réseaux de mots des dissertations*), nous condensons les résultats du PVM de la façon suivante :

- La dissertation 1 présente des paragraphes qui sont très peu alignés ; la densité d'alignement est donc plus faible par rapport aux autres dissertations. Cette caractéristique est liée à la couverture d'information des RG et des RS dans une dissertation.
- Lorsqu'on examine l'alignement avec les documents des RS, nous voyons une fenêtre de valeurs plus ample, dénotant ainsi une meilleure couverture de la terminologie. Ce patron est également observé lorsqu'on compare toutes les dissertations.

- Les dissertations 2, 3 et 4 semblent avoir un flux de l'alignement avec les RG et les RS ; elles sont donc plus cohérentes et consistantes. La dissertation 1 semble ne pas présenter ces dernières caractéristiques. Conséquemment, les trois autres dissertations sont mieux élaborées.

Concernant les figures de la section 4.1.2, (*L'influence des RG et des RS sur la production des dissertations*), le PVM a révélé que les stratégies de similarité cosinus et coefficient de Dice évaluent mieux la dissertation 1 que la dissertation 4. Ceci peut être expliqué par la symétrie de ces deux mesures. En effet, elles placent la dissertation, les RG et les RS au même niveau ; il devient dès lors impossible de savoir quel document est plus similaire, comme le soulignent, d'ailleurs, Shivakumar & Garcia-Molina (1995). Une approche asymétrique est capable de faire cette distinction et de fournir une analyse plus cohérente de la couverture.

Les participants du PVM ont aussi observé qu'avec notre approche, il est aussi possible de voir l'influence de certains documents des RG et des RS sur la production des dissertations.

Lors de la dernière séance du PVM, les enseignants nous ont informés qu'une évaluation des documents les plus utilisés dans l'élaboration des dissertations avait été tenue aupres des étudiants dont nous avons analysé les dissertations. Nous avons utilisé cette évaluation pour corroborer en fait que les documents les mieux alignés avec les dissertations correspondent aussi à ceux qui ont été évalués par les étudiants comme étant les plus utilisés. Nous avons identifié ces documents comme les références considérées (RC) dans le diagramme de la fig. 4.1 repris ici dans la fig. 4.19. Nous y avons ajoutons les index des références pour illustrer leur provenance. Nous présentons l'évaluation des RC dans le graphique de la fig. 4.20.

Dans la fig 4.20, nous voyons les documents qui ont été les plus utilisés par les étudiants pour la réalisation de leur dissertation. Parmi les documents visualisés, nous avons inclus dans notre analyse les documents 15, 16, 17 et 18 (qui font à la fois partie des RS) ; leur utilisation varie d'une dissertation à l'autre. En ce qui concerne les documents en orange, 9, et 11, leur utilisation apparait de façon égale pour les quatre dissertations.

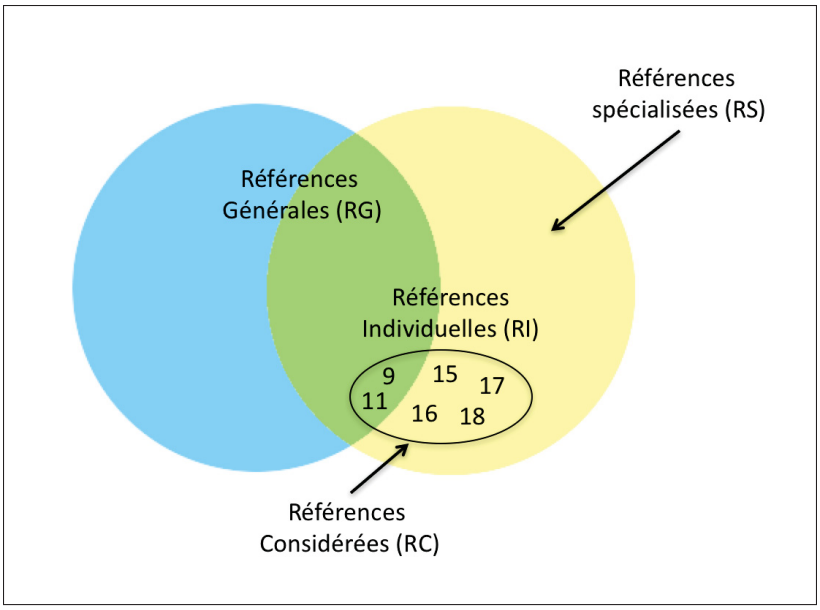


Figure 4.19 Diagramme de la distribution des documents des références et la nomenclature utilisée.

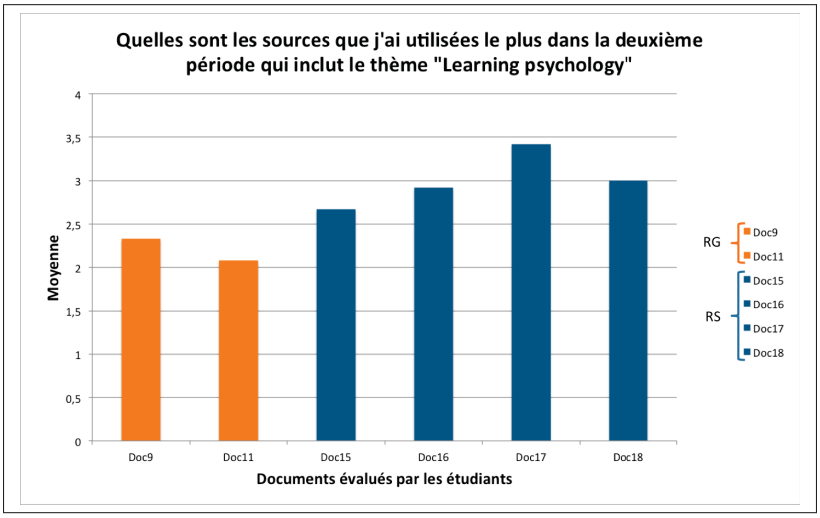


Figure 4.20 Évaluation de l'impact des références considérées par le groupe d'étudiants.

La fig 4.11 est reprise à la fig 4.21. Les documents 9 et 11 correspondent aux barres qui apparaissent en vert et en bleu foncé, respectivement.

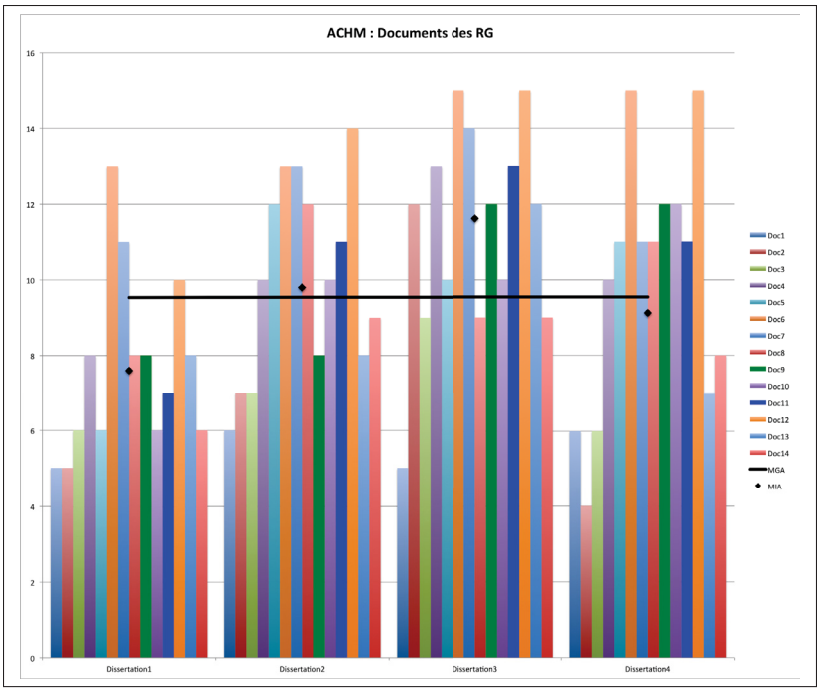


Figure 4.21 Alignements des dissertations et les documents des RG. Approche ACHM .

Comme nous pouvons le voir dans la dissertation 1, les documents 9 et 11 n'atteignent pas la MGA ; seulement le document 9 (en bleu) atteint la MIA du de la dissertation 1. Dans la dissertation 2, le document 11, en vert, surmonte la MGA et la MIA. Pour les dissertations 3 et 4, nous remarquons que les documents 9 et 11 dépassent la MGA ainsi que la MIA.

En utilisant la MGA et la MIA comme repères pour déterminer l'influence d'un document des RC sur les dissertations, nous constatons que notre analyse (fig. 4.21) fait ressortir que le document 9 et le document 11 ont une certaine influence sur les dissertations.

En ce qui concerne les documents des RG, qui sont inclus dans l'évaluation de la figure 4.20, (documents 15, 16, 17, et 18), nous nous attendions à ce que ces documents soient au-dessus de la MGA. Nous reprenons le graphe de la fig 4.12 dans la figure 4.22 ; les documents évalués par les étudiants apparaissent en couleur jaune, orange, rouge et mauve. La dissertation 1 utilise les références 16, 17 et 18. Ces trois documents atteignent la MGA et dépassent la MIA. La dissertation 2 utilise les références 16 et 18. À la fig 4.22, nous observons que le document

16 dépasse légèrement la MGA et la MIA, tandis que le document 18 est parmi les deux documents qui ont été les plus alignés avec les paragraphes de la dissertation 2. La dissertation 3 inclut les documents 15, 16 et 18 pour son élaboration. Dans la fig. 4.21, nous remarquerons que ces trois documents arrivent à dépasser la MGA et la MIA. Le document 15 est parmi les documents avec le nombre d'alignements le plus élevés pour cette dissertation. Finalement, la dissertation 4, dans la figure fig 4.22, utilise les documents 16 et 17 pour son élaboration. Ces deux documents dépassent la MGA et la MIA ; ils sont, tous les deux, les documents qui ont le plus d'alignements avec les paragraphes du document 4.

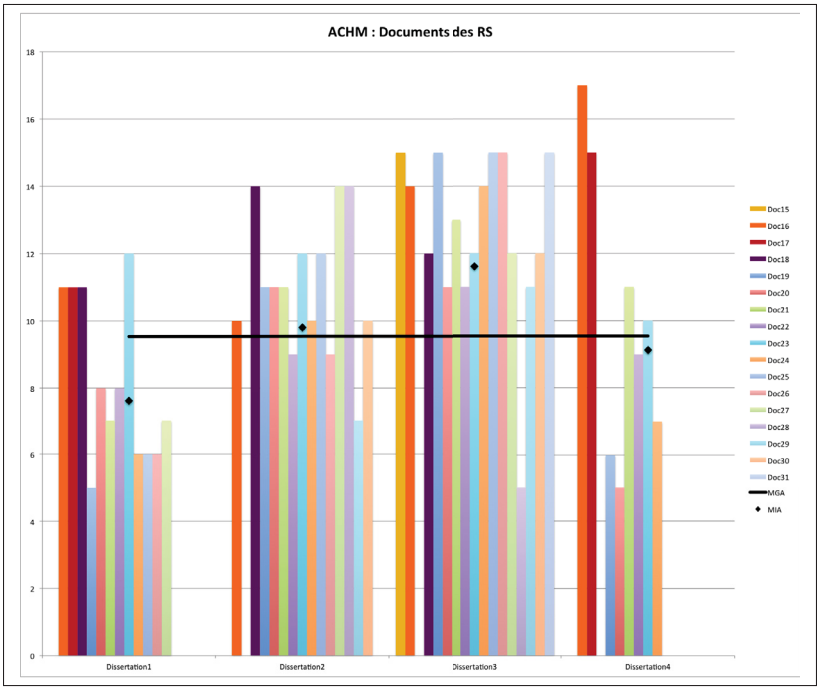


Figure 4.22 Alignements de dissertations et des documents des RS. Approche ACHM.

Cette évaluation nous a permis de corroborer que les documents des RC seraient parmi les documents ayant un nombre élevé d'alignements dans notre analyse. Il est vrai que nous n'avons pas couvert toutes les références individuelles (RI) par les étudiants pour des raisons de disponibilité des ressources, mais nous croyons que les résultats avec cet échantillon pourront être élargis pour l'ensemble des RI.

4.2 Scénario 2 : La couverture d'information dans les textes journalistiques

Dans cette section, nous reportons les résultats concernant les expériences pour la couverture d'information dans les textes journalistiques. Puisque nous avons réalisé une nouvelle annotation sur une partie du corpus *novelty TREC*, nous présentons aussi les résultats de l'accord entre les annotateurs. Nous décrivons une évaluation pour chaque thématique. Nous avons utilisé le résultat de cette annotation pour évaluer notre deuxième proposition en termes quantitatifs.

4.2.1 L'accord des annotateurs sur le corpus

Pour l'interprétation des valeurs du coefficient κ, nous utilisons le tableau 4.2. Ce tableau est proposé par McHugh (2012).

Tableau 4.2 Interprétation des valeurs du coefficient κ

Valeur κ	Niveau d'accord	% de données qui sont fiables
0.0 – 0.20	Null	0 – 4%
0.21 – 0.39	Minimum	4 – 15 %
0.40 – 0.59	Faible	15 – 35 %
0.60 – 0.79	Modéré	35 – 63 %
0.80 – 0.90	Fort	64 – 81 %
Plus de 0.90	Presque parfait	82 – 100 %

Puisque nous avons eu quatre annotateurs pour étiqueter le corpus, nous avons, au départ, utilisé le coefficient κ Fleiss. Ce dernier est une adaptation du coefficient κ original pour des contextes où il y a plus de deux annotateurs (Pustejovsky & Stubbs, 2012). Comme nous l'observons au tableau 4.3, les valeurs *Kappa* Fleiss sont très basses (avec faible accord selon le tableau 4.2) ou presque nulles dans le cas de quelques thématiques. Nous avons aussi testé un autre coefficient pour exprimer l'accord entre les annotateurs, *ICC (Intraclass Correlation Coefficient)*. Les valeurs de deux coefficients étaient similaires (voir tableau 4.3).

Nous avons observé dans quelque cas, que les valeurs du coefficient κ Fleiss et du ICC étaient très faibles. En regardant les annotations nous avons remarqué un accord plus élevé de ce que ces deux coefficients exprimaient. Nous avons décidé d'analyser la distribution des données des annotations pour déterminer quel était le problème.

Pour ce faire, nous avons utilisé la moyenne et la médiane. Puisque les annotations contiennent des valeurs binaires seulement, nous ajoutons une analyse de la distribution valeurs (les 1s et les 0s) en pourcentages de l'ensemble des données. Nous savons qu'une moyenne exacte de 0.5 exprimerait que la moitié des valeurs est de 0 et l'autre moitié de 1 ; dans ce cas la médiane serait aussi de 0.5. Si la valeur de moyenne est majeure à 0.5, les valeurs de l'annotation seraient majoritairement des 1 ; la valeur de la médiane serait dont égale à 1. Une valeur inférieure à 0.5 de la moyenne exprimerait que la plupart des valeurs contenues dans l'échantillon sont des 0 ; la valeur de la médiane serait égale à 0. Une moyenne égale à 1 exprime que toutes les valeurs de l'annotation seraient égales à 1 ; la valeur de la médiane serait 1.

Tableau 4.3 Mesures d'accord entre annotateurs et analyse de distribution de donnnées.

					Proportion en %	
Topic	Kappa Fleiss	ICC	Moyenne	Médiane	1	0
1	0.17	0.18	0.9	1	90.32	9.68
2	0	0	0.99	1	99	1
3	0.757	0.767	0.44	0	44	56
4	0.46	0.461	0.95	1	94.58	4.58
5	0.166	0.19	0.9	1	90.21	9.78

Dans le tableau 4.3, nous constatons que la thématique 2 présente une valeur nulle pour le coefficient κ et aussi pour le coefficient ICC. Selon l'interprétation des valeurs du coefficient *kappa* (dans le tableau 4.2) la valeur obtenue pour la thématique 2 exprime un accord nul entre les annotateurs. Le même cas d'interprétation s'applique pour les thématiques 1 et 5. La thématique 4 présente un accord faible entre les annotateurs ; seulement la thématique 3 présente un accord modéré.

En observant le tableau 4.3, nous constatons que la thématique 2 présente une moyenne de 0.99, ce qui implique que la majorité des valeurs sont des 1. Nous pouvons constater cela aussi à l'aide de la médiane qui est égale à 1. Dans le même tableau, nous avons ajouté deux colonnes pour exprimer la proportion des 1 et des 0 dans l'annotation. Dans le cas de la thématique 2, le 99% des valeurs sont des 1. Si le 99% des valeurs sont des 1, il y a forcement un accord élevé. Pourtant, les deux coefficients ne semblent pas l'exprimer.

Tel que décrit dans la méthodologie, nous avons utilisé les coefficients proposés par Cicchetti & Feinstein (1990) pour exprimer l'accord positif et négatif entre les annotateurs. Ces coefficients répondent aux deux paradoxes du coefficient Kappa identifiés dans (Feinstein & Cicchetti, 1990). Ces paradoxes surviennent lorsqu'il existe un accord élevé entre les annotateurs (voir section 3.3.6 pour plus de détails). De plus, ces coefficients permettent d'évaluer l'accord des étiquettes positives et négatives d'un même annotateur, ce qui nous permettra de vérifier la compréhension de ce qui est classifié comme nouveau ou pas par les annotateurs.

Nous avons illustré les valeurs issues de ces coefficients dans des tableaux. Pour chaque tableau, nous présentons le coefficient P_{pos} pour mesurer l'accord positif, le coefficient P_{neg} pour mesurer l'accord négatif entre les annotateurs, et la valeur du coefficient κ. Dans les cas où nous jugerons le nécessaire, nous ajoutons aux tableaux les valeurs de p_0 et p_e qui sont utilisées pour le calcul du coefficient κ.

Dans la colonne *Annotateurs* de chaque tableau, nous avons inscrit la combinaison de chacun de nos experts. Par exemple, la combinaison 1 – 1 correspond à la combinaison du même annotateur ; nous y attendons un accord parfait. La combinaison 1 – 4 exprime l'accord entre l'annotateur 1 et l'annotateur 4.

La thématique 1 (62 phrases) aborde le sujet des *tests nucléaires au Pakistan*. Les valeurs de 1 dans le tableau 4.4 correspondent à l'accord d'un annotateur avec lui même. Nous pouvons remarquer que les annotateurs 1 et 2 ont les plus grandes valeurs d'accord pour les classes positives et négatives. Pour cette thématique en particulier, il existe un fort accord entre les

Tableau 4.4 Valeurs des coefficients d'accord pour la thématique 1

Annotateurs	P_{pos}	P_{neg}	κ	p_0	p_e
1 – 1	1	1	1	1	0.775
1 – 2	0.964	0.666	0.635	0.935	0.823
1 – 3	0.920	0.181	0.119	0.854	0.835
1 – 4	0.859	0.117	-0.021	0.758	0.763
2 – 2	1	1	1	1	0.879
2 – 3	0.957	0.285	0.243	0.919	0.893
2 – 4	0.882	0	-0.0980	0.79	0.809
3 – 3	1	1	1	1	0.907
3 – 4	0.928	0.333	0.281	0.87	0.82
4 – 4	1	1	1	1	0.751

annotateurs sur ce qui est nouveau. Par contre, il n'existe pas de consensus sur ce qui n'est pas nouveau.

En observant le tableau 4.4, nous remarquerons une majorité de valeurs du coefficient κ très bas. Par exemple entre les annotateurs 1 – 3 la valeur du coefficient κ est de *0.119* (un accord très bas selon l'interprétation du tableau 4.2). Nous remarquons aussi des valeurs négatives entre les annotateurs 1 – 4 (*-0.021*). Une valeur négative du coefficient κ représente un accord (ou désaccord) encore pire dont l'on peut attendre (McHugh, 2012). Les valeurs négatives apparaissent quand p_e est supérieur à p_0.

La thématique 2 aborde le sujet de l' *Embargo cubain imposé par les États-Unis*. Cette thématique comporte 25 phrases. Encore une fois, nous avons décidé de montrer l'accord entre le même annotateur, voir les index 1 – 1, 2 – 2, 3 – 3, et 4 – 4 du tableau 4.5. Nous remarquons des valeurs *NaN*, ce qui veut dire une division par zéro ; cela dénote que l'annotateur n'a pas trouvé de classes négatives, donc il a trouvé seulement des phrases apportant de la nouvelle information. Ceci est le cas pour les annotateurs 1, 2 et 4. L'annotateur 3 a considéré la présence de quelques phrases n'apportant pas de nouveauté. La présence de valeurs nulles pour les coefficients P_{neg} et κ indique qu'il n'y a pas d'accord entre les annotateurs sur l'annotation des classes négatives. Cette thématique présente les valeurs les plus élevées en termes d'accord

entre les annotateurs de la classe positive. Nous observons aussi que lorsque les valeurs de p_0 (pourcentage total d'accord entre les annotateurs) et p_e (la proportion d'accord espérée) sont égales à 1, le coefficient κ n'arrive pas à couvrir ce comportement. Quand la valeur de p_0 et p_e est la même, le coefficient κ est zéro. La seule façon de constater l'accord de la classe positive et de la classe négative entre les annotateurs, est de regarder les valeurs de P_{pos} et de P_{neg} .

Tableau 4.5 Valeurs des coefficients d'accord pour la thématique 2

Annotateurs	P_{pos}	P_{neg}	κ	p_0	p_e
1 – 1	1	*NaN*	*NaN*	1	1
1 – 2	1	*NaN*	*NaN*	1	1
1 – 3	0.979	0	0	0.96	0.96
1 – 4	1	*NaN*	*NaN*	1	1
2 – 2	1	*NaN*	*NaN*	1	1
2 – 3	0.979	0	0	0.96	0.96
2 – 4	1	*NaN*	*NaN*	1	1
3 – 3	1	1	1	1	0.923
3 – 4	0.979	0	0	0.96	0.96
4 – 4	1	*NaN*	*NaN*	1	1

La thématique 3 (25 phrases) aborde *la mort de la princesse Diana*. Le tableau 4.6 montre que cette thématique présente les valeurs les plus élevées après la thématique 2. Nous remarquons également que l'étiquetage entre les annotateurs 1 et 3 est parfait. Les valeurs les plus basses surviennent avec la combinaison des annotateurs 1 – 4, 2 – 4 et 3 – 4.

La thématique 4 (119 phrases) aborde *la vie et mort de Charles Schultz*. En regardant le tableau 4.7, nous remarquons que cette thématique présente aussi un équilibre sur ce qui est nouveau et ce qui ne l'est pas. D'après les résultats contenus dans les tableaux 4.6 et 4.7, il semble que le type de nouvelle est déterminant pour l'identification de la nouveauté. Identifier la nouveauté d'une histoire qui porte sur la vie d'un personnage semble plus facile pour un annotateur. La quantité d'information pourrait aussi être déterminante pour l'identification de la nouveauté. Plus d'information implique un travail cognitif plus exigent pour garder en mémoire toute l'information déjà lue.

Tableau 4.6 Valeurs des coefficients d'accord pour la thématique 3

Annotateurs	P_{pos}	P_{neg}	κ
1 – 1	1	1	1
1 – 2	0.956	0.962	0.919
1 – 3	**1**	**1**	**1**
1 – 4	0.761	0.827	0.595
2 – 2	1	1	1
2 – 3	0.956	0.962	0.919
2 – 4	0.7	0.8	0.503
3 – 3	1	1	1
3 – 4	0.761	0.827	0.595
4 – 4	1	1	1

Tableau 4.7 Valeurs des coefficients d'accord pour la thématique 4

Annotateurs	P_{pos}	P_{neg}	κ
1 – 1	1	1	1
1 – 2	0.973	0.5	0.474
1 – 3	0.973	0.4	0.373
1 – 4	0.973	0.4	0.373
2 – 2	1	1	1
2 – 3	0.973	0.5	0.474
2 – 4	0.991	0.982	0.833
3 – 3	1	1	1
3 – 4	0.964	0.2	0.164
4 – 4	1	1	1

La thématique 5 aborde une série de *tremblements de terre en Turquie*. Cette thématique comporte 23 phrases. Tous les tremblements de terre sont différents et ont lieu en différentes régions de la Turquie. Nous aurions pu attendre une identification facile de la nouveauté dans ce cas, puisque cette nouvelle contient des chiffres qui expriment l'intensité des tremblements de terre ainsi que le nom de la ville où le phénomène s'est produit. Par contre, cette thématique présente aussi des valeurs faibles d'accord pour la classe négative entre les annotateurs (voir le tableau 4.8). Nous observons dans ce tableau des valeurs nulles pour la classe négative (p_{neg}) ; la valeur nulle exprime qu'il n'y a pas d'accord entre les annotateurs sur les classes négatives

qu'ils ont identifiées. Finalement, l'accord sur ce qui est nouveau (exprimé par la valeur de p_{pos}) est toujours plus élevé par rapport à ce qui n'est pas nouveau.

Tableau 4.8 Valeurs des coefficients d'accord pour la thématique 5

Annotateurs	P_{pos}	P_{neg}	κ	p_0	p_e
1–1	1	1	1	1	0.916
1–2	0.954	0	-0.045	0.913	0.916
1–3	1	1	1	1	0.916
1–4	0.871	0.285	0.228	0.782	0.718
2–2	1	1	1	1	0.916
2–3	0.954	0	-0.045	0.913	0.916
2–4	0.82	0	-0.08	0.695	0.718
3–3	1	1	1	1	0.916
3–4	0.871	0.285	0.228	0.782	0.718
4–4	1	1	1	1	0.614

De la même façon que pour la thématique 2, dans la thématique 5 nous observons des valeurs négatives du coefficient κ (voir le tableau 4.8 pour les annotateurs 1 – 2, 2 – 3 et 2 – 4). Les valeurs négatives découlent du fait que la valeur de p_e est supérieure à la valeur de p_0.

4.2.2 Évaluation

Pour être en mesure d'évaluer la couverture d'information entre un référent et un sujet de comparaison, nous avons utilisé notre mesure à base de patrons linguistiques. Pour comparer les résultats de notre mesure, nous avons aussi utilisé la similarité cosinus et le coefficient de Dice, comme mesures *référentielles*. Nous avons établi que si une phrase du sujet de comparaison n'est pas couverte dans le référent, cela serait interprété comme une phrase qui apporte de la nouvelle information. Nous avons utilisé le corpus *novelty TREC* dans notre version annotée pour réaliser cet tâche (voir la section 3.3.2 du chapitre 3).

En ce qui concerne l'annotation des experts, nous considérons une phrase comme ayant de la nouvelle information quand celle-ci a été classifiée comme telle par 3 ou 4 annotateurs.

Pour chacune des cinq thématiques, nous avons une liste contenant la classification de toutes phrases. Cette liste indique si une phrase comporte de la nouvelle information ou non. Nous nous servons de ces listes pour faire une évaluation en termes de *précision, rappel, exactitude et mesure-F*.

Notre mesure de couverture à base de patrons linguistiques est conçue de telle façon que nous pouvons donner plus de poids à un patron en particulier. Nous avons conduit des expériences avec différentes valeurs pour le paramètre α.

Dans la série de tableaux que nous présentons par la suite, nous avons mis en relief les meilleures valeurs pour la mesure-F. La colonne *seuil* correspond à la valeur que nous avons fixée pour déterminer si une phrase était couverte ou pas. Si la valeur d'une des trois mesures était plus petite que ce seuil, nous considérons que la phrase correspondante amène de la nouvelle information. Rappelons que la taille du référent dans cette expérience est de taille variable, puisque nous y ajoutons une phrase chaque fois que celle-ci ne dépasse pas le seuil fixé.

Le tableau 4.9 montre les résultats pour la thématique 1 avec la mesure de similarité cosinus (qui est symétrique). La thématique 1 aborde « *les tests nucléaires au Pakistan* ». Les résultats sont les meilleurs avec un seuil à 0.7 et le même scénario se répète avec le seuil à 0.5 ; la valeur de la mesure-F est de 0.954. Un changement est observé avec le seuil à 0.3.

Tableau 4.9 Évaluation pour la thématique 1. Cosine

Seuil	**Exactitude**	**Précision**	**Rappel**	**Mesure-F**
0.7	**0.918**	**0.929**	**0.981**	**0.954**
0.5	0.918	0.929	0.981	0.954
0.3	0.819	0.957	0.833	0.891

Le tableau 4.10 montre les résultats avec le coefficient de Dice (également symétrique) pour la thématique 1. Les meilleurs résultats sont obtenus avec le seuil à 0.7 ; la valeur de mesure-F est de 0.954. Les autres seuils fixés génèrent une performance plutôt faible.

Tableau 4.10 Évaluation pour la thématique 1. Dice

Seuil	Exactitude	Précision	Rappel	Mesure-F
0,7	**0.918**	**0.929**	**0.981**	**0.954**
0,5	0.655	0.971	0.629	0.764
0,3	0.18	1	0.074	0.137

Le tableau 4.11 montre les valeurs d'*exactitude, de précision, de rappel,* et *de mesure-F* pour la thématique 1 et notre mesure. Nous voyons dans ce tableau que les meilleurs résultats correspondent à α ayant plus de poids pour le patron NV–NV, et un seuil de 0.7 ; la valeur de la mesure-F est de 0.946. La différence entre les approches symétriques et notre approche est très légère (0.008). **L'avantage de notre approche est sa capacité d'expliquer d'où vient la nouveauté ;** dans ce cas, la nouveauté vient du patron NV–NV, qui représente grossièrement la relation du sujet de la phrase. Les résultats sont bons, si l'on compare l'accord d'étiquetage du corpus de la part de nos experts.

Tableau 4.11 Évaluation pour la thématique 1. ACHM

Patron – α	Seuil	Exactitude	Précision	Rappel	Mesure-F
Tous $\alpha = 1$	0.7	0.885	0.927	0.944	0.935
Tous $\alpha = 1$	0.5	0.672	1	0.629	0.772
Tous $\alpha = 1$	0.3	0.245	0.9	0.166	0.281
PN–PN $\alpha = 10$	0.7	0.885	0.943	0.925	0.934
PN–PN $\alpha = 10$	0.5	0.508	0.928	0.481	0.634
PN–PN $\alpha = 10$	0.3	0.213	0.875	0.129	0.225
VN–VN α =10	0.7	0.885	0.927	0.944	0.935
VN–VN $\alpha = 10$	0.5	0.721	1	0.685	0.813
VN–VN $\alpha = 10$	0.3	0.393	1	0.314	0.478
VPN–NV $\alpha = 10$	0.7	0.885	0.927	0.944	0.935
VPN–NV $\alpha = 10$	0.5	0.885	0.912	0.962	0.936
VPN–NV $\alpha = 10$	0.3	0.262	0.909	0.185	0.307
NV–NV $\alpha = 10$	0.7	**0.901**	**0.913**	**0.981**	**0.946**
NV–NV $\alpha = 10$	0.5	0.672	0.925	0.685	0.787
NV–NV $\alpha = 10$	0.3	0.327	0.933	0.259	0.405
VN–NV $\alpha = 10$	0.7	0.885	0.927	0.944	0.935
VN–NV $\alpha = 10$	0.5	0.672	1	0.629	0.772
VN–NV $\alpha = 10$	0.3	0.245	0.9	0.166	0.281

Le tableau 4.12 présente les résultats pour la thématique 2 avec la mesure de similarité cosinus. La thématique 2 aborde « *l'embargo cubain imposé par les États-Unis* ». Les meilleurs résultats apparaissent avec les seuils à 0.7 et 0.5 ; la valeur de mesure-F est de 1. Le seuil à 0.3 comporte encore de bons résultat.

Tableau 4.12 Évaluation pour la thématique 2. Cosine

Seuil	Exactitude	Précision	Rappel	Mesure-F
0.7	**1**	**1**	**1**	**1**
0.5	1	1	1	1
0.3	0.958	1.0	0.958	0.978

Le tableau 4.13 présente les résultats pour la thématique 2 avec le coefficient de Dice. Les meilleurs résultats sont obtenus avec un seuil à 0.7 ; la valeur de la mesure-F est de 1, comme dans l'approche similarité cosinus. Cette approche comporte des résultats très bas avec un seuil de 0.3.

Tableau 4.13 Évaluation pour la thématique 2. Dice

Seuil	Exactitude	Précision	Rappel	Mesure-F
0.7	**1**	**1**	**1**	**1**
0.5	0.791	1	0.791	0.883
0.3	0.083	1	0.083	0.153

Le tableau 4.14 présente les résultats d'*exactitude, de précision, de rappel,* et *de mesure-F* pour la thématique 2 avec notre mesure. Dans les résultats, nous remarquons la même performance par rapport aux mesures symétriques quand le paramètre α est pondéré également pour tous les patrons ; la valeur de mesure-F est de 1.

Pondérer tous les patrons avec le même poids rend notre mesure similaire aux mesures symétriques. Par contre, nous remarquons que des valeurs pour la mesure-F de 1 apparaissent aussi pour les patrons NV–NV et VPN–NV. Ces deux patrons pourraient se dire équivalents, car NV–NV vise à attraper grossièrement la relation du sujet et VPN–NV représente, sommai-

rement, la conversion de la voix passive à la voix active. Pour cette thématique, les valeurs de précision sont, la plupart du temps, égales à 1 (sauf pour le patron VN–NV et les seuils 0.7 et 0.3) ; cela veut dire que notre stratégie est capable de classifier toutes les nouvelles phrases comme les experts l'ont fait. Nous devrons donc voir la valeur du rappel ; celle-ci exprime les classes positives que notre algorithme a classifiées comme telles et qui correspondent aussi aux classes positives classifiées par les experts.

Tableau 4.14 Évaluation pour la thématique 2

Patron – α	**Seuil**	**Exactitude**	**Précision**	**Rappel**	**Mesure-F**
Tous $\alpha = 1$	0.7	**1**	**1**	**1**	**1**
Tous $\alpha = 1$	0.5	0.875	1	0.875	0.933
Tous $\alpha = 1$	0.3	0.25	1	0.25	0.4
PN–PN $\alpha = 10$	0.7	0.875	1	0.875	0.933
PN–PN $\alpha = 10$	0.5	0.625	1	0.625	0.769
PN–PN $\alpha = 10$	0.3	0.083	1	0.083	0.153
VN–VN $\alpha = 10$	0.7	0.958	1	0.958	0.978
VN–VN $\alpha = 10$	0.5	0.833	1	0.833	0.909
VN–VN $\alpha = 10$	0.3	0.541	1	0.541	0.702
VPN–NV $\alpha = 10$	0.7	**1**	**1**	**1**	**1**
VPN–NV $\alpha = 10$	0.5	0.875	1	0.875	0.933
VPN–NV $\alpha = 10$	0.3	0.25	1	0.25	0.4
NV–NV $\alpha = 10$	0.7	**1**	**1**	**1**	**1**
NV–NV $\alpha = 10$	0.5	0.875	1	0.875	0.933
NV–NV $\alpha = 10$	0.3	0.416	1	0.416	0.588
VN–NV $\alpha = 10$	0.7	0.885	0.927	0.944	0.935
VN–NV $\alpha = 10$	0.5	0.672	1	0.629	0.772
VN–NV $\alpha = 10$	0.3	0.245	0.9	0.166	0.281

Dans le tableau 4.14, nous remarquons des valeurs de 1 pour la mesure-F. Selon le tableau 4.5, qui montre l'accord d'étiquetage entre les experts, presque toutes les phrases apportent de la nouvelle information. Les trois stratégies que nous avons utilisées sont capables d'identifier les phrases apportant de la nouvelle information. Dans le cas de notre mesure, si le paramètre α est le même pour tous les patrons, nous obtiendrons les mêmes résultats qu'une mesure symétrique. De plus, notre mesure est capable d'identifier de la nouvelle information si les patrons NV–NV et VPN–NV sont pondérés. Cette pondération permettrait aussi d'appliquer un

filtre lors du calcul de la couverture d'information, qui prioriserait une relation grammaticale en particulier. Par exemple, le lecteur pourrait s'intéresser à capturer toutes les nouvelles où Barack Obama apparaît comme l'entité déclenchant tous les événements d'une thématique. Pour obtenir ce résultat, la pondération du patron NV–NV devrait être élevée.

Le tableau 4.15 montre les résultats de l'évaluation pour la thématique 3 avec la mesure de similarité cosinus. La thématique 3 aborde « *la mort de la princesse Diana* ». Les meilleurs résultats sont obtenus avec un seuil à 0.7 ; la valeur de la mesure-F est de 0.916.

Tableau 4.15 Évaluation pour la thématique 3. Cosine

Seuil	Exactitude	Précision	Rappel	Mesure-F
0.7	**0.916**	**0.846**	**1**	**0.916**
0.5	0.916	0.909	0.909	0.909
0.3	0.458	0.458	1	0.628

Le tableau 4.16 montre les résultats pour la thématique 3 avec le coefficient Dice. Comme nous pouvons le voir, la meilleur valeur pour la mesure-F est obtenue avec un seuil à 0.7.

Tableau 4.16 Évaluation pour la thématique 3. Dice

Seuil	Exactitude	Précision	Rappel	Mesure-F
0.7	**0.958**	**0.916**	**1**	**0.956**
0.5	0.75	0.857	0.545	0.666
0.3	0.625	1	0.181	0.307

Le tableau 4.17 montre les résultats d'*exactitude, de précision, de rappel,* et *de mesure-F* pour la thématique 3. Dans ce cas, le patron PN–PN obtient les meilleurs résultats ; la valeur de la mesure-F est de 1. Ce patron vise à attraper grossièrement les adjoints du prédicat ; ces éléments peuvent représenter des locatifs qui expriment l'endroit où l'événement aurait eu lieu. D'autres patrons obtiennent des valeurs de mesure-F qui sont aussi bonnes. Par exemple, le patron VPN–NV, le patron VN–NV, les deux avec une valeur de 0.956 pour la mesure-F. Dans le tableau tableau 4.17, nous pouvons remarquer aussi des valeurs parfaites en termes de précisions, mais

aussi des valeurs basses pour le rappel. Cela indique que notre approche classsifie plus de phrases contenant de la nouvelle information que les annotateurs l'aurait déterminé.

Tableau 4.17 Évaluation pour la thématique 3

Patron – α	**Seuil**	**Exactitude**	**Précision**	**Rappel**	**Mesure-F**
Tous $\alpha = 1$	0.7	0.958	0.916	1	0.956
Tous $\alpha = 1$	0.5	0.833	1	0.636	0.777
Tous $\alpha = 1$	0.3	0.625	1	0.181	0.307
PN–PN $\alpha = 10$	0.7	**1**	**1**	**1**	**1**
PN–PN $\alpha = 10$	0.5	0.791	1	0.545	0.705
PN–PN $\alpha = 10$	0.3	0.583	1	0.09	0.166
VN–VN $\alpha = 10$	0.7	0.875	0.9	0.818	0.857
VN–VN $\alpha = 10$	0.5	0.791	0.875	0.636	0.736
VN–VN $\alpha = 10$	0.3	0.625	1	0.181	0.307
VPN–NV $\alpha = 10$	0.7	0.958	0.916	1	0.956
VPN–NV $\alpha = 10$	0.5	0.833	1	0.636	0.777
VPN–NV $\alpha = 10$	0.3	0.625	1	0.181	0.307
NV–NV $\alpha = 10$	0.7	0.875	0.833	0.909	0.869
NV–NV $\alpha = 10$	0.5	0.875	1	0.727	0.842
NV–NV $\alpha = 10$	0.3	0.666	1	0.272	0.428
VN–NV $\alpha = 10$	0.7	0.958	0.916	1	0.956
VN–NV $\alpha = 10$	0.5	0.833	1	0.636	0.777
VN–NV $\alpha = 10$	0.3	0.625	1	0.181	0.307

Le tableau 4.18 montre les résultats pour la thématique 4 avec la mesure de similarité cosinus. La thématique 4 aborde « *la vie et mort de Charles Schultz* ». Les meilleurs résultats pour la mesure-F sont obtenus avec un seuil de 0.7 ; la valeur de la mesure-F est de 0.973.

Tableau 4.18 Évaluation pour la thématique 4. Cosine

Seuil	**Exactitude**	**Précision**	**Rappel**	**Mesure-F**
0.7	**0.949**	**0.949**	**1**	**0.973**
0.5	0.949	0.949	1	0.973
0.3	0.949	0.949	1	0.973

Le tableau 4.19 montre les résultats pour la thématique 4 avec le coefficient de Dice. Les meilleurs résultats pour la mesure-F sont obtenus avec un seuil de 0.7 ; la valeur est de 0.982.

Tableau 4.19 Évaluation pour la thématique 4. Dice

Seuil	Exactitude	Précision	Rappel	Mesure-F
0.7	**0.966**	**0.965**	**1**	**0.982**
0.5	0.881	0.962	0.91	0.935
0.3	0.144	0.866	0.116	0.204

Le tableau 4.20 montre les valeurs d'*exactitude, de précision, de rappel,* et *de mesure F* pour la thématique 4 avec notre mesure. Nous observons que notre mesure a des résultats similaires aux mesures symétriques lorsque les patrons sont pondérés par le même poids du paramètre α. Le même résultat, en termes de mesure-F, est obtenu avec le patron VPN-NV ; la valeur est de 0.982. Cette thématique aborde l'histoire d'un personnage (Charles Schulz) et nous observons que les valeurs de précision ne varient pas significativement avec le patron NV–NV, même si ce dernier ne présente pas les valeurs les plus élevées de l'ensemble des patrons.

Tableau 4.20 Évaluation pour la thématique 4

Patron – α	Seuil	Accuracy	Precision	Rappel	Mesure-F
Tous $\alpha = 1$	0.7	**0.966**	**0.965**	**1**	**0.982**
Tous $\alpha = 1$	0.5	0.576	0.955	0.58	0.722
Tous $\alpha = 1$	0.3	0.144	0.866	0.116	0.204
PN–PN $\alpha = 10$	0.7	0.906	0.963	0.937	0.95
PN–PN $\alpha = 10$	0.5	0.525	0.951	0.526	0.678
PN–PN $\alpha = 10$	0.3	0.144	0.923	0.107	0.192
VN–VN $\alpha = 10$	0.7	0.923	0.955	0.964	0.96
VN–VN $\alpha = 10$	0.5	0.669	0.962	0.678	0.795
VN–VN $\alpha = 10$	0.3	0.254	0.928	0.232	0.371
VPN–NV $\alpha = 10$	0.7	**0.966**	**0.965**	**1**	**0.982**
VPN–NV $\alpha = 10$	0.5	0.576	0.955	0.58	0.722
VPN–NV $\alpha = 10$	0.3	0.144	0.866	0.116	0.204
NV–NV $\alpha = 10$	0.7	0.906	0.954	0.946	0.95
NV–NV $\alpha = 10$	0.5	0.644	0.96	0.651	0.776
NV–NV $\alpha = 10$	0.3	0.22	0.916	0.196	0.323
VN–NV $\alpha = 10$	0.7	0.966	0.965	1	0.982
VN–NV $\alpha = 10$	0.5	0.576	0.955	0.58	0.722
VN–NV $\alpha = 10$	0.3	0.144	0.866	0.116	0.204

Le tableau 4.21 présente les résultas pour la thématique 5 et la mesure de similarité cosinus. La thématique 5 aborde des « *tremblements de terre en Turquie* fg Les meilleur résultats sont obtenus avec un seuil à 0.7 ; la valeur de la mesure-F est de 0.976.

Tableau 4.21 Évaluation pour la thématique 5. Cosine

Seuil	**Exactitude**	**Précision**	**Rappel**	**Mesure-F**
0.7	**0.954**	**0.954**	**1**	**0.976**
0.5	0.954	0.954	1	0.976
0.3	0.954	0.954	1	0.976

Le tableau 4.22 montre les résultats pour la thématique 5 et le coefficient de Dice. Les meilleur résultats sont obtenus avec un seuil à 0.7 ; la valeur de la mesure-F est de 1.

Tableau 4.22 Évaluation pour la thématique 5. Dice

Seuil	**Exactitude**	**Précision**	**Rappel**	**Mesure-F**
0.7	**1**	**1**	**1**	**1**
0.5	0.772	1	0.761	0.864
0.3	0.136	1	0.095	0.173

En ce qui concerne la thématique 5, le tableau 4.23 montre les valeurs d'*exactitude, de précision, de rappel,* et *de mesure F*. Nous observons que les meilleurs résultats sont atteints avec les patrons VN–VN, VPN-NV et VN–NV. Les mêmes résultats sont obtenus si le paramètre α est pondéré également pour tous les patrons. La différence réside toujours dans le patron qui est pondéré et qui vise à attraper sommairement une relation grammaticale fournissant aussi un indice de l'origine de la nouveauté.

Tableau 4.23 Évaluation pour la thématique 5

Patron – α	**Seuil**	**Exactitude**	**Precision**	**Rappel**	**Mesure-F**
Tous $\alpha = 1$	0.7	**0.909**	**1**	**0.904**	**0.95**
Tous $\alpha = 1$	0.5	0.636	1	0.619	0.764
Tous $\alpha = 1$	0.3	0.318	1	0.285	0.444
PN–PN $\alpha = 10$	0.7	0.818	1	0.809	0.894
PN–PN $\alpha = 10$	0.5	0.636	1	0.619	0.764
PN–PN $\alpha = 10$	0.3	0.272	1	0.238	0.384
VN–VN $\alpha = 10$	0.7	**0.909**	**1**	**0.904**	**0.95**
VN–VN $\alpha = 10$	0.5	0.818	1	0.809	0.894
VN–VN $\alpha = 10$	0.3	0.409	1	0.38	0.551
VPN–NV $\alpha = 10$	0.7	**0.909**	**1**	**0.904**	**0.95**
VPN–NV $\alpha = 10$	0.5	0.636	1	0.619	0.764
VPN–NV $\alpha = 10$	0.3	0.318	1	0.285	0.444
NV–NV $\alpha = 10$	0.7	0.863	1	0.8571	0.923
NV–NV $\alpha = 10$	0.5	0.772	1	0.761	0.864
NV–NV $\alpha = 10$	0.3	0.772	1	0.761	0.864
VN–NV $\alpha = 10$	0.7	**0.909**	**1**	**0.904**	**0.95**
VN–NV $\alpha = 10$	0.5	0.636	1	0.619	0.764
VN–NV $\alpha = 10$	0.3	0.318	1	0.285	0.444

4.3 Derniers mots sur les résultats

Après avoir présenté les résultats, nous voulons les synthétiser dans cette dernière section. En ce qui concerne le cas de la couverture d'information dans les textes d'étudiants, (voir section 4.1), le PVM nous permet de valider nos résultats directement avec les enseignants. De cette façon nous sommes arrivés à un accord sur l'interprétation de nos résultats. Cet exercice nous a aussi permis de corroborer que notre stratégie asymétrique décrit mieux l'assignation de notes des dissertations. De plus, notre stratégie utilise des informations provenant de *WordNet* qui contient des relations lexico-sémantiques comme la synonymie. Cette information aurait pu jouer un rôle important pour capturer la couverture de la terminologie dans les dissertations, les références générales et les références spécialisées.

Nous sommes encouragés à croire que notre stratégie s'applique bien dans le contexte de couverture d'information sur ce premier scénario.

Dans la section 4.2 nous montrons d'abord l'accord des annotateurs sur le corpus utilisé dans l'expérience pour le scénario 2 (la couverture d'information dans les textes journalistiques). Dans les cinq thématiques, nous remarquons un accord élevé entre les annotateurs sur ce qui est nouveau. L'accord sur le concept de ce qui n'est pas nouveau semble différer beaucoup plus entre les annotateurs que le concept de nouveauté.

Les thématiques 1 et 5 se ressemblent dans le sens où l'accord entre les annotateurs sur ce qui est nouveau et ce qui ne l'est pas. Les thématiques 1 et 5 abordent les tests nucléaires du Pakistan et les suites du tremblements de terre en Turquie respectivement.

La thématique 2 ne semble que présenter de la nouvelle information ; les annotateurs n'ont quasiment pas reporté d'information non nouvelle . La thématique 2 aborde l'embargo imposé par les États-Unis à Cuba. Cependant, ces résultats pourraient être liés au fait que certains des annotateurs ont des origines latino-américaines ; culturellement, ce sujet était plus proche par rapport aux autres. Il semble que nos annotateurs avaient déjà construit un référent de ce sujet, puisque culturellement parlant, cette thématique était plus proche que les autres. Cette connaissance a *priori* d'un évènement pourrait aussi avoir un impact sur l'identification de l'information nouvelle par les lecteurs. Nous reviendrons sur cet aspect au chapitre suivant.

Les deux thématiques 3 et 4 sont du même type, elles abordent les événements d'un personnage (*l'accident de la Princesse Diana* et *la vie et mort de Charles Schulz*). La seule différence est que la thématique 3 comporte 25 phrases alors que la thématique 4 comporte 119 phrases. La quantité d'information pourrait jouer un rôle majeur dans la détermination d'information nouvelle dans ce type de tâche. D'après nos résultats, nous croyons que dans les nouvelles qui abordent les histoires des personnages, l'information nouvelle est plus facile à détecter que dans les nouvelles qui abordent des événements reliés à des entités inanimées (chiffres, noms de pays, de villes, etc.). Des expériences en science cognitive avec un groupe d'annotateurs seraient nécessaires pour confirmer cette hypothèse.

Nous avons évalué les résultats obtenus pour les deux scénarios de façon différent. Pour la couverture de l'information des textes des étudiants, nous avons opté pour une évaluation qua-

litative, tandis que pour la couverture d'information des textes journalistiques, nous avons opté pour une évaluation quantitative. Si un PVM nous permet de concilier une interprétation qualitative des résultats avec les producteurs directs des données, une évaluation quantitative nous permet d'évaluer la précision de l'algorithme à travers les résultats empiriques.

La définition de la tâche de couverture d'information dans le contexte des textes journalistiques est, par elle-même, asymétrique. Nous avons utilisé deux mesures symétriques et notre coefficient asymétrique de couverture pour réaliser la même tâche. Nous avons remarqué que dans les cas des mesures symétriques les résultats sont très bons. Nous avons aussi démontré que notre coefficient asymétrique peut produire les mêmes résultats que les mesures symétriques avec la différence que la stratégie de pondération de patrons de notre coefficient permet d'expliquer l'origine de la nouveauté de l'information.

Notre coefficient asymétrique de couverture utilise des patrons linguistiques linéaires qui visaient à capturer certaines relations grammaticales. Le patron qui présente les meilleures valeurs en termes de la mesure-F correspond à VPN-NV. Ce patron vise à attraper le changement d'une phrase écrite à la voix passive et sa phrase correspondante à la voix active. En deuxième lieu le patron NV-NV apparaît en termes de mesure F ; le patron NV-NV vise à attraper la relation du sujet dans la phrase.

Le seuil 0.7 présente les valeurs les plus élevées en termes de précision et du rappel [4]. Ceci peut dénoter que la nouvelle information qui circule dans un Mi-E peut correspondre à un élément lexical (ou une combinaison de ces éléments) et non à la phrase entière. Conséquement, notre vision du processus de couverture d'information dans les textes journalistiques implique :

- la phrase comme unité de traitement,
- la décomposition de la phrase en groupes syntaxico-sémantiques
- l'identification de la nouvelle information soit par une surface lexicale différente ou une combinaison des éléments dans les groupes syntaxico-sémantiques.

4. Le seuil a été fixé à 0.3, 0.5 et 0.7. L'interprétation de ces valeurs correspond au degré de couverture entre deux phrases.

CHAPITRE 5

DISCUSSION

Dans ce chapitre, nous présentons la discussion des résultats que nous avons obtenus. En même temps, nous établissons des liens avec les travaux de la littérature. Nous couvrons les deux scénarios que nous avons établis dans le cadre de cette thèse.

Nous avons discuté dans la revision de la littérature que l'asymétrie est une caractéristique qui est présente dans une comparaison. Les éléments impliqués dans ce processus prennent nécessairement un rôle spécifique : le *référent* et le *sujet de comparaison* (Tversky, 1977). L'auteur mentionne qu'il existe une direction dans la comparaison ; l'inversion des rôles entraîne un changement de la valeur de similarité que nous octroyons aux objets comparés.

Pour différentes raisons, il arrive des fois que l'on décide de changer la direction de la comparaison. Analysons les deux cas possibles :

a. Direction du *référent* vers le *sujet de comparaison* que nous appelons R–S. Les caractéristiques du *référent* sont couvertes partiellement par celles du *sujet de comparaison*. Ceci entraîne une valeur petite de similarité.

b. Direction du *sujet de comparaison* vers le *référent* que nous appelons S–R. Les caractéristiques du *sujet de comparaison* sont couvertes complètement par celles du *référent*. Ceci entraîne une valeur élevée de similarité.

Quelle est la bonne direction à prendre ? Cela dépend, premièrement, des critères sur lesquels nous nous basons pour identifier les caractéristiques proéminentes des objets à comparer et ainsi identifier celui qui portera le rôle de référent et celui de sujet de comparaison. Deuxièmement, la similarité de ce qu'on veut observer : la similarité du référent dans le sujet de comparaison ou celle du sujet de comparaison dans le référent. Pour le premier scénario, nous utilisons la direction de R–S [1]. Pour le deuxième scénario, nous utilisons la direction S–R [2]. La

1. Le référent est couvert par le sujet de comparaison.
2. Le sujet de comparaison est couvert par le référent.

justification de nos choix a été déjà expliquée dans la méthodologie, mais nous ferons un petit rappel dans les sections où nous discutons chaque scénario.

Comme nous pouvons le constater, l'asymétrie est le pont qui unit les deux scénarios que nous présentons dans cette thèse. Nous avons implémenté deux approches asymétriques pour analyser la couverture d'information, d'abord dans les dissertations d'étudiants et par la suite dans les textes journalistiques.

Notre discussion est menée séparément. Dans la section 5.1, nous présentons la discussion concernant la couverture d'information dans les dissertations d'étudiants et dans la section 5.2, nous présentons la couverture d'information dans les textes journalistiques. Nous exposons chacun des aspects de notre approche en les comparant avec les autres études dans la littérature.

5.1 Scénario 1 : couverture d'information dans les dissertations d'étudiants

5.1.1 La direction de la comparaison

Pour l'analyse de la couverture d'information des dissertations, nous avons utilisé deux mesures symétriques (similarité cosinus et le coefficient de Dice). Pour la même analyse, nous avons proposé un coefficient asymétrique de couverture, le coefficient ACHM (Asymetric Coverage Hybrid Mesure). Pour voir une description de ces mesures, voir la section 3.2.2 (*Mesures de similarité lexicale et de couverture*) du chapitre 3.

En ce qui concerne la direction de la comparaison, les mesures symétriques traitent au même niveau les éléments de la comparaison ; le rôle du référent ou du sujet de comparaison n'est plus pris en considération et donc, la direction non plus. Par contre, une mesure asymétrique considère cet effet, puisque l'inversion des rôles des éléments se voit reflétée dans la valeur de similarité.

Dans le contexte de la couverture d'information dans des dissertations d'étudiants, respecter l'ordre des éléments à comparer est important, car notre intérêt est de connaître l'influence des

références sur la production des dissertations. Alors nous avons établi que les références contenaient les caractéristiques les plus proéminentes et que celles-ci devraient être présentes dans les dissertations. Voyons en termes pratiques, l'influence de la direction de la comparaison.

Kalz *et al.* (2014) ont proposé d'utiliser l'Analyse de Sémantique Latente (ASL) comme méthode d'évaluation des apprentissages déjà acquis. L'objectif de leur étude était d'offrir aux étudiants un curriculum personnalisé en fonction de leurs connaissances et de leurs expériences précédentes. Pour ce faire, les auteurs utilisent l'information contenue dans le portfolio de chaque étudiant et le comparent avec le contenu du cours. Le curriculum proposé a été élaboré en termes de similarité sémantique calculée avec ASL (qui fait des comparaisons symétriques). Kalz *et al.* (2014) affirment que une analyse avec ASL ne peut pas discriminer, avec succès, les documents pertinents des documents non-pertinents, qui feront partie du curriculum. De plus, leur évaluation qualitative a révélé que la discrimination de ces documents chez l'être humain est basée sur une similarité sémantique, mais que ce processus est plus compliqué dans un point de vue cognitif. Dans le même contexte de Kalz *et al.* (2014), nous pourrions dire que le contenu du portfolio d'un étudiant est capital pour la génération d'un curriculum adapté aux besoins de l'étudiant. Ainsi, les unités du cours deviennent les référents et leur contenu devrait être couvert par le portfolio (qui devient le sujet de comparaison). Il existe donc un ordre à respecter dans cette comparaison, mais une approche symétrique est incapable de le faire. Par contre, une approche de similarité asymétrique est capable de caractériser le *référent* et le *sujet de comparaison* et ainsi mieux refléter le processus cognitif de comparaison.

Nous avons conduit notre analyse de couverture d'information sur quatre dissertations d'étudiants avec une approche asymétrique, afin de montrer l'influence des Références Générales (RG) ou des Références Spécialisées (RS) sur la production des dissertations. Les figures de la section 4.1.2, *L'influence des RG et des RS sur la production des dissertations*, montrent le nombre d'alignements des références par dissertation. Pour l'analyse de ces graphes, nous interprétons le nombre d'alignements comme la possible influence des références (RG ou RS) sur une dissertation. Dans la section 4.1.2, nous avons abordé le cas où les références (RG ou RS) détiennent le rôle de référent et les dissertations se comportent comme le sujet de compa-

raison. Puisque le référent comporte des caractéristiques que le sujet de comparaison couvre partiellement, nous nous attendions à des valeurs de similarité inférieures en utilisant la direction R–S [3]. Ces valeurs, même si elles sont inférieures, peuvent mieux décrire l'influence des références (ici ayant le rôle de référent) sur la production des dissertations. Des valeurs supérieures sont obtenues avec la direction S–R [4], car le sujet de comparaison comporte moins de caractéristiques que le référent couvre quasi complètement. Si l'on calcule la similarité avec la direction S–R, toutes les caractéristiques du sujet de comparaison sont donc couvertes par le référent.

Le cas de la direction S–R (le sujet de comparaison est couvert par le référent) n'a pas été présenté dans la section des résultats. Cependant, nous considérons prudent d'aborder le concept de la direction de la comparaison, car elle explique les limitations des approches symétriques à exprimer l'influence des références (RG et RS) sur la production des dissertations. Nous avons inclus dans l'annexe III les graphiques de l'analyse avec l'ACHM où les dissertations détiennent le rôle de référent et les références (RG et RS) tiennent le rôle du sujet de comparaison. En guise d'exemple, nous incluons, ici, l'un de ces graphes à la fig 5.1, qui affiche les alignements des RG pour chaque dissertation.

Si nous utilisons la direction S–R, nous ne pouvons pas déterminer l'influence des références sur la production de dissertations. Comme nous pouvons l'observer dans la fig. 5.1, toutes les RG sont alignées aux paragraphes des dissertations ; il est difficile de voir quelle est la référence qui a eu le plus d'influence sur la production des dissertations.

5.1.2 Les relations lexico-sémantiques pour capturer la couverture des concepts

Les travaux actuels en *LA* qui analysent les textes des étudiants utilisent des approches basées sur les probabilités et sur la fréquence des mots (Kalz *et al.*, 2014; Scheihing *et al.*, 2016). Par exemple, Kalz *et al.* (2014) utilisent un modèle vectoriel et une technique de réduction de dimensions inspirée de l'ASL pour calculer la similarité entre les documents des étudiants et

3. Le référent est couvert par le sujet de comparaison.
4. Le sujet de comparaison est couvert par le référent.

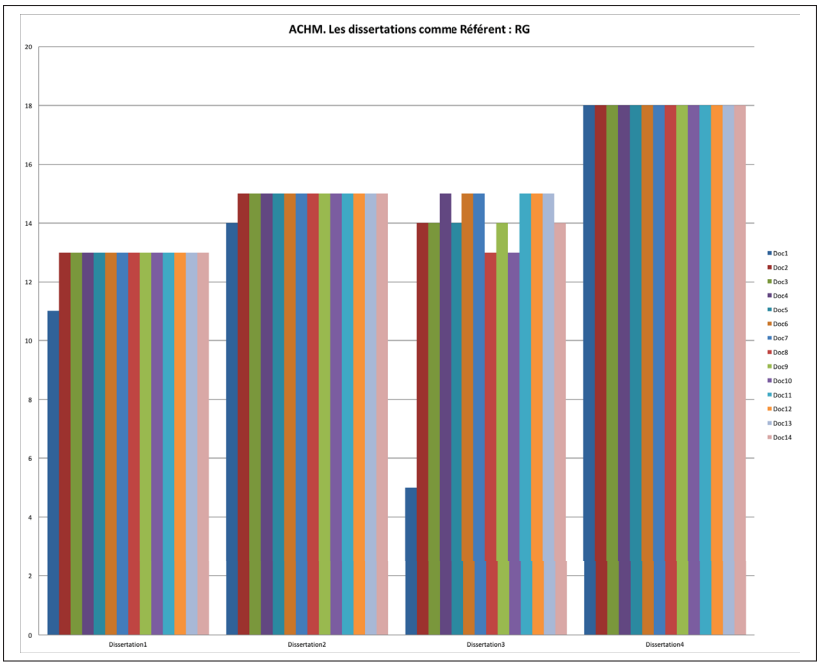


Figure 5.1 Alignement des RG avec les 4 dissertations.
Direction S–R.

le contenu des unités du cours. De leur côté, Scheihing *et al.* (2016) font la classification des messages entre les étudiants et les enseignants avec deux modèles statistiques. Le premier est basé sur l'Allocation de Dirichlet Latente (ADL), et le deuxième est basé sur une technique de Machine à Vecteurs de Support (MVS). Toutes les techniques mentionnées sont construites sur une représentation vectorielle du texte, ce qui implique l'utilisation de la fréquence des mots. Par conséquent, ces approches se contentent d'utiliser uniquement les informations lexicales du texte. Notre position par rapport à de telles approches est la suivante :

a. Les approches basées sur les probabilités et sur la fréquence des mots ne sont pas en mesure de traiter complètement la sémantique de la langue naturelle ; leur analyse est réduite à de cas de co-occurence de mots. Par exemple, *bois-charpentier*, *école-étudiant* ; ces mots partagent une certaine relation sémantique.

b. Une analyse purement lexicale n'est pas adaptée pour le contexte de textes des étudiants, puisque ces derniers ont souvent recours à des surfaces lexicales différentes pour exprimer un même concept.

Dans notre cas, nous utilisons les relations lexico-sémantiques dans le calcul de la similarité de texte en nous servant de l'information sur WordNet. Nous considérons que les relations lexico-sémantiques présentes dans WordNet (la synonymie, l'hyperonymie, l'hyponymie, etc.) doivent être considérées pour évaluer la couverture de la terminologie dans les textes produits par des étudiants (soit des dissertations ou d'autres textes reflétant la compréhension d'un sujet), car elles permettent de rendre compte de l'expression d'un même concept grâce à des formes de surface lexicale distincte. La considération de ces relations lexico-sémantiques a eu un impact double sur notre analyse :

1. Obtenir un nombre supérieur d'alignements entre les dissertations et les références (RG et RS). Conséquemment, les valeurs de la Moyenne Générale d'Alignement (MGA) et la Moyenne Individuelle d'alignement (MIA) dans le cas de notre approche, sont plus élevées que pour les approches purement lexicales. Dans les figures de la section 4.1.2, *L'influence des RG et des RS sur la production des dissertations*, nous observons que les valeurs de la MGA et de la MIA est plus élevée avec notre approche. Par contre, les valeurs de la MGA et de la MIA avec les autres approches sont inférieures. Notre méthode permet de donner aux valeurs de la MGA de la MIA un point d'ancrage pour interpréter comment les étudiants ont utilisé les concepts.
2. Générer des réseaux de mots pour montrer l'interaction des paragraphes des dissertations avec les RG ou les RS. Cette interaction permet de comprendre comment les concepts sont utilisés tout au long d'une dissertation et la source (RG ou RS) d'où le concept est extrait. Les mots qui sont utilisés dans la construction des réseaux de mots sont obtenus lors du calcul de similarité de mot à mot avec WordNet. Ces mots représentent les éléments les plus similaires entre les paragraphes d'une dissertation et les RG ou les RS. Les réseaux de mots ont été présentés dans les figures de la section 4.1.3 *Réseaux de mots des dissertations*.

Finalement, l'utilisation des relations lexico-sémantiques nous semble pertinente dans le contexte de couverture des concepts chez les étudiants, puisqu'une analyse purement lexicale nous paraît limitative, car elle comprend seulement l'appariement des lettres dans les mots. Une analyse

uniquement lexicale se rapproche de celles qui pourraient être conduites pour la détection de plagiat. Dans ce cas, une approche lexicale est bien justifiée, car l'intérêt du plagiat est de trouver des copies exactes du textes. Dans le cas de la couverture de concepts, une approche lexicale se justifie mal puisque les étudiants ont nécessairement recours à des formes de surfaces lexicales différentes qui sont de la synonymie ou des variantes différentes pour exprimer un même concept.

5.1.3 Évaluation

Nous avons conduit un *Processus de Vérification par les Membres* (PVM) avec les enseignants du cours à deux reprises. Dans la première session du PVM, nous avions réalisé l'analyse de la couverture d'information avec les RG. Pour la deuxième session du PVM, nous avons ajouté les RS dans notre analyse. Réaliser un PVM est une pratique assez commune en projets d'ingénierie (Dubé & Paré, 2003). Ceci nous a permis d'évaluer notre approche directement avec les enseignants impliqués dans l'obtention et l'interprétation des données. Le PVM implique une négociation des découvertes entre le chercheur et les autres membres du projet. Avec le PVM, nous avons obtenu des informations complémentaires par rapport à la thématique des dissertations et l'évaluation de nos résultats.

– En ce qui concerne la thématique des dissertations :

- Rappelons que les dissertations 1 et 2ont été produites par l'étudiant A ; ces deux dissertations portaient sur la *motivation*,
- La dissertation 3 qui appartient à l'étudiant B portait sur l'*acquisition des connaissances*
- La dissertation 4 de l'étudiant C abordait le sujet de la *curiosité*.

Ceci est un premier aspect, que nous considérons comme intéressant du PVM, puisque l'intéraction avec les membres révèle plus d'information sur l'origine des données.

– En ce qui concerne l'interprétation de nos résultats :

En observant les résultats de notre analyse, nous avons conclu que l'étudiant A avait présenté des difficultés à exprimer les concepts lors de la rédaction de sa dissertation. Cette observation a été confirmée par les participants lors du PVM. En effet, ils ont confirmé que l'étudiant A, l'auteur des dissertations 1 et 2, a eu du mal à saisir les concepts, ceci se reflète par le fait que les dites dissertations ont moins d'alignements que les dissertations 3 et 4. En ce qui concerne l'étudiant B, il avait une bonne compréhension des concepts, mais il n'avait pas très bien réussi à établir une connexion entre les concepts ou à renforcer la cohérence. L'étudiant C, l'auteur de la dissertation 4, avait une meilleure compréhension des concepts et était cohérent. Nous avons aussi appris que la thématique de cet étudiant n'a pas été couverte par les RG. Elle était seulement couverte par les RS que l'étudiant C avait proposées par lui même. L'absence de cette thématique dans les RG pourrait expliquer le nombre inférieur d'alignements dans les approches purement lexicales (similarité cosinus et coefficient de Dice). Puisque notre approche utilise les relations lexico-sémantiques disponibles en WordNet, elle a été en mesure de faire plus d'alignements avec la dissertation 4.

Nous avons construit un réseau de mots à partir de la terminologie qui contribue le plus à la couverture de concepts entre une dissertation et les RG ou les RS. Le but de cette visualisation était de montrer l'interaction entre la dissertation, la terminologie, et les RG ou les RS. Dans les deux sessions du PVM, notre analyse a révélé, pour l'étudiant A, que la terminologie reliée à la motivation était plus proéminente dans la dissertation 2 que dans la dissertation 1. Les participants du PVM étaient en mesure de tirer des conclusions en observant les réseaux de mots et en utilisant leur propre expérience avec les étudiants.

Dans notre rencontre, nous avons aussi appris que les réseaux de mots, présentés sous la forme d'une image statique, posaient des problèmes d'interprétation. L'une des remarques faites par les participants du PVM était d'implémenter une animation pour présenter les réseaux de mots. Car dû au chevauchement des arêtes, il était difficile de comprendre les interactions entre les paragraphes, les concepts et les références. Nous savons que les représentations à base de

graphes sont conseillées pour des graphes ayant moins de 20 nœuds (Ghoniem *et al.*, 2004), mais nous avons estimé que l'interaction disponible sur *Ghephi* serait suffisante. À l'avenir, nous considérerons une animation pour présenter d'une façon plus claire le réseau de mots. Ceci est un autre aspect des PVM que nous trouvons intéressants ; leurs suggestions sont aussi une forme d'évaluation par rapport à la forme que nous utilisons pour présenter les résultats.

5.1.4 Cohésion

Lors de nos sessions de PVM, nous avons corroboré que pour la dissertation 1, notre approche a trouvé une couverture faible de la terminologie des RG et des RS par rapport aux autres dissertations. En observant les graphes des figures de la section 4.1.1 *Nombre de documents couverts par chaque dissertation*, l'un des participants a déclaré : *« Ce que je vois est que dans les dissertations 2, 3 et 4, il y a plus de consistances dans l'utilisation des RG et des RS, car on y voit plus de documents alignés. Il semble avoir un flux et c'est ce que vous obtenez quand vous lisez un texte. La dissertation 2 a plus de ce patron, de ce flux ; elle est plus condensée, plus cohérente que la dissertation 1 »*.

Les commentaires de cohésion de la part des participants aux PVM, nous permettent de faire un lien avec la théorie d'Halliday & Hassan (1976) à propos de la cohésion du texte ; nous avons présenté cette théorie dans la *Revue de la littérature*, section 1.4.4. Cette théorie est, d'une certaine façon, liée à notre analyse, voyons pourquoi.

– La théorie de cohésion de texte

Le concept de cohésion du texte est purement sémantique ; il fait référence aux relations de signification qui existent dans le texte. Plus formellement, la cohésion du texte est définie comme l'ensemble des relations lexicales, grammaticales et sémantiques reliant l'ensemble des unités textuelles. Halliday & Hassan (1976, p. 1) définissent en termes linguistiques le concept de texte : *« any passage, spoken or written, of whatever length, that does form a unified whole »*. En tant que locuteurs d'une langue, nous avons l'habilité de déterminer, si une collection de

phrases représente un texte ou non. Ceci inclut les cas où l'on serait incertain de la distinction entre un texte et un ensemble de phrases non liées. Alors, cette distinction est en dernier recours une question de degré. Assigner un degré de *cohésion* est probablement un exercice familier aux enseignants quand ils lisent les compositions de leurs étudiants (Halliday & Hassan, 1976).

Halliday & Hassan (1976) mentionnent différentes stratégies pour achever la cohésion du texte. L'une de ces stratégies correspond à la cohésion lexicale. Celle-ci, selon Halliday & Hassan (1976), est accomplie par la sélection du vocabulaire approprié. Pour les auteurs la sélection du vocabulaire correspond à une paire d'éléments lexicaux qui correspondent au même concept. Une telle paire reçoit le nom d'attache ou *tie* en anglais. Par exemple :

(1) There was a large *mushroom* growing near her, about the same height as herself ; and, when she had looked and see what was on the top of it. She stretched herself up on tiptoe, and peeped over the edge of the *mushroom*,...

(2) Accordingly ... I took leave, and turned to the *ascent* of the peak. The *climb* is perfectly easy...

(3) Henry's bought himself a new *Jaguar*. He practically lives in the *car*. [5]

Comme nous pouvons le voir, dans l'exemple, (1) l'attache se fait par une répétition du mot *mushroom*. Une simple approche lexicale est capable de détecter ce type d'attache. Dans l'exemple (2), l'attache se fait par l'utilisation d'un synonyme d'*ascent* c'est-à-dire *climb*. Dans le cas de l'exemple (3), l'attache se fait en utilisant un mot d'ordre supérieur ce qui est le cas de *car* pour *Jaguar*. Halliday & Hassan (1976) parlent de la cohésion d'un même texte. Dans notre cas, la cohésion se réalise entre les dissertations et les références (les RG ou les RS). Dans ce contexte, si on étend le concept d'attache (entre une dissertation et les références)

5. Ces exemples ont été tirés de Halliday & Hassan (1976, p. 278).

notre approche est capable de couvrir les attaches de type 1, 2, et 3. La capture des attaches de type 2 et 3 (exemples (2) et (3)) est impossible pour une approche lexicale.

Puisque l'évaluation de la cohésion du texte est une tâche à laquelle les enseignants sont très habitués, un système de détection de la cohésion du texte semble pertinent. Par exemple, Dascalu *et al.* (2015) analysent les discussions des étudiants selon deux perspectives : le dialogisme et la cohésion du texte, en utilisant la théorie d'Halliday & Hassan (1976) que nous venons d'expliquer dans les paragraphes précédents. L'objectif de Dascalu *et al.* (2015) était d'analyser la cohésion des conversations des étudiants. Pour ce faire, les conversations doivent passer par un processus de transcription automatique. À partir du texte obtenu de la transcrition, les auteurs construisent des *chaînes sémantiques* qui sont obtenues par une fonction de cohésion. Cette fonction est en réalité un *pipeline* d'outils en NLP (LSA, LDA et des mesures de distance sémantique en WordNet), ce qui leur permet d'obtenir une chaîne de concepts sémantiquement liés. La fréquence de ces chaînes sémantiques dans tout le discours d'un étudiant dénoterait la cohésion de son discours. Dans notre cas, nous avons utilisé les dissertations finales des étudiants afin de vérifier la couverture des concepts des RG ou des RS dans les dissertations ; le concept de *cohésion* est apparu lors de nos deux PVM.

En vérifiant la couverture des concepts des RG et des RS dans les dissertations des étudiants, nous avons implémenté une stratégie hybride. Celle-ci utilise de l'information lexicale du texte et l'information linguistique provenant de WordNet. Cette information linguistique provenant de WordNet comprend les relations lexico-sémantiques que sont référées par Halliday & Hassan (1976) comme les directives pour identifier les *attaches* (du type 2 et 3). Aussi, dans notre analyse, nous assurons la cohésion entre les dissertations et les références (RG et RS) par la couverture de concepts.

5.1.5 La différence entre prédire et expliquer une note

Lors des sessions des PVMs, nous avons corroboré que notre approche était en mesure de présenter la dissertation 1, qui avait eu une note d'échec, comme ayant une moindre couverture

des concepts des RG et des RS. Le degré de couverture des concepts pouvait expliquer la note attribuée aux dissertations. Généralement, quand les études en *LA* sont menées pour analyser les notes l'objectif est de prédire les notes des étudiants (Figueira, 2016; Al-Barrak & Al-Razgan, 2016; Mueen *et al.*, 2016). Quel est donc le chemin à prendre lorsque nous analysons des textes d'étudiants, prédire une note ou plutôt l'expliquer ? Y a-t-il une différence entre ces deux concepts ?

Il existe un très vaste débat sur la différence entre prédire et expliquer, commençant en philosophie (Rescher, 1958; Forster, 2002; Hitchcock & Sober, 2004) et plus récemment en recherche scientifique (René, 1993; Shmueli, 2010). Rescher (1958) mentionne qu'il existe une *asymétrie* inhérente dans le temps entre une prédiction et une explication : la première concerne le futur et la seconde concerne le passé. Cette asymétrie temporelle est d'une portée importante et fondamentale, puisque le passé porte une supériorité évidente sur le futur, en ce qui concerne l'accès à l'information fiable. Cet accès entraînerait la suppression d'incertitude et de la contingence. En d'autres mots, on connaît bien le passé et dans le moment présent nous pouvons évaluer les retombées de toute action entreprise dans passé. Le futur n'est pas encore connu, il n'est donc pas possible de prévoir avec certitude un événement.

Rescher (1958) indique aussi que dans une explication, l'événement est déjà connu, mais les conditions qui l'ont produit doivent être éclairées. Le cas contraire concerne les prédictions, où l'événement n'est pas connu, mais où l'on est confronté à des conditions qui pourraient provoquer quelque chose ; un possible événement doit donc être proposé.

Voyons les implications de prédire et expliquer en termes plus technologiques. Selon Shmueli (2010), la plupart des modèles que nous trouvons généralement en TI (surtout en forage de données) seraient classifiés dans deux catégories : des « *modèles prédictifs* » et des « *modèles explicatifs* ». Selon Shmueli (2010), un modèle explicatif permet de tester des hypothèses causales sur des constructions théoriques. Par contre, un modèle prédictif est toute méthode qui produit des prédictions, indépendamment de son approche sous-jacente (un exemple serait une approche basée sur le théorème de Bayes). Ce type d'approches prédictives prolifèrent en LA.

De nombreuses études ont analysé les documents réalisés par des étudiants. La plupart d'entre elles se concentrent sur différents types de productions élaborées pendant les cours (dialogues et discours sur les plates-formes en apprentissage collaboratif supporté par ordinateur (ACSO)), mais pas sur les dissertations (Dascalu *et al.*, 2015; Scheihing *et al.*, 2016). Par exemple, Sorour *et al.* (2014) ont présenté une méthode basée sur des réseaux neuronaux et l'ASL pour prédire les notes des étudiants. Les auteurs ont utilisé uniquement des commentaires de *« type libre »* écrits par les étudiants après chaque leçon pour faire les prédictions. Si le but est de prédire la note obtenu par un étudiant, il est plus approprié d'utiliser les dissertations car elles représentent mieux la compréhension de l'étudiant à la fin du cours. Les discours de type libre pourront contenir des sujets très divers qui peuvent ne pas être liés à la thématique du cours. En revanche, nous avons utilisé les dissertations des étudiants qui ont été livrées à la fin du cours pour leur notation. Nous les avons comparées avec les références générales (RG) et les références spécialisées (RS) sans l'intention de prédire les notes octroyées aux documents par les enseignants. Nous avons utilisé le concept de couverture d'information pour expliquer la note des dissertations. Conséquemment, une dissertation avec une bonne couverture des concepts des RG ou des RS aurait nécessairement une note de réussite. Alors, la couverture d'information est l'*hypothèse causale* à laquelle Shmueli (2010) fait référence et que notre approche aurait testée pour expliquer une note. Nous sommes conscients qu'une note ne dépend pas exclusivement de la couverture d'information ; il existe d'autres facteurs qui peuvent influencer celle-ci (style d'écriture, argumentation, respect de la date de remise, etc).

En tenant compte de notre analyse, nous avons basé notre alignement sur la couverture d'information. En le faisant, nous avons établi le nombre d'alignements d'une dissertation avec les documents des références comme le paramètre expliquant les notes. Nous expliquons ici notre démarche interprétative :

- Le nombre d'alignements avec les documents des références : Plus le nombre d'alignements est élevé plus la note est élevée.
- La MGA constitue un repère de couverture d'un groupe d'étudiants. Si les enseignants gardent un historique des MGA des groupes précédents, l'enseignant pourrait s'attendre

à ce que les meilleurs élèves produisent des dissertations avec une MIA supérieure à la moyenne historique des MGAs.

- La MIA constitue le repère de couverture individuelle d'un étudiant. Elle est dépendante du nombre d'alignements de la dissertation analysée. Elle pourrait être utilisée pour déduire l'octroi d'une note à une dissertation.

5.2 Scénario 2 : couverture d'information de textes journalistiques

5.2.1 Remarques sur la direction de la comparaison

Dans le cas de la couverture d'information de textes journalistiques, nous avons abordé la direction de la comparaison S–R [6], où les caractéristiques du sujet doivent être couvertes le plus possible par le référent. Notre approche utilise des patrons linéaires qui correspondent aux caractéristiques du sujet de comparaison et du référent.

En ce qui concerne la couverture d'information des textes journalistiques, nous avons établi, dans notre méthodologie, que le référent est de taille variable. Au début, le référent contient l'information de la première nouvelle d'une thématique. Les autres nouvelles de la même thématique, qui ont été générées par la suite, prennent à leur tour, le rôle de sujet de comparaison. La nouvelle information qui est ajoutée au référent provient de tout sujet de comparaison qui n'arrive pas à être couvert par le référent. De cette façon, nous considérons l'évolution chronologique de l'événement.

5.2.2 Problèmes avec TREC

Pour conduire l'analyse de couverture d'information de textes journalistiques, nous avons réalisé une nouvelle annotation du corpus *novelty TREC*. Dans la section 3.3.2, nous avons expliqué notre façon de conduire la nouvelle annotation. Nous en présentons ici un bref rappel :

6. Le sujet de comparaison est couvert par le référent.

- Le processus d'annotation du corpus a été remis en question par quelques participants depuis la première année de la compétition (Schiffman, 2002; Tsai & Chen, 2002; Collins-Thompson *et al.*, 2002).
- Les annotateurs devaient identifier les phrases pertinentes, puis les phrases contenant de la nouvelle information. Le critère de la pertinence dépend de chaque annotateur ; dans les comptes-rendus de la conférence(Harman, 2002; Soboroff & Harman, 2003; Soboroff, 2004) n'offrent pas une définition claire de la pertinence, pourtant la nouveauté dépend d'elle.
- La charge cognitive des annotateurs lors de l'annotation était relativement grande si l'on considère la capacité de mémoire chez l'être humain (Baddeley & Hitch, 1974; Miller, 1956; Shiffrin & Nosofsky, 1994; Ma *et al.*, 2014). Pour chaque thématique les annotateurs avaient au minimum 25 nouvelles à traiter ; dans le cas le plus extrême, ils avaient 69 nouvelles. Cette surcharge cognitive semble avoir joué un rôle important lors de l'identification de l'information nouvelle.
- Dans la tâche TREC, les annotateurs devaient considérer la phrase dans son entier pour identifier la nouvelle information sans se soucier des éléments qui la rendent porteuse de la nouveauté.
- Finalement, les compétitions *TREC novelty Track* en 2002 (Harman, 2002), 2003 (Soboroff & Harman, 2003) et en 2004 (Soboroff, 2004) attestent que des problèmes dans la définition de la tâche ont eu un impact négatif sur les résultats.

Également, dans la section 3.3.1, nous avons exposé les raisons qui nous ont menées à refaire l'annotation. Nous en présentons aussi un bref rappel :

Pour construire notre *ground truth*, issu de notre processus d'annotation, nous devions mesurer l'accord entre les annotateurs. Habituellement pour ce type de tâches, le coefficient κ est utilisé, mais nous avons rencontré des problèmes dans son application. D'abord, quatre experts ont annoté les textes. Le coefficient κ étant conçu pour exprimer l'accord entre deux annotateurs, nous avons d'abord utilisé une version adaptée pour exprimer l'accord entre plusieurs

annotateurs, le coefficient κ Fleiss. Ce dernier présentait les mêmes problèmes que sa version à deux annotateurs. Nous les exposons par la suite.

– Le coefficient *kappa* : ses paradoxes et notre stratégie de solution

Le coefficient κ présente deux paradoxes qui apparaissent quand il y a un accord élevé sur une classe (Cicchetti & Feinstein, 1990; Feinstein & Cicchetti, 1990). Nous avons exposé les dits paradoxes à la section 3.3.6. Des travaux récents ont proposé de nouveaux coefficients pour mesurer l'accord entre les annotateurs. Par exemple, Power (2012) conseille l'utilisation de la corrélation de Mathiew. Auparavant, le même auteur conseillait d'utiliser le *Informedness coefficient* (Power, 2003). Nous avons opté pour la proposition de Cicchetti & Feinstein (1990), parce qu'elle permet d'exprimer l'accord des classes positives et des classes négatives. De cette façon, nous avons pu évaluer l'accord entre les annotateurs sur ce qui est nouveau et ne l'est pas.

Lors de l'application du coefficient κ sur les données de l'annotation des thématiques 1 et 2, nous avons relevé ces paradoxes. Reprenons le cas de la thématique 1 pour illustrer l'un des paradoxes du coefficient κ. Quand l'accord est mesuré entre l'étiqueteur 1 et 3, la valeur du coefficient κ est de 0.119, ce qui exprime un accord très bas. Si nous observons la valeur de P_{pos} (0.92), celle-ci dénote un accord élevé entre les annotateurs. La valeur de P_{neg} est très basse (0.181). Il existe donc une disparité d'accord entre la classe négative et la classe positive entre les deux annotateurs, disparité que le coefficient κ n'est pas en mesure d'exprimer. De plus, cette même thématique présente des valeurs négatives du coefficient κ; ceci exprime un accord ou désaccord pire de ce à quoi l'on peut s'attendre.

Le fait d'utiliser les coefficients P_{pos} et P_{neg} de Cicchetti & Feinstein (1990) pour exprimer l'accord entre les annotateurs a été bénéfique pour deux raisons :

- D'abord, éviter des erreurs interprétatives de l'accord entre les annotateurs si le coefficient κ rencontre ses paradoxes.

- Ensuite, mesurer l'accord des classes positives et des classes négatives séparément permet de faire une parallèle entre une phrase qui porte la nouveauté et une phrase qui ne la porte pas (classe positive et classe négative respectivement).

– L'unité de traitement : phrase contre les documents segmentés

Selon Schiffman (2002); Soboroff & Harman (2005), la phrase ne compte pas assez d'information pour prendre une décision et déterminer ce qui est nouveau, car celle-ci n'est pas une bonne unité pour l'analyse. Par exemple, Soboroff & Harman (2005) préfèrent la décomposition d'un document en morceaux (paragraphes) pour découvrir les relations entre des entités dans une collection de documents d'une même thématique. Habituellement, nous associons la notion de paragraphe aux étendues d'une idée. Si notre intérêt est de traiter des textes journalistiques, et sachant que ces derniers rapportent des événements, la phrase semble plus adaptée que le paragraphe, puisque la structure de la phrase comprend aussi la structure d'un événement.

À différence de ces approches qui prônent le traitement d'un document entier ou en paragraphes pour détecter la nouvelle information, nous retrouvons dans la phrase, à condition qu'elle soit décomposée, les informations nécessaires pour identifier la nouvelle information. Conséquemment, cette décomposition nous permet de :

- Considérer que la nouvelle information circule dans la phrase et peut prendre la forme d'un argument du prédicat ou encore d'un adjoint.
- Trouver la nouvelle information quand les éléments lexicaux sont complètement différents de ceux qui ont été déjà rencontrés précédemment.
- Considérer jusqu'à un certain point la structure de la phrase. Ceci est envisageable grâce à l'utilisation des patrons linguistiques.

Dans notre évaluation quantitative de la thématique 3, nous avons trouvé que l'un des patrons les plus performants était PN–PN. Ce patron vise à attraper grossièrement les locatifs spatio-temporels. La thématique 3 aborde la mort de la princesse Diana dans un accident de voiture

à Paris. Dans le contenu de cette thématique, les médias abordent la thématique et les déclarations des personnages en des lieux distincts, à des moments différents. Par exemple, dans les premières nouvelles, la mort de la princesse est rapportée seulement dans un accident de voiture. Plus tard, on apprend que cet accident a eu lieu *« sur le pont de l'Alma » « à Paris »*, etc. Ainsi, à mesure qu'une thématique évolue, dans le temps, l'information devient plus accessible aux médias, et conséquemment, la couverture d'information devient plus précise. Indépendamment du patron le plus performant dans nos expériences, le fait d'utiliser des patrons pour capturer sommairement des « *petits morceaux syntaxiques* », correspond à la façon dont la nouvelle information apparaît dans les nouvelles.

– Le traitement des *mots fréquents*

Un autre aspect qui nous différencie des travaux présentés à la compétition *novelty TREC* est le traitement des mots fréquents. Plusieurs travaux ont utilisé des stratégies d'élimination de *mots fréquents* et de *mots vides* (Collins-Thompson *et al.*, 2002; Dkaki *et al.*, 2002; Zhang *et al.*, 2002). D'autres ont implémenté seulement une stratégie de réduction de mots fréquents (Collins-Thompson *et al.*, 2002; Tsai & Chen, 2002) ; certaines basées sur la mesure TF-IDF (Term frequency Inverse document Frequency). Le contenu des listes de mots dits vides ou de mots fréquents peut inclure les prépositions ou les verbes auxiliaires. Pourtant, les mots dits « *fréquents* » ont une fonction primordiale pour identifier les relations sémantiques. En effet, ces mots peuvent aussi s'avérer nécessaires pour différencier deux événements exprimés par deux phrases qui ont une même surface lexicale, mais un ordre différent. Par exemple :

(4) *Jean* demande à *Marie* de faire la vaisselle.

(5) *Marie* demande à *Jean* de faire la vaisselle.

Comme nous pouvons le voir, dans (4) et (5), le contenu lexical est le même, mais il y a un changement dans l'ordre des mots. Dans (4), *Jean* est l'agent de la phrase et *Marie* est le patient. Dans (5), les rôles sont inversés.

Pourquoi est-il important de prend soin des mots vides ou des mots fréquents quand on analyse les textes de nouvelles ? Il y a certains éléments lexicaux communément inclus dans les listes de *mots vides* qui jouent un rôle important pour identifier, par exemple, la factualité d'un événement. Saurí & Pustejovsky (2012) mentionnent que la factualité d'un événement implique deux niveaux : la polarité et la certitude de la phrase. La première distingue les instances positives ou négatives des événements (s'ils ont eu lieu ou pas). Dans ce sens, si un événement n'a pas eu lieu, il n'y aurait pas de nouvelle information à identifier. La certitude joue aussi un rôle important pour déterminer la factualité des événements. Cette information est exprimée en anglais par des verbes modaux et en français par les modes grammaticaux comme le conditionnel. Si on inclut dans la listes de mots fréquents les verbes modaux, nous perdons la factualité de l'événement.

Dans notre cas, nous avons aussi entrepris une stratégie de réduction de mots fréquents. Cependant, nous avons retenu les verbes modaux pour conserver la factualité de l'événement, et les prépositions pour conserver les relations syntaxico-sémantiques. Avec ces prépositions, nous avons construit le patron PN–PN et le patron VPN–VN qui, dans ce dernier cas, capture grossièrement la transition d'une phrase exprimée à la voix passive vers la voix active. La voix passive est une caractéristique très utilisée par les journalistes dans la production de textes de nouvelles (Richardson, 2006, p. 54-55). Dans nos résultats, le patron VPN–VN est l'un des patrons obtenant les meilleures valeurs en termes de mesure-F pour les thématiques 2, 4 et 5.

– La détection de la nouveauté

Iacobelli *et al.* (2010a,b) font la *Détection de Nouveauté* (DN) grâce à un système appelé *Tell me more*. Leur système permet d'extraire les histoires similaires sur un fil de nouvelles en sélectionnant des fragments textuels des histoires qui offrent de la nouvelle information. Ces

fragments de texte constituent des paragraphes. Le système présente de la nouvelle information non répétée. *Tell me more* est basé sur la prémisse que les entités nommées, les nouveaux quantificateurs (chiffres) et les nouvelles citations [7] sont une forme importante de nouvelle information. La détection d'entités nommées est cruciale pour une bonne performance. Les outils de détection d'entités nommés ne sont pas disponibles pour toutes les langues ; de plus leurs informations doivent être entretenues pour être à jour. Par exemple, un outil développé en 2014 ne serait pas en mesure d'identifier *Donald Trump* comme le président des États-Unis en 2016. Dans notre cas, nous sommes en mesure de capturer les entités nommées dans la partie nominale de nos patrons. Par exemple, le patron N–V vise à capturer grossièrement le sujet du prédicat ; la partie nominale de ce patron (N) correspond bien à une entité nommée. Nous avons utilisé un étiqueteur de classes lexicales pour construire nos patrons. Ces outils sont relativement indépendants de la langue et ils sont bien connus pour avoir une bonne performance. Nos patrons sont facilement adaptables à d'autres langues comme le français ou l'espagnol, mais devraient être repensés pour des familles de langues morphologiquement éloignées, comme le chinois ou le japonais.

De plus, la thématique 5 aborde les *tremblements de terre en Turquie* sur une période de deux ans. Dans l'ensemble de nouvelles, des tremblements de terre avec différents degrés d'intensité et différents lieux ont été rapportés. Alors, cette thématique est le cas idéal à traiter par le système de Iacobelli *et al.* (2010a,b), mais lors de l'annotation nous avons remarqué que la thématique 5 comporte les niveaux les plus bas d'accord des classes positives et des classes négatives entre les annotateurs. Ceci nous laisse croire qu'implémenter une stratégie de DN basée sur les chiffres et les entités nommées comme le suggère le système *tell me more* peut être étrangère à la façon dont un lecteur détecte par lui-même la nouveauté. D'autres études axées sur le sujet devront être conduites pour corroborer cette hypothèse.

Aksoy *et al.* (2012) abordent la DN pour une thématique en particulier selon trois approches : la première est basée sur le concept de similarité cosinus, la deuxième sur un modèle statis-

7. Il s'agit d'une déclaration faite par une personnage : *"I will build that wall"*, Donald Trump declared yesterday.

tique et la troisième sur un coefficient de couverture. Aksoy *et al.* (2012) utilisent le concept d'asymétrie pour la DN dans deux de leurs approches, le modèle statistique et l'utilisation d'un coefficient de couverture. Le coefficient de couverture d'Aksoy *et al.* (2012) utilise une représentation matricielle pour un ensemble de documents. Cette matrice est carrée et elle contient des probabilités de couverture entre les documents. Dans cette matrice, la couverture d'un document doc_i par un document doc_j est la probabilité de sélection de tout terme du doc_i dans le doc_j. La matrice est dont asymétrique car elle contient les valeurs considérées par la couverture du doc_i dans le doc_j et la couverture du doc_j dans le doc_i. Les meilleurs résultats, en termes de mesure-F, sont obtenus avec le modèle statistique qui est basé sur *Dirichlet smoothing*, qui est symétrique. Des résultats inférieurs sont aussi obtenus avec l'utilisation du coefficient de couverture asymétrique. Dans notre cas, nous avons réalisé des expériences avec deux mesures symétriques et un coefficient de couverture asymétrique à base de patrons linéaires linguistiques. Nos résultats montrent que notre coefficient de couverture asymétrique est capable d'avoir les mêmes performances qu'une mesure symétrique avec la différence qu'il peut expliquer l'origine de la nouveauté. Nous reviendrons sur ce sujet un peu plus tard.

Karkali *et al.* (2013) font la DN par un algorithme basé sur la fréquence inverse du document (IDF). Cet algorithme n'utilise ni le concept de similarité ni une mesure de distance. Ainsi, pour détecter si un nouveau document contient de la nouvelle information, les auteurs proposent de capturer la différence de la représentation IDF d'un document actuel avec les représentations IDF des documents passés. Comme nous l'avons déjà mentionné, nous croyons que la nouveauté ne réside pas seulement dans l'apparition de nouveaux éléments lexicaux dans un fils de nouvelles. La nouveauté concerne aussi une combinaison différente de ces éléments dans les phrases qui appartiennent à d'autres nouvelles plus récentes. Par exemple, un locatif, c'est-à-dire le lieu d'un événement, peut être partagé par plusieurs Mi-E qui se déroulent à un moment distinct, mais qui appartient à un même Ma-E.

5.2.3 La couverture d'information : un type de biais

En journalisme, nous avons trois types de biais : le contrôle d'information, la déclaration et la couverture (pour plus de détails,voir la section 1.5.2). Des travaux pour mesurer la couverture d'information comme biais sur les textes de nouvelles ont été déjà proposés (Saez-Trumper *et al.*, 2013; Park *et al.*, 2009). Par exemple, Saez-Trumper *et al.* (2013) ont mesuré la couverture d'information selon trois critères : la longueur (en mots) des articles couvrant une histoire particulière, la répartition du nombre de mentions pour un personnage ciblé dans différents médias (journaux et réseaux sociaux), et les régions géographiques. La comparaison entre la couverture des médias traditionnels et la couverture des réseaux sociaux d'une histoire rend cette étude intéressante, mais leur méthode ne prend en considération ni les composants de l'événement, ni la structure des nouvelles. Considérant le premier élément, notre coefficient de couverture utilise des patrons linéaires qui visent à attraper sommairement les relations grammaticales de sujet, objet, et adjoints, ce qui permet, d'une certaine façon, de capturer les composants de l'événement. En ce qui concerne nos résultats, nous observons que les patrons NV–NV et VPN–NV reviennent toujours avec les meilleures valeurs de la mesure-F. En considérant le deuxième élément, la structure des nouvelles, notre proposition pourrait facilement être adaptée pour prendre en compte la structure d'une nouvelle en attribuant un poids plus élevé aux éléments qui se retrouvent au début de la nouvelle. Ces éléments sont d'ailleurs sensés contenir les informations les plus importantes d'une nouvelle (Richardson, 2006).

5.2.4 La structure des nouvelles

Park *et al.* (2009, 2010) fondent leur approche sur la structure pyramidale des nouvelles. La structure pyramidale que les auteurs utilisent diffère un peu de celle mentionnée par Richardson (2006). Pour Park *et al.* (2009, 2010) cette structure contient les éléments suivants : La « *tête* » et la « *sous-tête* » qui contiennent les éléments clés qui reflètent la thématique qui est discutée dans la nouvelle. Le « *chapeau* » [8] représente la première ou deuxième phrase d'un

8. En anglais *lead*. Le terme journalistique préféré en français est *chapeau* d'après le dictionnaire de terminologie de l'Office québécois de la langue française.

article de nouvelle ; il guide les lecteurs aux faits les plus intéressants de la nouvelle. Le dernier élément de la structure pyramidale ,« *texte principal* », est le reste du texte de la nouvelle. Park *et al.* (2009, 2010) utilisent des mots-clés pondérés pour mesurer la différence de couverture entre différentes sources. Dans leur proposition, le poids associé à un mot-clé spécifique diminue en fonction de sa position dans le texte : la *tête* (poids supérieur), la *sous-tête*, le *chapeau*, le *texte principal* (poids inférieur). Pour capturer l'agent d'un événement, les auteurs choisissent les noms propres et les pronoms sujets. Pour capturer l'agent d'un événement, nous avons plutôt choisi d'utiliser des patrons linguistiques linéaires. Le patron N–V vise à capturer grossièrement le sujet de la phrase, qui correspond à l'entité qui déclenche un événement, donc l'agent de l'événement. Notre approche est plus large, puisque l'ensemble de patrons permet aussi de capturer d'autres composants de l'événement qui correspond à des notions spatio-temporelles. Le paramètre α est utilisé pour pondérer les patrons. Cette stratégie de pondération peut être utilisée comme un filtre permettant de donner plus de poids aux éléments de l'événement auxquels nous sommes intéressés. De plus, nous pouvons emprunter la proposition de Park *et al.* (2009, 2010) en associant la pondération du paramètre α à la structure de la nouvelle.

5.2.5 Intérêt des patrons pour expliquer l'origine de la nouveauté

En ce qui concerne la performance de notre approche asymétrique par rapport aux autres approches symétriques que nous avons testées, nous pouvons observer, dans nos résultats, que pour les thématiques 2, 3 et 4, l'approche asymétrique est aussi précise que les deux autres approches. L'approche asymétrique n'arrive pas à performer également avec les deux autres thématiques. Pour la thématique 1 et la thématique 5, les valeurs de la mesure-F pour les approches symétriques dépassent les valeurs obtenues avec notre approche asymétrique par 0.008 et 0.05 respectivement. La différence de performance est minime. Alors, quelle est l'approche à choisir ? L'un qui est symétrique ou l'autre qui est asymétrique.

Le fait d'utiliser ces patrons rend notre approche plus avantageuse ; elle est capable d'exprimer l'origine de la nouveauté quand un patron en particulier est pondéré avec le paramètre α. Par

exemple, pour la thématique 2, la nouveauté vient principalement des patrons VPN–NV ou NV–NV. Ces deux patrons capturent sommairement la relation de l'agent de la phrase.

5.2.6 Les observations des annotateurs

Jusqu'à présent, nous avons mis l'accent sur les résultats de l'évaluation de notre approche asymétrique, mais ne peut être négligé le processus d'annotation du corpus. Il y a deux éléments qui ressortent par rapport au processus d'annotation :

- *La charge cognitive :* à ce propos, nos résultats suggèrent que l'accord entre les annotateurs pour les classes positives et négatives est plus élevé, particulièrement quand la quantité d'information consiste en peu de phrases (25), comme c'est le cas pour les thématiques 2 et 3. Nous aurions cru que le fait de raconter l'histoire d'un personnage particulier, comme c'est le cas dans les thématiques 3 et 4, pourrait relever l'accord entre les annotateurs sur ce qui est nouveau ou pas. Ceci est vrai pour la thématique 3 qui parle de l'accident de la Princesse Diana, mais les résultats de la thématique 4 ne permettent pas de corroborer cette hypothèse. La thématique 4 contenait 119 phrases, mais elle abordait l'histoire de Charles Schulz. Nous croyons que, dans ce cas, la quantité d'information a joué un rôle important qui explique la différence entre les résultats des thématiques 3 et 4.
- *Le contexte culturel des annotateurs :* à ce propos, nous avons observé que la thématique *2* présente des valeurs très élevées sur ce qui est nouveau. La thématique *2* aborde l'embargo de Cuba et les quatre annotateurs proviennent de l'Amérique latine. Le sujet de l'embargo est quelque chose qui est, culturellement parlant, plus proche d'eux que des tests nucléaires au Pakistan ou des tremblements de terre en Turquie. Les origines d'un lecteur peuvent aussi avoir une influence sur la construction d'un référent. En effet, nos annotateurs avaient déjà leur propre référent sur ce sujet. Le contexte culturel constitue donc une plateforme où le concept de nouveauté est partagé entre un groupe d'individus.

5.3 Derniers mots sur la discussion

Dans les deux scénarios, nous avons proposé une stratégie de similarité de texte asymétrique pour analyser la couverture d'information. Notre stratégie s'appuie sur des théories cognitives qui attestent la présence de l'asymétrie dans le processus de comparaison (Tversky, 1977; Tversky & Gati, 1978). La comparaison comporte deux éléments, le référent qui présente les caractéristiques jugées les plus proéminentes, et le sujet de comparaison qui a des caractéristiques moins proéminentes. Il existe donc un sens dans le processus de comparaison ; tout changement de rôle des objets à comparer entraîne un changement de la valeur de similarité. La direction de comparaison dépend de nos intérêts.

–La couverture d'information dans les textes des étudiants :

En ce qui concerne la couverture d'information de textes des étudiants (premier scénario), nous utilisons WordNet pour le calcul de similarité, cela nous permet de capturer des relations lexico-sémantiques (synonymie, hyperonymie, hyponymie) afin d'évaluer la couverture de la terminologie dans les écrits des étudiants. Ces relations correspondent, d'une certaine façon, aux attaches qui assurent la cohésion lexicale selon la théorie de cohésion de texte de Halliday & Hassan (1976). Notre proposition est donc en mesure d'évaluer d'une certaine façon, la cohésion du texte. Ceci se voit refléter dans l'alignement des paragraphes des étudiants avec les RG et les RS. S'il existe une cohésion avec les sources alignées tout au long du document, nous pourrions inférer que notre approche reflète la cohérence.

Avec notre analyse, nous n'avions pas l'intention de **prédire** les notes octroyées aux dissertations comme Dascalu *et al.* (2015). Nous avons plutôt utilisé le concept de couverture d'information afin d'**expliquer** la note de chaque dissertation. Une bonne couverture de la terminologie des RG et des RS dans les dissertations des étudiants correspond nécessairement à une note de réussite.

– La couverture d'information dans les textes journalistiques :

En ce qui concerne le scénario deux, la couverture d'information dans les textes journalistiques, nous avons mené un étiquetage d'une partie du corpus de la compétition *novelty track* de la conférence *TREC*. Nous avons observé que l'origine d'un annotateur pouvait aussi avoir une influence sur la construction d'un référent, à partir duquel, les annotateurs détectaient la nouveauté. La connaissance *a priori d'une thématique* constitue aussi un référent, qui semble être utilisée par les lecteurs au moment d'identifier la nouvelle information. Comme le montrent deux thématiques en particulier, la thématique 2 (l'embargo cubain) et la thématique 5 (des tremblements de terre en Turquie), la première étant culturellement près du profil des annotateurs (l'accord entre les annotateurs est élevé) alors que la seconde thématique étant culturellement éloignée (l'accord entre les annotateurs est bas).

Par rapport à la quantité d'information, nous observons que les accords sont plus élevés lorsque la quantité d'information est peu élevée. Par exemple, la thématique 3 qui comprenait 25 phrases (l'accord étant élevé) et la thématique 4 qui comprenait 119 phrases (l'accord étant plus bas). La quantité d'information pourrait donc aussi avoir un impact sur l'identification de la nouvelle information.

En termes linguistiques, l'asymétrie est aussi présente dans la structure du langage (Reinhart, 1976; Chomsky, 1993; Kayne, 1994; Chomsky, 2005; Di Sciullo, 2016). Par exemple, Di Sciullo (2013) mentionne que la relation entre un prédicat et ses arguments est asymétrique, car ces derniers ne peuvent pas être interchangés sans affecter l'interprétation de l'événement ; aussi, tout changement dans la structure argumentale induit une autre interprétation sémantique. Notre approche est capable de refléter, en partie, cette relation asymétrique grâce aux patrons linéaires. Une méthode de similarité basée sur le concept de distance (ou toute approche symétrique) ne peut pas exprimer la relation asymétrique entre les éléments de la phrase.

Le patron N–V N–V, peut capturer grossièrement la relation de sujet d'un prédicat. La position du sujet d'un prédicat a tendance à être occupée par une entité agentive (Palmer, 1994), déclencheur d'un événement. Dans ce patron il y a deux éléments à considérer : d'abord, la partie

agentive, c'est-à-dire le patron lui-même, ensuite la stratégie de fusion qui permet de capturer grossièrement des entités complexes.

Pondérer un patron en particulier (avec le paramètre α) permet de filtrer des Mi-Es selon les critères auxquels un lecteur pourrait s'intéresser. Par exemple, l'entité déclencheur d'un évènement pourrait être capturé en donnant un poids élevé au patron N–V N–V. Si notre intérêt est de capturer les informations spatio-temporelles des évènements, le poids élevé doit aller sur le patron P–N P–N.

CONCLUSION ET RECOMMANDATIONS

Dans cette thèse, nous avons abordé la couverture d'information. Nous avons défini la couverture d'information comme une comparaison entre un référent un sujet de comparaison. Donc notre stratégie pour analyser la couverture est fondée sur la comparaison, que nous avons envisagée avec une approche asymétrique. L'asymétrie fait partie de processus cognitifs comme la comparaison ; elle est aussi une partie fondamentale de la structure du langage. Une approche de comparaison de texte asymétrique s'apparie mieux à ces deux derniers concepts qu'une approche symétrique.

Nous avons appliqué notre analyse de couverture d'information dans deux scénarios différents :

Le premier scénario correspond à la couverture de concepts dans les dissertations d'étudiants en les comparant aux sources bibliographiques suggérées dans le syllabus du cours. Nous avons implémenté un coefficient de couverture asymétrique qui est en mesure de capturer les différentes surfaces lexicales d'un même concept. En utilisant la couverture de concepts, notre approche est capable d'expliquer la note octroyée aux dissertations.

Le deuxième scénario correspond à la couverture d'information dans les textes journalistiques de type narratif. Alors, les événements sont essentiellement rapportés dans les textes journalistiques. Nous avons proposé un coefficient de couverture d'information à base de patrons linéaires afin de capturer grossièrement les relations grammaticales, qui reflètent, d'une certaine manière, la structure d'un événement. Notre approche est en mesure d'expliquer l'origine de la nouvelle information, qui n'a pas été couverte par les sources journalistiques déjà rencontrées. Il nous parerait tout à fait justifié d'utiliser des informations linguistiques afin de refléter la nature de la langue quand nous réalisons des tâches en TALN.

Les travaux qui donneront une continuité à notre recherche incluent : Pour le premier scénario :

- Implémenter une animation pour la visualisation des réseaux de mots.

- Inclure des patrons dans l'analyse des textes d'étudiants pour mettre en contexte les concepts extraits.
- Utiliser d'autres textes (les notes des cours) que les étudiants ont utilisés pour la création de leurs dissertations.

Pour le deuxième scénario :

- Implémenter une stratégie de pondération du paramètre alpha afin de trouver une configuration optimale.
- Adapter les patrons pour une analyse en langue française, et d'autres langues.
- À partir du référent complet obtenu : d'abord détecter les contradictions ; ensuite, extraire les événements factuels.

ANNEXE I

DIFFUSION SCIENTIFIQUE

a. Velazquez, E. Ratté, S. Desrosiers, C. (14 Mai, 2014). *"Alignement informatif d'un corpus bilingue de nouvelles"*. 82e Congrès de l'Acfas. Montréal, Canada. Conférence.

b. Velazquez, E. Ratté, S. Desrosiers, C. (18 décembre, 2014). *"Alignment of student texts vs teachers texts"*. Knowledge Building Advanced Learning Analytics Colloquium/Hackathon. Wageningen, Pays-Bas. Conférence.

c. Velázquez, E. et Ratté, S. (1 avril, 2016). *"Comparing news story coverages : Light vs deep analysis."* Mid-Atlantic Student Colloquium on Speech, Language and Learning (MASC-SLL) Philadelphia, PA, USA. Poster.

d. Velazquez, Erick, Ratté, S. et de Jong, F. (Septembre, 2016). *"Analyzing Students' Knowledge Building Skills by Comparing Their Written Production to Syllabus"* International Conference on Interactive Collaborative Learning. Short Paper

e. Velazquez, E., Ratté, S., et de Jong, F. (Septembre, 2016). *"Analyzing Students' Knowledge Building Skills by Comparing Their Written Production to Syllabus"*. In International Conference on Interactive Collaborative Learning (pp. 345-352). Springer, Cham.

f. Velazquez, E., Ratté, S., et de Jong, F. (Avril, 2017). *"Coverage of syllabus terminology in students' written productions : An asymmetric approach based on linguistic and cognitive knowledge"*. Article soumis au journal *Education and Information Technologies*.

ANNEXE II

ANALYSE COMPLÉMENTAIRE DE LA COUVERTURE D'INFORMATION DANS LES DISSERTATIONS

1. Nombre de documents couverts par chaque dissertation : Dissertations 3 et 4

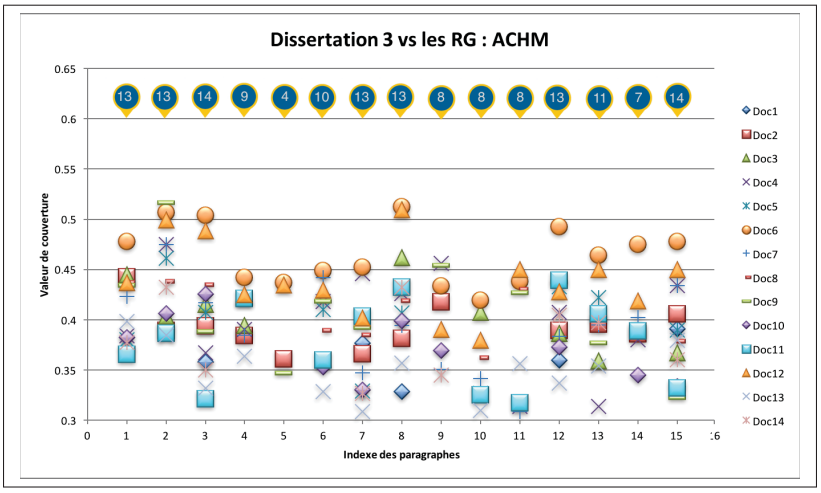


Figure-A II-1 Dissertation 1. Alignements des paragraphes avec les RG. Direction S–R.

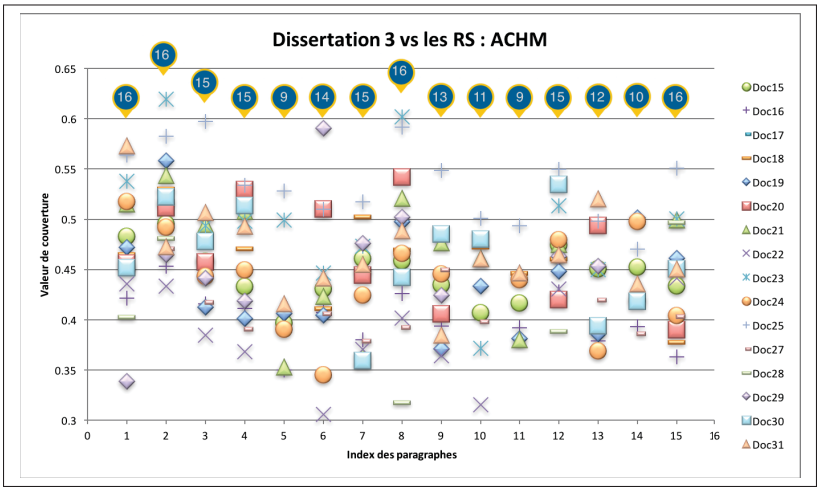


Figure-A II-2 Dissertation 1. Alignements des paragraphes avec les RG. Direction S–R.

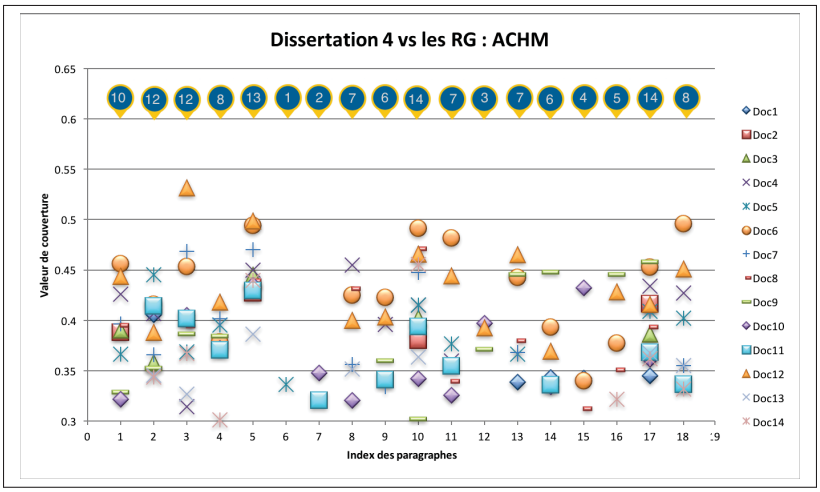


Figure-A II-3 Dissertation 1. Alignements des paragraphes avec les RG. Direction S–R.

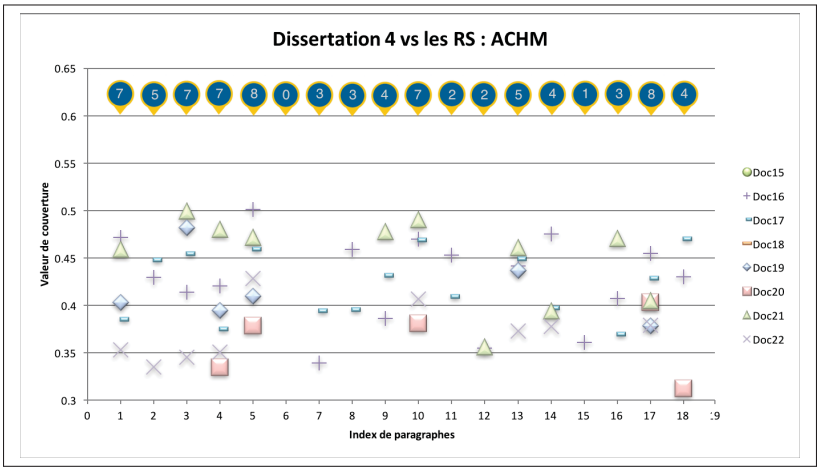


Figure-A II-4 Dissertation 1. Alignements des paragraphes avec les RG. Direction S–R.

2. Réseaux de mots : Dissertations 2 et 4

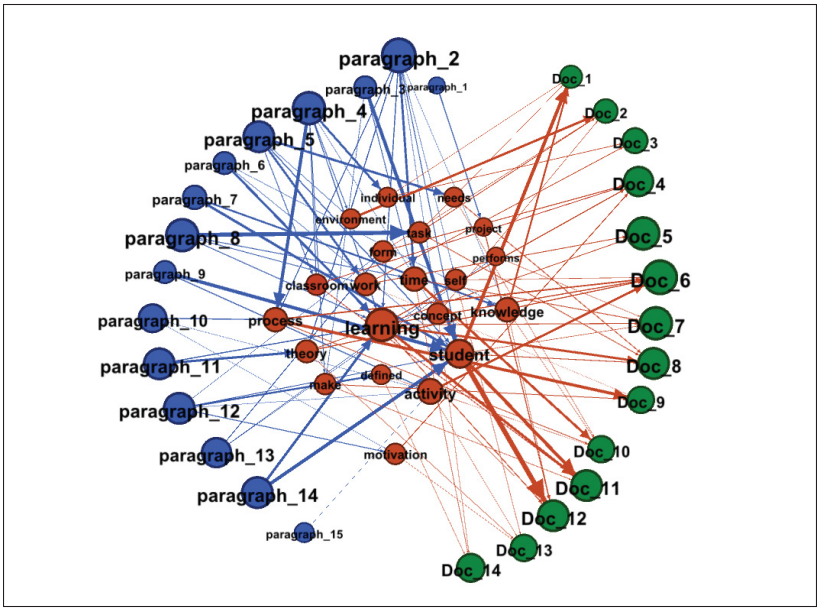


Figure-A II-5 Dissertation 2. Réseau de mots avec les RG.

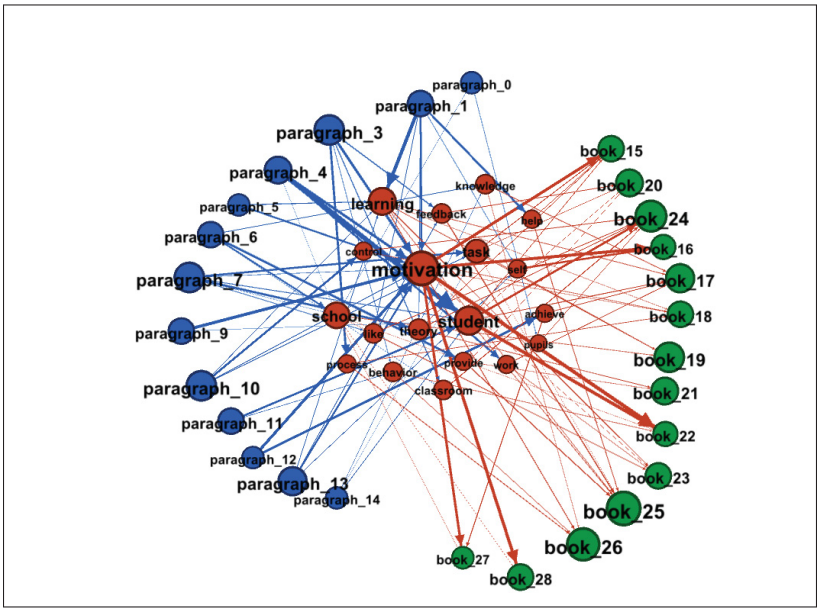


Figure-A II-6 Dissertation 2. Réseau de mots avec les RS.

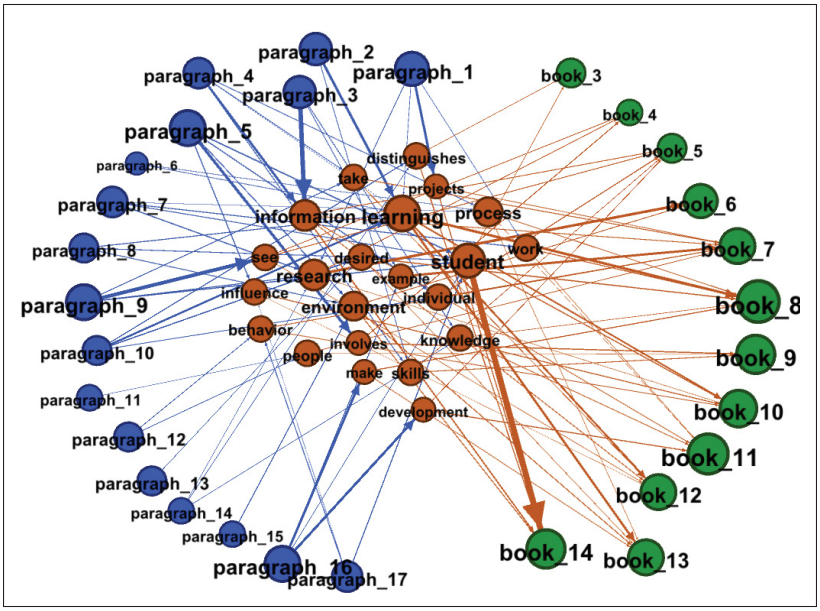


Figure-A II-7 Dissertation 4. Réseau de mots avec les RG.

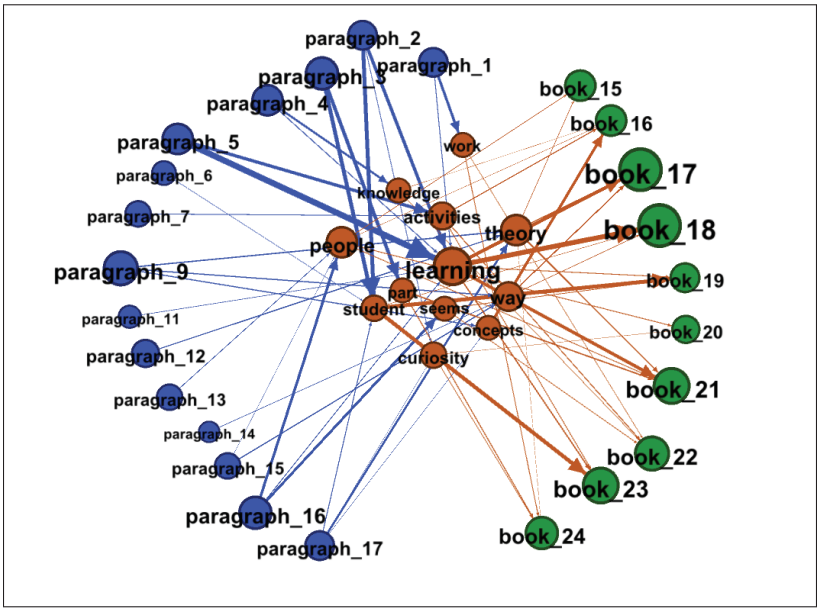


Figure-A II-8 Dissertation 4. Réseau de mots avec les RS.

ANNEXE III

GRAPHIQUES DE LA COUVERTURE D'INFORMATION DES DISSERTATIONS : DIRECTION SUJET-RÉFÉRENT

1. Nombre de documents couverts par chaque dissertation

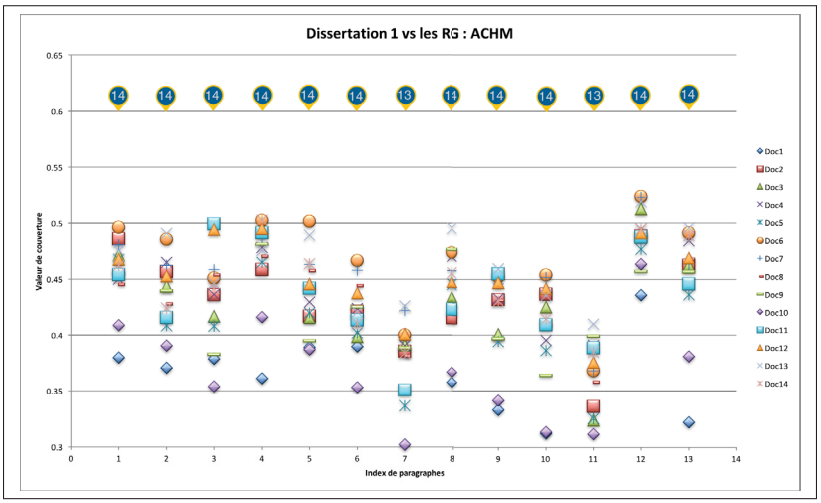


Figure-A III-1 Dissertation 1. Alignements des paragraphes avec les RG. Direction S–R.

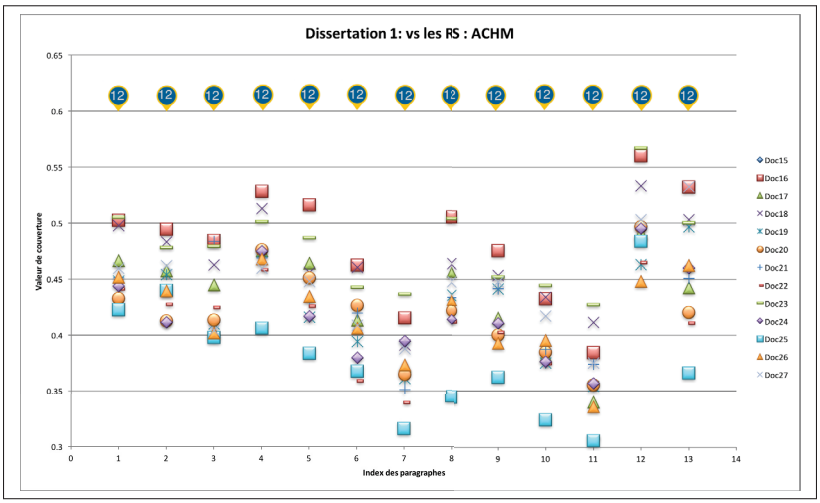


Figure-A III-2 Dissertation 1. Alignements des paragraphes avec les RS. Direction S–R.

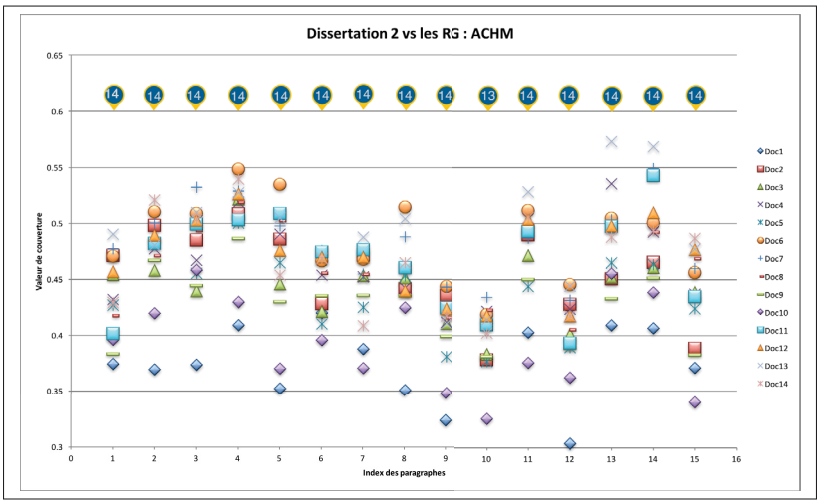


Figure-A III-3 Dissertation 2. Alignements des paragraphes avec les RG. Direction S–R.

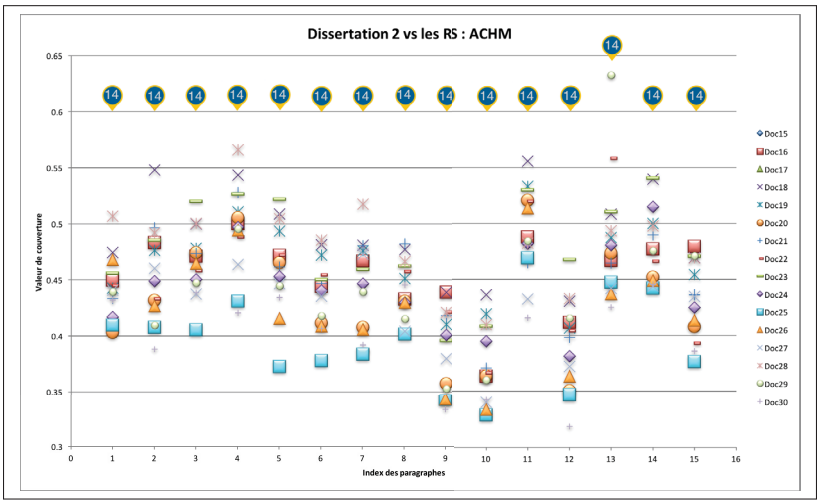


Figure-A III-4 Dissertation 2. Alignements des paragraphes avec les RS. Direction S–R.

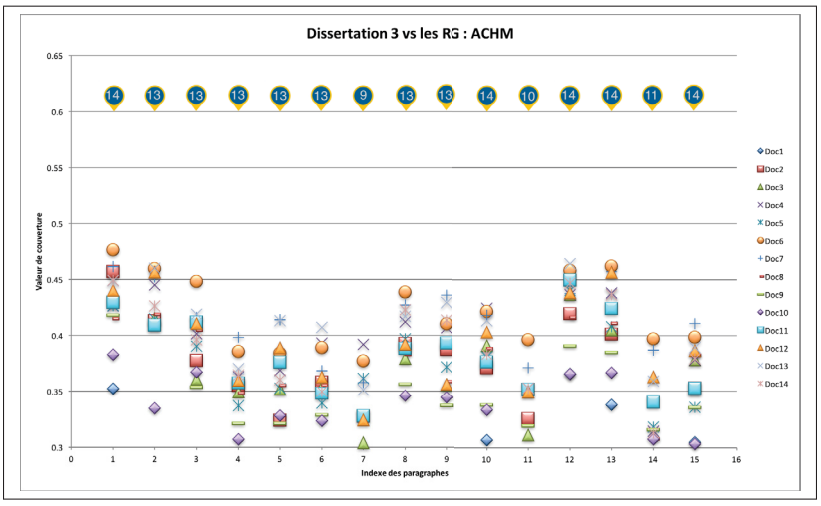


Figure-A III-5 Dissertation 3. Alignements des paragraphes avec les RG. Direction S–R.

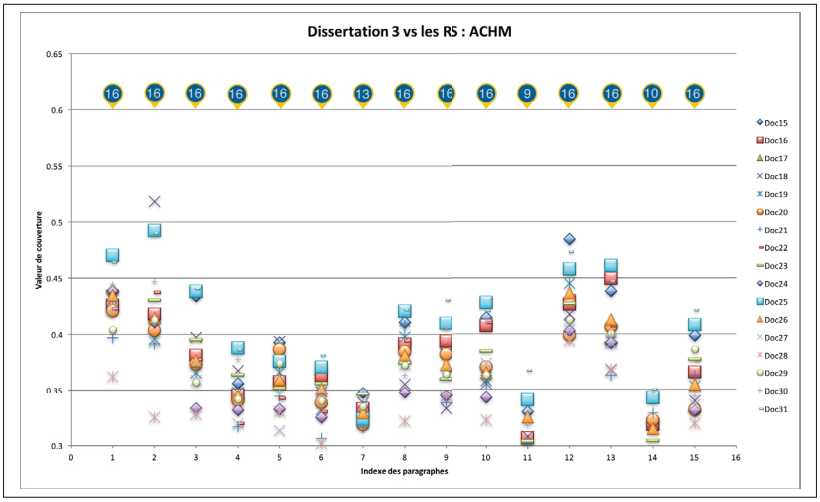


Figure-A III-6 Dissertation 3. Alignements des paragraphes avec les RS. Direction S–R.

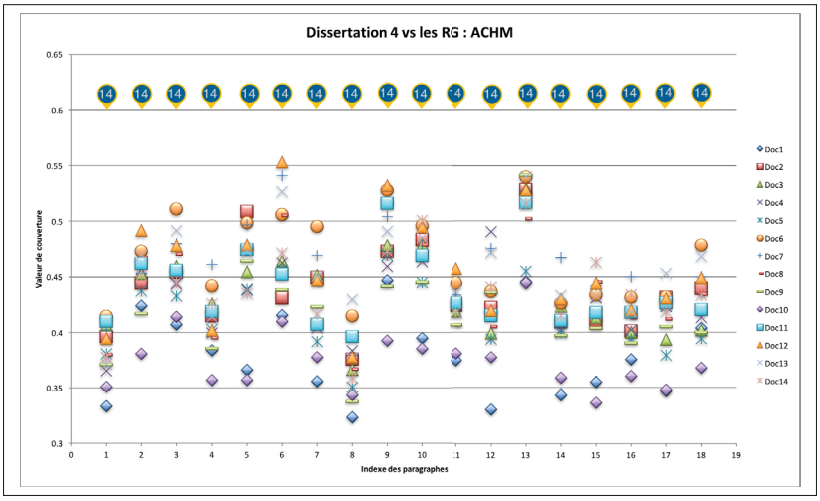


Figure-A III-7 Dissertation 4. Alignements des paragraphes avec les RG. Direction S–R.

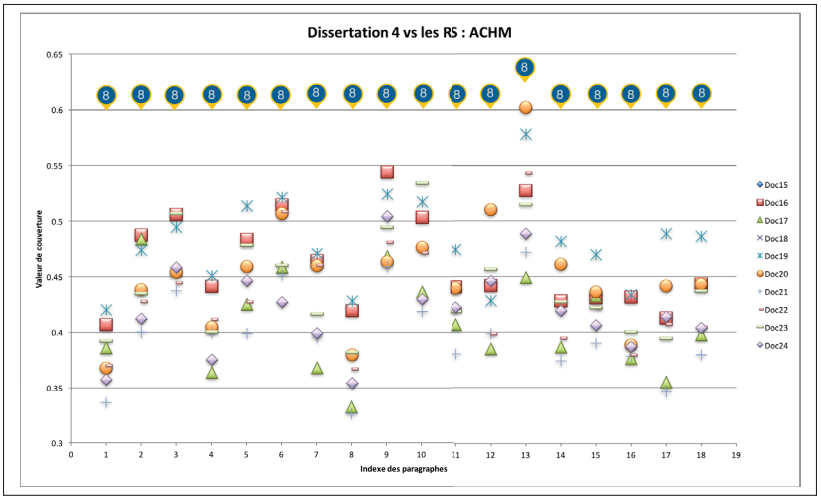


Figure-A III-8 Dissertation 4. Alignements des paragraphes avec les RS. Direction S–R.

2. L'influence des RG et des RS sur la rédaction des dissertations : Direction S–R.

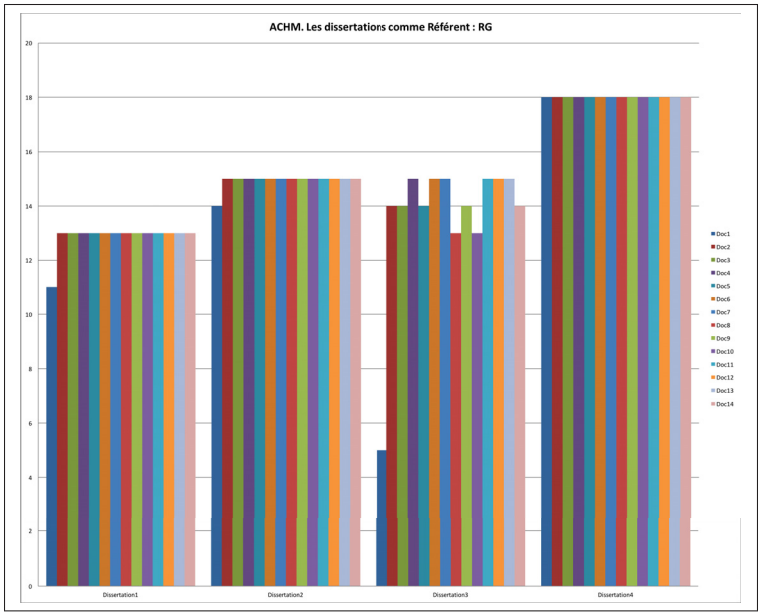


Figure-A III-9 Alignement des RG avec les 4 dissertations. Direction S–R.

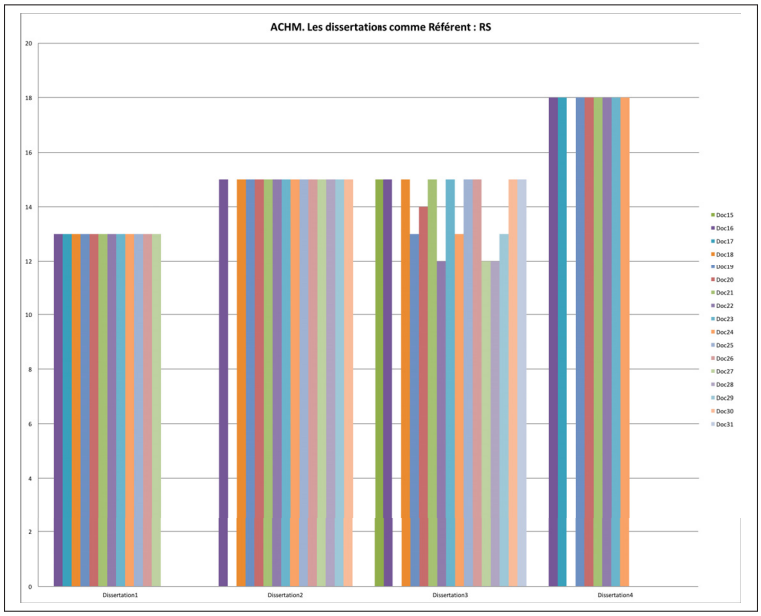


Figure-A III-10 Alignement des RS avec les 4 dissertations. Direction S–R.

ANNEXE IV

TABLEAUX DES TITRES DES DOCUMENTS DES RG ET DES RS

Tableau-A IV-1 Tableau des RG. Ces références sont partagées par les quatre dissertations.

Index	Auteurs	Titre	Type	Longueur
1	Scardamalia, M. and Bereiter, C.	A Brief History of Knowledge Building.	Article	16
2	Scardamalia, M. and Bereiter, C.	Knowledge Building : Theory, Pedagogy, and Technology.	Article	40
3	Richard, R. and Deci, E.	Self-Determination Theory and the Explanatory Role of Psychological Needs in Human Well-being.	Article	56
4	Argyris, C.	Teaching Smart People How to Learn.	Article	12
5	Illeris, K.	Contemporary Theories of Learning : Learning theorist in their words.	Livre	244
6	Scardamalia, M. and Bereiter, C.	A Brief History of Knowledge Building : Extended	Article	42
7	Cianciolo, A. and Sternberg, R.	Intelligence : A brief history.	Livre	181
8	Jossberger, H. et al.	The Challenge of Self-Directed and Self-Regulated Learning in Vocational Education : A Theoretical Analysis and Synthesis of Requirements	Article	53
9	Paavola, S. et al.	Models of Innovative Knowledge Communities and Three Metaphors of Learning.	Article	31
10	Piaget, J.	Cognitive Development in Children : Piaget Development and Learning.	Article	11
11	Hattie, J. and Timperley, H.	The Power of Feedback.	Article	33
12	Jossberger, Helen	Toward Self-Regulated Learning in Vocational Education : Difficulties and Opportunities.	Livre	161
13	Deci, E. and Richard, R.	The “What” and “Why” of Goal Pursuits : Human Needs and the Self-Determination of Behaviour.	Article	42
14	van Woerkom, M.	Critical Reflection as a Rationalistic Ideal.	Article	19

Tableau-A IV-2 Tableau des RS pour la dissertation 1.

Index	Auteurs	Titre	Type	Longueur
15	Non incluse	–	–	–
16	Bransford, J. et al.	How People Learn : Brain, Mind, Experience, and School.	Livre	386
17	Illeris, K.	How We Learn Learning and non-learning in school and beyond.	Livre	304
18	Ruijters, M	Liefde voor leren. Over de diversiteit van leren en ontwikkelen in en van organisaties.	Article	23
19	Rubens, W.	E-learning Trends en ontwikkelingen.	Livre	234
20	Wigfield, A.	Expectancy–Value Theory of Achievement Motivation.	Article	14
21	Mayer, R. and Alexander, P.	Handbook of Research on Learning and Instruction.	Livre	516
22	Richard, R. and Deci, E.	Intrinsic and Extrinsic Motivations : Classic Definitions and New Directions.	Article	14
23	Gillet, N. et al.	Intrinsic and extrinsic school motivation as a function of age : the mediating role of autonomy support.	Article	19
24	de Brabander, K. and Martens, R.	Ontwerp van een conceptueel kader.	Article	32
25	Eccles, J. and Wigfield, A.	Motivational Beliefs, Values, and Goals.	Article	24
26	R. L. Martens.	Positive learning met multimedia Onderzoeken, toepassen and generaliseren.	Livre	80
27	Simons, R. and Ruijters, M.	Varieties of work related learning.	Article	10

Tableau-A IV-3 Tableau des RS pour la dissertation 2.

Index	Auteurs	Titre	Type	Longueur
15	Non incluse	–	–	–
16	Bransford, J. et al.	How People Learn : Brain, Mind, Experience, and School.	Livre	386
17	Non incluse	–	–	–
18	Martens, R.L.	Liefde voor leren. Over de diversiteit van leren en ontwikkelen in en van organisaties.	Article	23
19	Perry, N. et al.	Classrooms as Contexts for Motivating Learning.	Article	22
20	Korthagen, F.	“Ik heb er veel van geleerd !” Een reflectie over effectief opleiden en krachtgericht coachen.	Article	20
21	van der Veen, T.P.	Een tweede onderzoek naar de beïnvloeding van motivatie bij vmbo-leerlingen.	Rapport	58
22	Wigfield, A.	Expectancy–Value Theory of Achievement Motivation.	Article	14
23	Bergmann, J. and Sams, A.	Flip YOUR Classroom Reach Every Student in Every Class Every Day Reach.	Livre	124
24	Mayer, R. and Alexander, P.	Handbook of Research on Learning and Instruction.	Livre	516
25	Richard, R. and Deci, E.	Intrinsic and Extrinsic Motivations : Classic Definitions and New Directions.	Article	14
26	Gillet, N. et al.	Intrinsic and extrinsic school motivation as a function of age : the mediating role of autonomy support.	Article	19
27	Schuit, H. et al.	Leerlingen motiveren : een onderzoek naar de rol van leraren.	Rapport	70
28	Eccles, J. and Wigfield, A.	Motivational Beliefs, Values, and Goals.	Article	24
29	Fei-Yin, F., et all.	Children’s Achievement Moderates the Effects of Mothers’ Use of Control and Autonomy Support	Article	67
30	Sanneke Bolhuis.	Leerstrategieën, leren en verantwoordelijkheid.	Rapport	40

Tableau-A IV-4 Tableau des RS pour la dissertation 3.

Index	Auteurs	Titre	Type	Longueur
15	Bereiter, C.	Education and Mind in the Knowledge Age.	Livre	541
16	Bransford, J. et al.	How People Learn : Brain, Mind, Experience, and School.	Livre	386
17	Non incluse	–	–	–
18	Ruijters, M.	Liefde voor leren. Over de diversiteit van leren en ontwikkelen in en van organisaties.	Article	23
19	Lave, J. and Wenger, E.	Situated Learning Legitimate Peripheral Participation.	Livre	139
20	Sloep, P. et al.	Leernetwerken.	Livre	201
21	Berger, M. et al.	Actieplan Professionalisering Jeugdzorg.	Rapport	63
22	Bereiter, C.	Can Children Really Create Knowledge ?	Article	24
23	Bood, R. and Coenders, M.	Communities of Practice : Bronnen van inspiratie	Article	3
24	Simons, R. and Nijmegen, K.U.	Competentiegerichte leeromgevingen in organisaties en hoger beroepsonderwijs.	Article	14
25	de Jong, R	Doen, Leren en Kenniscreatie : Verstand en Competentie.	Livre	89
26	Bolhuis, S.	Leerstrategieën, leren en verantwoordelijkheid.	Article	42
27	Simons, R.	Mindshifting : (Hoe) kunnen we mindsets veranderen ?	Article	26
28	Lammersen, G. and Vlaar, P.	Naar een eigentijds Hbo- arrangement voor de gehandicaptenzorg.	Rapport	52
29	Paavola, S. et al.	Models of Innovative Knowledge Communities and Three Metaphors of Learning.	Article	21
30	Van Yperen, T. and Westering, Y.	Pijlers voor nieuw jeugdbeleid.	Article	12
31	Van Biene, M.	Wederkerig leren. Onderzoek naar georganiseerde leerondersteuning voor mensen met een verstandelijke beperking én professionals.	Livre	463

Tableau-A IV-5 Tableau des RS pour la dissertation 4.

Index	Auteurs	Titre	Type	Longueur
15	Non incluse	–	–	–
16	Bransford, J. et al.	How People Learn : Brain, Mind, Experience, and School.	Livre	386
17	Illeris, K.	How We Learn Learning and non-learning in school and beyond.	Livre	304
18	Non incluse	–	–	–
19	Engel, S.	The case for curiosity.	Article	14
20	Litman, J.	Curiosity and the pleasures of learning : Wanting and liking new information.	Article	23
21	Ruijters, M.	Love of Learning About diversity in learning and development.	Article	8
22	Paavola, S. et al. .	Models of Innovative Knowledge Communities and Three Metaphors of Learning.	Article	20
23	Lucas, B. et al..	Progression in Student Creativity in School.	Article	46
24	Chak, A.	Understanding Children's Curiosity and Exploration through the Lenses of Lewin's Field Theory : On Developing an Appraisal Framework.	Article	12

ANNEXE V

SURVOL SUR L'HISTOIRE DE LA SYMÉTRIE VS L'ASYMÉTRIE

Le concept de symétrie [1] a été employé depuis l'antiquité principalement dans le domaine de l'architecture et de l'art (peinture, sculpture et aussi musique). En architecture, nous trouvons le texte de Vitruvius qui définit la symétrie comme : « ... *a proper agreement between the members of the work itself, and relation between the different parts and the whole general scheme, in accordance with a certain part selected as standard* [2] » (Vitruvius, 2009). De plus,Vitruvius (2009) considère que le corps humain présente une harmonie symétrique entre les avant-bras, les pieds, les paumes, les doigts et les autres petites parties du corps. Marcus *Vitruvius* Pollio était un architecte romain, ayant vécu au premier siècle av. J.C., et qui a fait un effort pour rassembler toutes les normes de constructions architecturales de l'antiquité. Dans son œuvre en dix volumes, *De architectura*, Vitruvius (2009) présente la symétrie comme un moyen pour atteindre l'eurythmie d'une œuvre :

> « *Eurythmy is beauty and fitness in the adjustments of the members. This is found when the members of a work are of a height suited to their breadth, of a breadth suited to their length, and, in a word, when they all correspond symmetrically* [3]. »
> (Livre I)

Selon Vitruvius, la beauté d'un bâtiment dépend de la bonne utilisation des concepts de symétrie : « ... *and beauty (will be assured), when the appearance of the work is pleasing and in good taste, and when its members are in due proportion according to correct principles of symmetry* [4] » (Vitruvius, 2009, Livre I). Vitruvius assure que la gloire de l'accomplissement d'une œuvre est octroyée de la façon suivante : quand l'œuvre est réalisée somptueusement, c'est le propriétaire qui devrait être reconnu pour les grandes dépenses qu'il a autorisées. Si

1. Du mont grec συμμετρία, symmetria.
2. N.T. ... l'accord approprié entre les membres d'une même œuvre, et la relation entre les différentes parts et le schème général du tout, en accord avec une certaine partie sélectionnée comme standard
3. N.T. L'eurythmie c'est beauté et finesse dans l'ajustement des membres. Ceci est trouvé quand les membres d'une œuvre correspondent à une hauteur adapté à leur largeur, à une largeur adaptée à leur longueur, c'est à dire quand tous correspondent symétriquement.
4. N.T. ...et la beauté, (est assurée) quand l'apparence du travail est plaisante et du bon goût, quand ses membres sont conçus proportionnellement aux principes corrects de symétrie.

l'œuvre est réalisée délicatement, alors c'est le maître d'œuvre qui sera reconnu, mais quand la proportion et la symétrie prêtent un imposant effet, la gloire appartient à l'architecte (Vitruvius, 2009, Livre VI).

Aristote (2016) aborde la capacité des mathématiques d'exprimer le concept de beauté, selon lui, les principales formes de beauté sont l'ordre, la définitude et la symétrie pour lesquelles les sciences mathématiques portent un intérêt spécial. *Cosmos* est le mot grec [5] pour designer l'ordre, et les anciens Grecs étaient fascinés par le concept de formes. Dans un univers de formes, les Grecs ont défini une distinction entre *chaos* et *cosmos*, c'est-à-dire l'*ordre*, ce qui n'est pas beau et ce qui l'est (Mitchell, 1990). En sciences, les Grecs ont fait des efforts pour découvrir une cosmologie, c'est-à-dire un système sous-jacent de la forme du monde, Mitchell (1990). Par exemple, Pythagore explique la structure du *cosmos* en termes de nombres et de géométrie. Il croyait que l'élégance des mathématiques qu'il avait trouvée dans les figures abstraites était aussi présente dans le monde naturel (Buckingham *et al.*, 2011). Pythagore croyait aussi que les nombres exprimaient la réalité ultime de l'univers (Gull, 2016). Ainsi la beauté, en utilisant les concepts de ration et symétrie, pourrait être exprimée en termes mathématiques (Gull, 2016).

En ce qui concerne la musique, les mathématiciens grecs qui ont suivi la tradition de Pythagore, ont observé des ratios et des proportions en faisant sonner une corde pincée. La hauteur de la note produite était proportionnelle à la longueur de la corde. Des combinaisons de notes harmonieuses étaient obtenues par les longueurs des cordes respectant certains ratios de petits nombres entiers, 1 :2 (octave), 2 :3 (quinte) et 3 :4 (quarte), Mitchell (1990). Voici comment la symétrie est aussi présente dans le domaine de la musique.

Dans le domaine des arts, les Grecs ont aussi développé des conventions explicites pour les compositions (Mitchell, 1990). Une théorie artistique peut être retracée dans le *canon* [6] de *Polyklète* qui applique les concepts de la géométrie, ratio, proportion et symétrie. Son système

5. La transcription phonétique kosmos du mot κόσμος.

6. *Canon* est la transcription phonétique (kanon) et anglaisisée du mot grec κανον) . La traduction du mot grec est "règle".

utilise une moyenne géométrique en progression continue, (Tobin, 1975). Alors la symétrie d'une sculpture était atteinte en rapportant les dimensions de toutes les parties d'une statue à chacune, et à l'ensemble par le moyen d'un système approprié de ratios, Mitchell (1990). Selon, Tobin (1975) ce *canon* aurait pu avoir une influence importante chez les mathématiciens qui ont suivi la tradition de Pythagore.

Pendant l'époque de la renaissance, la redécouverte du texte de Vitruvius a permis aussi aux artistes d'ancrer le concept de symétrie dans leurs propres œuvres. Par exemple, *Léonard Da Vinci* dessine l'homme de Vitruve, qui correspond aux descriptions de proportion et de symétrie du corps humain proposées par Vitruvius (2009). En 1673, dans la traduction française du texte de Vitruvius, Claude Perrault ajoute une note à la définition de symétrie. Il présentait la symétrie comme la relation dans laquelle, les parts du côté gauche sont similaires avec celles du côté droit, celles en haut avec celles en bas, et finalement celles en avant avec celles en arrière, (Mitchell, 1990). Cette définition s'aligne avec la signification contemporaine de symétrie, qui réfère à une symétrie bilatérale, le type de symétrie présente le corps humain. Cela a influencé la composition d'architecture classique que nous pouvons observer au Parthénon, (Mitchell, 1990) et au palais de Versailles, construit par Louis XIV, admirateur de la culture grecque classique.

Les mathématiques modernes ont formalisé le concept de symétrie géométrique en termes d'un ensemble de transformations géométriques possibles. Ces transformations sont des translations, rotations, réflexions (Mitchell, 1990). Dans un plan cartésien, par exemple, ces opérations peuvent se réaliser à partir d'un point de repère, d'un axe ou même sur l'arête d'une figure géométrique.

Comme nous venons de le voir, il existe une longue tradition qui prône pour une utilisation de la symétrie. Cette tradition a eu un impact aussi dans d'autres domaines que l'art ou l'architecture. L'idée d'une symétrie naturelle, grâce à la composition du corps humain, a aussi mené à son application pour décrire des processus cognitifs tels que la comparaison. Pour ce faire, le concept géométrique de distance a dû être emprunté. La distance est utilisée pour déterminer

le degré de similarité entre deux objets, A et B, qui ont été projetés dans un espace coordonné. La distance est une fonction symétrique, car la distance entre les objets A et B est la même qu'entre B et A.

Cependant, Tversky (1977) postule que la similarité est une relation asymétrique et qu'elle est mieux décrite comme une correspondance entre caractéristiques (ou un processus d'appariement) plutôt qu'un calcul de la distance entre deux points. Dans sa proposition, Tversky (1977) considère que chaque élément à comparer détient un rôle différent. C'est ainsi qu'il distingue le référent et le sujet de comparaison. Le référent c'est l'objet de comparaison qui détient les caractéristiques ou les stimuli les plus proéminents. Le choix de l'objet de comparaison qui jouera le rôle de référent dépend de l'importance qui est attribuée aux caractéristiques de l'objet. Le sujet de comparaison est généralement l'objet ayant des caractéristiques moins proéminentes, (Tversky, 1977). Il existe donc une direction dans la comparaison qui dépend de la proéminence des caractéristiques des objets à comparer, (Tversky, 1977). Pour mieux comprendre la différence entre référent et sujet de comparaison, Tversky mentionne que les jugements de similarité peuvent être envisagés comme l'extension de l'énoncé de similarités, tel que : *A est comme B*; où *A* est le sujet de comparaison et *B* le référent. Dans ce sens, nous aurons aussi des énoncés en langue naturelle comme : *Le fils ressemble à son père*, ou *le portrait ressemble à la personne*. Le choix des énoncés de similarité est associé avec la symétrie/asymétrie en jugement de similarité, (Tversky, 1977).

Formellement, Tversky (1977) définit la similarité de la façon suivante :

$$S(A,B) = F(A \cap B, A - B, B - A) \qquad \text{(A V-1)}$$

Où $F()$ est une fonction de similarité, $A \cap B$ représente les caractéristiques communes entre A et B. $A - B$ représente les caractéristiques qui appartiennent seulement à A. $B - A$ représente les caractéristiques qui appartiennent seulement à B. Cette formalisation est illustrée sous la forme d'un diagramme à la figure V-1.

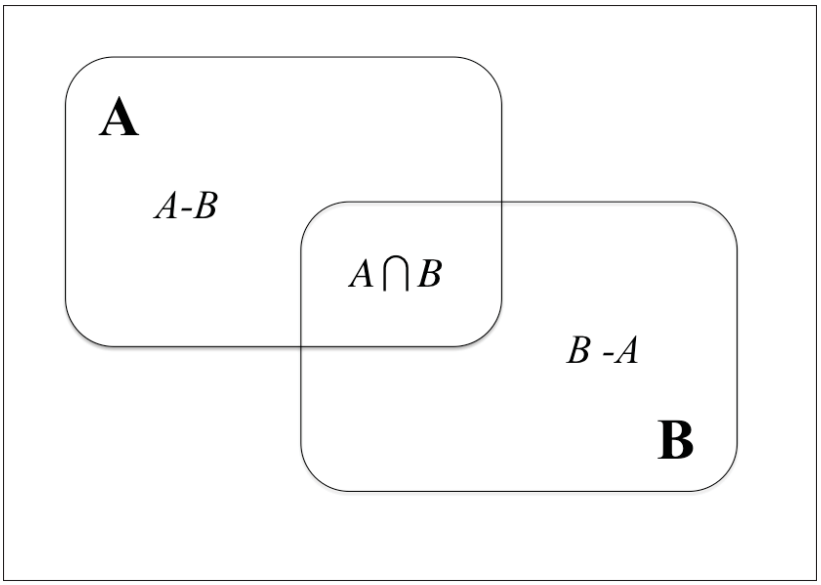


Figure-A V-1 Diagramme de similarité de Tversky (1977).

Tversky (1977) indique également que la notion de similarité symétrique ne doit pas être rejetée complètement ; elle est valable dans de nombreux contextes, et dans beaucoup d'autres, il s'agit d'une approximation utile. Par ailleurs, il souligne que la similarité symétrique ne peut être acceptée comme un principe universel de similarité en psychologie. En outre, Tversky (1977) montre que la notion de similarité asymétrique a été observée dans les tâches de comparaison où les gens comparent deux objets pour déterminer leur degré de similarité, (Tversky, 1977).

En 1992, Leyton utilise les concepts de symétrie et asymétrie dans sa théorie sur la perception et la cognition dans son œuvre *Symmetry, Causality, Mind*. Il présente la symétrie comme un élément nécessaire à toute activité cognitive quotidienne. Leyton (1992) a recours au problème de récupération du processus pour expliquer sa théorie. Ainsi, Leyton (1992) pose le problème de récupération comme : supposons qu'un individu observe un état, qui est appelé *moment présent*. Une certaine caractéristique structurelle de ce moment permet à la personne de reculer dans le temps et déduire les processus qui ont mené à ce moment présent. Le problème de récupération du procès représente donc les efforts d'une personne à récupérer les processus passés d'un moment repère. Comme solution à ce problème, Leyton (1992) présente deux principes :

- le principe de symétrie : Une symétrie dans le présent est comprise comme ayant existée depuis toujours. La symétrie est l'absence de processus-mémoire.
- le principe d'asymétrie : Une asymétrie dans le présent est interprétée comme provenant à partir d'une symétrie passée. L'asymétrie est la mémoire qu'un processus laisse sur un objet.

Alors, Leyton (1992) considère la mémoire, qu'un processus laisse sur un objet, comme l'élément principal pour identifier la symétrie et l'asymétrie. Prenons l'exemple de Leyton (1992) pour mieux le comprendre : Supposons qu'un réservoir de gaz reste stable dans une chambre et que le gaz aie atteint son équilibre dans un premier temps, voir fig V-2-A. Pour chaque position dans le réservoir, la concentration de gaz est équivalente, si l'on trace un axe sur la ligne verticale juste au milieu du réservoir. Maintenant, au temps 2, supposons que nous utilisons un aimant sur le côté gauche. Cet aimant entraîne le déplacement du gaz ainsi qu'une augmentation dans les particules du gaz du même côté du réservoir, voir V-2-B. La distribution du gaz est devenue asymétrique. Leyton (1992) mentionne que si une personne rentre dans la chambre, elle pourrait conclure qu'il y a eu un changement qui a provoqué une concentration de gaz sur le côté gauche de la chambre, cela même si la personne n'a pas vu de mouvement. L'asymétrie agit donc comme une mémoire du mouvement. Si dans un temps 3, voir V-2-C, le gaz atteint encore une fois l'équilibre dans le réservoir, et qu'une personne, qui n'a pas encore été dans la chambre, y accède, elle ne pourrait pas dire que quelque chose s'est passée. Ainsi, le gaz revenu en état symétrique a effacé toute mémoire de l'événement passé. Alors la symétrie dans le présent ne permet pas de déduire une différence dans le passé, Leyton (1992).

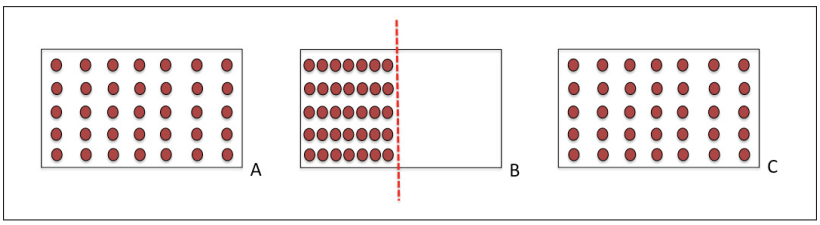


Figure-A V-2 Reproduction du diagramme d'un réservoir de gaz en trois temps, A, B, et C. Example pris de Leyton (1992, p. 8).

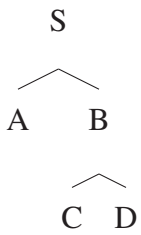
Leyton (1992) applique aussi sa théorie en linguistique et en art. En ce qui concerne la linguistique, Leyton (1992) affirme que les arguments des linguistes, en grammaire générative, qui justifient l'existence des opérations de mouvement dans les structures syntaxiques, sont basés sur les principes de symétrie et d'asymétrie. Le principe d'asymétrie agit en distinguant les capacités dans l'information positionnelle. Le principe de symétrie, quand à lui, est instancié par le principe de projection.

Dans les théories syntaxiques actuelles, qui sont basées sur la grammaire générative, les expressions linguistiques peuvent être représentées en termes de graphes orientés, Di Sciullo (2013). En syntaxe, les concepts de symétrie et d'asymétrie sont identifiés dans les relations structurelles de la phrase. Ces relations peuvent être la préséance, la domination et la *C-command*, Carnie (2015). La *C-command* est peut-être l'une des relations structurelles de la phrase les plus importantes Carnie (2015). Un nœud *C-commande* [7] ses sœurs et toutes les filles et les petites-filles de ses sœurs. Carnie (2015) présente la définition de deux types de *C-command* :

- *C-command* symétrique : Relation entre deux nœuds soeurs. Un nœud *A C-commande* symétriquement *B* si *A C-commande B* et *B C-commande A*, (Carnie, 2015).
- *C-command* asymétrique : Relation entre un nœud tante et ses nièces et les descendantes de celle-ci. Le nœud *A C-commande* asymétriquement *B* si *A C-commande B* mais *B* ne *C-commande* pas *A*. (Carnie, 2015).

Dans l'exemple suivant, emprunté de Carnie (2015), le nœud *A C-commande* symétriquement le nœud *B* et vice-versa. Le nœud *A C-commande* asymétriquement les nœuds *C* et, *D* mais ces derniers ne *C-commandent* pas le nœud A.

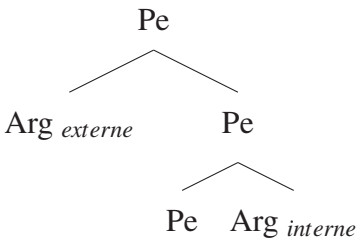
7. Nous ajoutons un C majuscule comme préfixe au verbe commander pour indiquer le même sens que *C-command en anglais.*



Le concept d'asymétrie a été largement discuté en linguistique, par exemple, dans la Théorie d'Asymétrie (TA), originalement proposée en morphologie par Di Sciullo (2005). La TA tient compte du fait qu'un changement dans les relations asymétriques dans un objet morphologique provoque soit un charabia ou une interprétation sémantique différente, (Di Sciullo, 2005). Les relations syntaxiques sont aussi asymétriques ; une inversion des constituants n'entraîne pas un charabia, mais l'altération des relations sémantiques et de l'information, Di Sciullo (2013). À propos de la sélection des arguments du prédicat, (Di Sciullo, 2013) mentionne :

> « *Argument structure relations are asymmetric in the sense that a predicate asymmetrically selects an argument, whereas the inverse relation does not hold : an argument does not asymmetrically select a predicate.* » [8]

La figure suivante [9] présente la structure d'un prédicat dyadique avec son argument externe (habituellement appelé sujet) et son argument interne (habituellement appelé objet) :



Nous constatons la relation asymétrique sur cet arbre en appliquant les concepts de *C-command* asymétrique. Si nous nous alignons avec Di Sciullo (2013), la structure des prédicats dénote des événements et le noyau d'un événement peut être modifié par des adjoints, c'est à dire par d'autres arguments internes au prédicat, par exemple des adjoints de localisation spatiale

8. N.T. La relation de la structure argumentale est donc asymétrique dans le sens qu'un prédicat asymémtriquement sélectionne un argument, tandis que l'inverse ne peut pas avoir lieu. Un argument ne sélectionne asymétriquement pas son prédicat.

9. L'arbre syntaxique et l'exemple fut extrait de Di Sciullo (2013)

ou temporelle. Ainsi, l'asymétrie est également une propriété de la localisation spatiale d'un événement, (Di Sciullo, 2013). La TA prédit correctement qu'il devrait y avoir une asymétrie entre le point d'origine d'un événement et le point final de cet événement (Di Sciullo, 2013).

Chomsky (2005) propose que le développement du langage chez un individu est déterminé par la génétique, l'expérience ainsi que par des principes d'efficacité computationnelle. Di Sciullo (2016) présente de l'évidence de la présence de l'asymétrie dans la faculté du langage humain en discutant deux principes d'efficacité computationnelle : *minimize symmetrical relations* et *minimize externalization*. En particulier, le principe *minimize symmetrical relations* s'applique aussitôt que possible dans les dérivations syntaxiques et élimine les relations symétriques (Di Sciullo, 2016). Ce qu'il est nécessaire de comprendre ici, c'est que l'application du principe *minimize symmetrical relations* a pour effet d'éliminer les relations symétriques en déplaçant un ou plusieurs constituants. D'après la théorie de Leyton (1992), l'application de ce principe correspond ainsi au concept de mémoire, qui se trouve ici instanciée par la trace des déplacements et agit comme une récupération du processus. On rejoint, ainsi, le principe asymétrique de Leyton, puisqu'on agit sur une symétrie "du passé" pour générer une asymétrie dans la structure "dans le present".

En traitement de langues naturelles TLN, la comparaison de textes se fait aussi, la plus part du temps, avec des approches symétriques. Ceci est dû à l'utilisation des modèles d'espaces géométriques, comme *Vector Space Model*, et le modèle de sac à mots pour calculer la similarité cosinus, voir formule 1.2. Si l'objectif d'une comparaison est de déterminer le degré de similarité entre deux textes, nous devons considérer les propriétés du langage. Pour les raisons exposées précédemment, l'asymétrie est une propriété structurelle du langage donc une approche de comparaison asymétrique se rapproche donc plus de la réalité du langage qu'une approche symétrique.

Les concepts de symétrie et asymétrie sont utilisés aussi dans la conception de techniques d'interaction humain-machine [10]. Dans ce sens, nous trouverons alors des manipulations symé-

10. Nous faisons référence ici aux techniques de manipulation sur des interfaces tactiles, soit un écran tactile ou un téléphone intelligent

triques et asymétriques sur les objets dans une interface. La manipulation symétrique se produit quand la main dominante et la main non dominante [11] partage le même espace de travail dans un espace de temps dit coordonné . L'exemple le plus commun d'une manipulation symétrique c'est un mouvement de zoom et de rotation que nous faisons sur une image, Velazquez Godinez (2012). La manipulation asymétrique, moins reperdue en techniques d'interaction, se présente comme les plus naturelles aux utilisateurs lors la réalisation des tâches comme le dessin de figures, voir Velazquez Godinez (2012) pour une discussion plus profonde à se sujet.

11. Le terme de main dominante ou non dominante est utilisé pour dessiner la préférence d'un utilisateur à se servir d'une main en particulier pour réaliser une tâche, soit la main droite ou la main gauche. La main dominante est celle qui réalise les détails les plus fins lors d'une tâche. Par conséquent, la fonction de la main non dominante est de fournir un support supplémentaire pour que la main dominante puisse réaliser la tâche. Pensons quand nous écrivons une lettre sur papier : la main dominante tient le stylo et la main non dominante tient le cahier pour que celui-ci ne bouge pas.

ANNEXE VI

INSTRUCTIONS POUR LA NOUVELLE ANNOTATION DU CORPUS *NOVELTY* TREC

Reading the first file :

1.-Every sentence in the file will have the following format :

<s docid="APW19980601.1458" num="27"> { NP_1 Security Council 's permanent members} {VP_2 meet} {NP_3 later this week} {PP_4 in NP_5 Geneva} .</s>

- NP only contains nouns, proper names and pronouns.
- VP only contains verbs.
- PP contains a preposition fallowed by a NP. Note that in this case the { } contains both the PP and the NP.
- Please ignore groups that are not listed or in { }.

As you can see, the sentence has been divided into groups. Put attention on these groups, the order in which they appear, or their combinations.

2.- For each topic you will have a first document. Please take the time to read it and keep it in mind as much as possible.

3.-In the case where two NPs are separated by the conjunction "and" please consider them as one group.

4.-Consider this first document as a referent to detect the new information that you will read in the others documents.

5.-The goal of this task is to find the sentences that bring new information. This new information could appear in the sentences as one group, a combination of two groups or more that did not appear in the reference document.

How to fill up the file of tagging :

6.-The annotation steps will be done as follows :

- In the document topic_new_information.txt you have 3 columns :
  - docid num What's new (a group, pair, combination)
- The columns name corresponds to the information of each sentence in the <s> tag. Inside you will find the docid and the num. Finaly the tagid correspond to the number after the underscore. For example, if we tag the information of the first group NP + VP in sentence in the bullet point 1 we could have :
  - APW19980601.1458 27 NP_1,VP_2
- For the first group of VP + NP we have :
  - APW19980601.1458 27 VP_2,NP_3
- Specially, we consider the tag PP as a group please put the information in the topic_new_-information.txt file as follows :
  - APW19980601.1458 27 PP_4,PP_4

Please separate the elements that you consider as new by a comma.

Reading the rest of files

7.- Now when you read the second document please select the group of tags (NP +VN, etc, see The format of the sentence section) that you consider as new. Please consider a new group of tags a different order of the same tags that were not presented in the referent document. For example :

In the referent document you have the following sentence : <s docid="APW19980601.1458" num="20"> NP_1 Barack Obama VP_2 congratulated NP_3 Hillary Clinton.</s>

In the second file you have : <s docid="APW19980605.1458" num="2"> NP_1 Hillary Clinton VP_2 congratulated NP_3 Barack Obama.</s>

Both sentences contain the same lexical items but in different order, so please consider the configuration in the second file as new information.

As you read a new file, consider this new file a part of the referent in order to determine the new information of the next file.

Example of the first and second file :

First file (referent)

a. <s docid="XIE19970831.0129" num="4"> PARIS, August 31 (Xinhua) ñ NP_1 Britain NP_2 ’s Princess Diana VP_3 died PP_4 in NP_5 hospital ADVP here NP_6 early Sunday PP_7 after NP_8 a car crash , VP_9 said NP_10 French Interior Minister Jean-Pierre Chevenement .</s>

b. <s docid="XIE19970831.0129" num="5"> NP_1 Chevenement VP_2 was speaking PP_3 at NP_4 the Pitie-Salpetriere Hospital PP_5 in NP_6 south-east Paris ADVP where NP_7 Diana VP_8 was treated PP_9 after NP_10 the accident NP_11 which VP_12 occurred PP_13 at NP_14 midnight (NP_15 2200 GMT).</s>

c. <s docid="XIE19970831.0129" num="6"> NP_1 The 36-year-old princess VP_2 died PP_3 at NP_4 4 a.m. local time PP_5 after VP_6 going PP_7 into NP_8 cardiac arrest , NP_9 doctors VP_10 told NP_11 reporters.</s>

d. <s docid="XIE19970831.0129" num="7"> NP_1 The accident VP_2 happened PP_3 at NP_4 midnight PP_5 at NP_6 a Paris road tunnel PP_7 under NP_8 the Alma bridge PP_9 on NP_10 the Seine river bank .</s>

e. <s docid="XIE19970831.0129" num="8"> NP_1 Her companion Egyptian millionaire Dodi el-Fayed , NP_2 41 , and NP_3 the driver VP_4 were killed ADVP outright .</s>

f. <s docid="XIE19970831.0129" num="9"> NP_1 Her bodyguard VP_2 was ADVP also seriously UCP injured .</s>

g. <s docid="XIE19970831.0129" num="10"> NP_1 A police spokesman VP_2 said NP_3 there VP_4 were NP_5 four people PP_6 in NP_7 the the car and NP_8 the accident VP_9 occurred SBAR while NP_10 the princess NP_11 ’s car VP_12 was being pursued PP_13 by NP_14 press photographers PP_15 on NP_16 a motorcycle .</s>

h. <s docid="XIE19970831.0129" num="11"> NP_1 The car , NP_2 a black Mercedes 600 , VP_3 slewed PP_4 into NP_5 a road pillar PP_6 under NP_7 the Alma bridge PP_8 on NP_9 the north bank PP_10 of NP_11 the River Seine and VP_12 smashed PP_13 into NP_14 a wall.</s>

i. <s docid="XIE19970831.0129" num="12"> NP_1 Police VP_2 have detained NP_3 five press photographers PP_4 for NP_5 questioning and VP_6 seized NP_7 two motorcycles and NP_8 a motor scooter PP_9 after NP_10 the crash .</s>

j. <s docid="XIE19970831.0129" num="13"> NP_1 The princess and Al Fayed VP_2 arrived PP_3 in NP_4 Paris PP_5 on NP_6 Saturday afternoon .</s>

k. <s docid="XIE19970831.0129" num="14"> NP_1 Princess Diana VP_2 is NP_3 the mother PP_4 of NP_5 two royal princes , NP_6 William and NP_7 Harry .</s>

Second file :

a. <s docid="XIE19970831.0162" num="4"> PARIS, August 31 (Xinhua)óNP_1 British Princess Diana VP_2 was seriously injured and NP_3 her friend Egyptian millionaire Dodi el-Fayed VP_4 was killed ADVP when NP_5 their car VP_6 crashed PP_7 in NP_8 a Paris road tunnel PP_9 on NP_10 Saturday night , NP_11 Agence France-Presse (NP_12 AFP) VP_13 reported NP_14 today .</s>

b. <s docid="XIE19970831.0162" num="5"> NP_1 The driver VP_2 was also killed , NP_3 the agency VP_4 quoted NP_5 officials PP_6 at NP_7 police headquarters PP_8 as VP_9 saying .</s>

c. <s docid="XIE19970831.0162" num="6"> NP_1 Officials VP_2 said NP_3 the Princess PP_4 of NP_5 Wales , and NP_6 her bodyguard NP_7 who VP_8 was ADVP also ADJP seriously injured , VP_9 had been taken PP_10 to NP_11 a hospital .</s>

d. <s docid="XIE19970831.0162" num="7"> NP_1 The accident VP_2 took NP_3 place PP_4 in NP_5 the tunnel PP_6 under NP_7 the Alma bridge PP_8 on NP_9 the Seine river bank.</s>

e. <s docid="XIE19970831.0162" num="8"> NP_1 Al Fayed and NP_2 the princess VP_3 arrived PP_4 in NP_5 Paris PP_6 on NP_7 Saturday afternoon .</s>

Let's consider that we find some new information. The file must look like that :

Tableau-A VI-1 Table en Annexe.

docid	**num**	**What's new**
XIE19970831.0162	4	VP_2, NP_5, VP_6, PP_9, NP_11, NP_12, VP_13, NP_14
XIE19970831.0162	5	NP_3, VP_4, VP_4, NP_5, PP_6, PP_6
XIE19970831.0162	8	NP_1 NP_2, VP_3, PP_4, PP_6

BIBLIOGRAPHIE

Aksoy, C., Can, F. & Kocberber, S. (2012). Novelty detection for topic tracking. *Journal of the american society for information science and technology*, 63(4), 777–795.

Al-Barrak, M. A. & Al-Razgan, M. (2016). Predicting students final gpa using decision trees : a case study. *International journal of information and education technology*, 6(7), 528–533.

Aristote. (2016). *Méthaphisique*. Arcadia ebook.

Baddeley, A. D. & Hitch, G. (1974). Working memory. *Psychology of learning and motivation*, 8, 47–89.

Bao, J.-P., Shen, J.-y., Liu, X.-d. & Liu, H.-y. (2003, Novembre). Quick asymmetric text similarity measures. *Machine learning and cybernetics, 2003 international conference on*, 1, 374-379.

Barzilay, R. & Elhadad, N. (2003). Sentence aligment for monolingual comparable corpora. *Proceedings of the 2003 conference on empirical methods in natural language processing*, pp. 25–32.

Beamer, B. & Girju, R. (2009). Investigating automatic alignment methods for slide generation from academic papers. *Proceding of the thirteenth conference on computational natural language learning*, pp. 111–119.

Beaudoin-Bégin, A.-M. (2015). *La langue rapaillée. combatre l'insécurité des québecois*. Somme toute.

Bell, A. (1991). *The language of news media*. Oxford : Blackwell.

Blanco, E. & Moldovan, D. (2015). A semantic logic-based approach to determine textual similarity. *Ieee/acm transactions on audio, speech, and language processing*, 23(4), 683–693.

Brin, S., Davis, J. & García-Molina, H. (1995). Copy detection mechanisms for digital documents. *Proceedings of the 1995 acm sigmod international conference on management of data*, (SIGMOD '95), 398–409.

Buckingham, W., Burnham, D., Peter, J., Hill, C., Marcus, W. & Marenbon, J. (2011). *The philosophy book (big ideas simple explained)*. DK.

Burstein, J., Chodorow, M. & Leacock, C. (2004). Automated essay evaluation : The criterion online writing service. *Ai magazine*, 25(3), 27–36.

Bygstad, B. & Munkvold, B. E. (2007). The significance of member validation in qualitative analysis : experiences from a longitudinal case study. *System sciences, 2007. hicss 2007. 40th annual hawaii international conference on*.

Carnie, A. (2015). *Syntax : A generative introduction* (éd. Third). John Wiley& Sons.

Chen, X., Vorvoreanu, M. & Madhavan, K. P. C. (2014). Mining social media data for understanding students' learning experiences. *Learning technologies, ieee transactions*, 7(3), 246–259.

Chomsky, N. (1993). *Lectures on government and binding : The pisa lectures*. Walter de Gruyter.

Chomsky, N. (2002). *Syntactic structures*. Walter de Gruyter.

Chomsky, N. (2005). Three factors in language design. *Linguistic inquiry*, 36(1), 1–22.

Cicchetti, D. V. & Feinstein, A. R. (1990). High agreement but low kappa : Ii. resolving the paradoxes. *Journal of clinical epidemiology*, 43(6), 551–558.

Collins-Thompson, K., Ogilvie, P., Zhang, Y. & Callan, J. (2002). Information filtering, novelty detection, and named-page finding. *Trec*.

Cortes, V. (2004). Lexical bundles in published and student disciplinary writing : Examples from history and biology. *English for specific purposes*, 23(4), 397–423.

Creswell, J. W. (2013). *Qualitative inquiry and research design : Choosing among five approaches* (éd. Third). SAGE.

D'Alessio, D. & Allen, M. (2000). Media bias in presidential elections : A meta-analysis. *Journal of communication*, 50(4), 133–156.

Dascalu, M., Trausan-Matu, S., McNamara, D. S. & Dessus, P. (2015). Readerbench : Automated evaluation of collaboration based on cohesion and dialogism. *International journal of computer-supported collaborative learning*, 10(4), 395–423.

Davidson, D. (2001). *Essays on actions and events : Philosophical essays*. Oxford University Press.

De Jong, F. (2015). *Understanding the difference : Responsive education : A search for a difference which makes a difference for transition, learning and education*. STOAS Wageningen, The Netherlands.

Di Sciullo, A. M. (2005). *Asymmetry in morphology*. MIT Press.

Di Sciullo, A. M. (2013). A reason to optimize information processing with a core property of natural language. *Intelligent software methodologies, tools and techniques (somet), 2013 ieee 12th international conference on*, pp. 21–28.

Di Sciullo, A. M. (2016). On the domain specificity of the human language faculty and the effects of principles of computational efficiency : Contrasting language and mathematics. *Revista linguistica*, 11(1).

Dice, L. R. (1945). Measures of the amount of ecologic association between species. *Ecology*, 26(3), 297–302.

Dkaki, T., Mothe, J. & Augé, J. (2002). Novelty track at irit-sig. *Trec*.

Dubé, L. & Paré, G. (2003). Rigor in information systems positivist case research : current practices, trends, and recommendation. *Mis quarterly*, 27(4), 597–635.

Durrant, P. (2014). Discipline and level specificity in university students' written vocabulary. *Applied linguistics*, 35(3), 328–356.

Elliot, S. (2003). Intellimetric : From here to validity. *Automated essay scoring : A cross-disciplinary perspective*, 71–86.

Erkan, G. & Radev, D. R. (2004). Lexrank : graph-based lexical centrality as salience in text summarization. *Journal of artificial intelligence research*, 457–479.

Feinstein, A. R. & Cicchetti, D. V. (1990). High agreement but low kappa : I. the problems of two paradoxes. *Journal of clinical epidemiology*, 43(6), 543–549.

Ferreira, R., Dueire Lins, R., J. Simske, S., Freitas, F. & Riss, M. (2016). Assessing sentence similarity through lexical, syntactic and semantic analysis. *Computer speech and language*, 39(C), 1–28.

Figueira, A. (2016). Predicting grades by principal component analysis : A data mining approach to learning analyics. *Advanced learning technologies (icalt), 2016 ieee 16th international conference on*, pp. 465–467.

Fillmore, C. J. (1967). The case for the case. *Universals in linguistic theory*, 1–88.

Forster, M. R. (2002). Predictive accuracy as an achievable goal of science. *Philosophy of science*, 69(3), S124–S134.

Gardner, D. & Davies, M. (2013). A new academic vocabulary list. *Applied linguistics*, 35(3), 305–327.

Ghoniem, M., Fekete, J.-D. & Castagliola, P. (2004). A comparison of the readability of graphs using node-link and matrix-based representations. *Iee symposium on information visualization (infovis)*, pp. 17–24.

Gull, K. (2016). WTF IS ART ? PLATO'S REFLECTIONS ON BEAUTY AND LOVE. Repéré à https://www.visualnews.com/2016/05/11/wtf-art-platos-reflections-beauty-love/.

Halliday, M. & Hassan, R. (1976). *Cohesion in english*. London : Longman.

Halliday, M. A. K. (1978). *Language as social semiotic*. London Arnold.

Harman, D. (2002). Overview of the trec 2002 novelty track. *Proceedings of the 10th text retrieval conference (trec 2004)*.

Hitchcock, C. & Sober, E. (2004). Prediction versus accommodation and the risk of overfitting. *The british journal for the philosophy of science*, 55(1), 1–34.

Huff, A. S. (2009). *Designing research for publication*. SAGE.

Iacobelli, F., Birnbaum, L. & Hammond, K. J. (2010a). Tell me more, not just more of the same. *Proceedings of the 15th international conference on intelligent user interfaces*, pp. 81–90.

Iacobelli, F., Nichols, N. D., Birnbaum, L. & Hammond, K. J. (2010b). Finding new information via robust entity detection. *Fall symposium : Proactive assitant agents*.

Jackendoff, R. S. (1992). *Semantic structures*. MTI press.

Jain, G. P., Gurupur, V. P., Schroeder, J. L. & Faulkenberry, E. D. (2014). Artificial intelligence-based student learning evaluation : a concept map-based approach for analyzing a student's understanding of a topic. *Learning technologies, ieee transactions on*, 7(3), 267–279.

Jakobson, R. (1972). Linguistics and poetics. *Style in language*, 350-377.

Jiang, J. J. & Conrath, D. W. (1997). Semantic similarity based on corpus statistics and lexical taxonomy. *Proceedings on international conference on research in computational linguistics*, pp. 19–33.

Julinda, S., Boden, C. & Akbik, A. (2014). Extracting a repository of events and event references from news clusters. *Proceedings of the first aha !-workshop on information discovery in text*, pp. 14–18.

Kalz, M., Van Bruggen, J., Giesbers, B., Waterink, W., Eshuis, J. & Koper, R. (2014). A study about placement support using semantic similarity. *Educational technology and society*, 17(3), 54-64. JSTOR.

Karkali, M., Rousseau, F., Ntoulas, A. & Vazirgiannis, M. (2013). Efficient online novelty detection in news streams. *International conference on web information systems engineering*, pp. 57–71.

Kayne, R. S. (1994). *The antisymmetry of syntax*. MIT Press.

Kessler, R., Tannier, X., Hagège, C., Moriceau, V. & Bittar, A. (2012). Extraction de dates saillantes pour la construction de chronologies thématiques. *Revue traitement automatique des langues*, 53(2), 57–86.

Landauer, T. K., Laham, D. & Foltz, P. W. (2003). *Automated scoring and annotation of essays with the intelligent essay assessor*. Lawrence Erlbaum.

Leacock, C. & Chodorow, M. (1998). Combining local context and wordnet similarity for word sense identification. *Wordnet : An electronic lexical database*, 49(2), 265–283.

Leyton, M. (1992). *Symmetry, causality, mind*. MIT Press.

Lin, D. (1998). An information-theoretic definition of similarity. *International conference on machine learning*, 98, 296–304.

Ma, W. J., Husain, M. & Bays, P. M. (2014). Changing concepts of working memory. *Nature neuroscience*, 17(3), 347–356.

Manning, C. D. & Hirich, S. (1999). *Foundations of statistical natural language processing*. MIT Press.

Martinez, R. & Schmitt, N. (2012). A phrasal expressions list. *Applied linguistics*, 33(3), 299–320.

McHugh, M. L. (2012). Interrater reliability : the kappa statistic. *Biochemia medica*, 22(3), 276–282.

Mihalcea, R., Corley, C. & Strapparava, C. (2006). Corpus-based and knowledge-based measures of text semantic similarity. *Aaai*, pp. 775-780.

Miller, G. A. (1956). The magical number seven, plus or minus two : Some limits on our capacity for processing information. *Phychcological review*, 63(2), 343–352.

Miller, G. A. (1995). Wordnet : a lexical database for english. *Communications of the acm*, 38(11), 39–41.

Mitchell, W. J. (1990). *The logic of architecture : Design, computation, and congnition*. MIT Press.

Morse, J. M. (1998). Validity by commitee. *Qualitative health research*, 8, 443–445.

Mueen, A., Zafar, B. & Manzoor, U. (2016). Modeling and predicting students' academic performance using data mining techniques. *International journal of modern education & computer science*, 8(11), 36–42.

Nelken, R. & Shieber, S. M. (2006). Towards robust context-sensitive sentence alignment for monolingual corpora. *proceedings eacl*, pp. 161–166.

of Encyclopædia Britannica, T. E. (2016). Acta, ANCIENT ROMAN PUBLICATION. Repéré à https://www.britannica.com/topic/Acta.

Palmer, F. R. (1994). *Grammatical roles and relations*. Cambridge.

Palmer, M., Gildea, D. & Kingsbury, P. (2005). The proposition bank : An annotated corpus of semantic roles. *Computational linguistics*, 31(1), 71–106.

Park, S., Kang, S., Chung, S. & Song, J. (2009). Newscube : delivering multiple aspects of news to mitigate media bias. *Proceedings of the special interest group on computer-human interaction, conference on human factors in computing systems*, pp. 443–452.

Park, S., Lee, S. & Song, J. (2010). Aspect-level news browsing : understanding news events from multiple viewpoints. *Proceedings of the 15th international conference on intelligent user interfaces*, pp. 41–50.

Pedersen, T., Patwardhan, S. & Michelizzi, J. (2004). Wordnet : : Similarity : measuring the relatedness of concepts. *Demonstration papers at hlt-naacl 2004*, pp. 38–41.

Power, D. M. W. (2003). Recall and precision versus the bookmarker. *Proceedings of the international conference on cognitive science (icsc-2003)*, pp. 529–534.

Power, D. M. W. (2012). The problem of kappa. *Proceedings of the 13th conference of the european chapter of the association for computational linguistics*, pp. 345–355.

Pustejovsky, J. (1991). The syntax of event structure. *Cognitiontion*, 41(1), 47–81.

Pustejovsky, J. & Stubbs, A. (2012). *Natural language annotation for machine learning*. O'Reilly Media, Inc.

Ratté, S. (1995). *Interprétations des structures syntaxiques : une analyse computationelle de la structure des événements*. (Thèse de doctorat, Université du Québec à Montréal).

Reinhart, T. M. (1976). *The syntactic domain of anaphora*. (Thèse de doctorat, Massachusetts Institute of Technology).

René, T. (1993). *Prédire n'est pas expliquer*. Paris : Champs Flammarion.

Rescher, N. (1958). On prediction and explanation. *The british journal for the philosophy of science*, 8(32), 281–290.

Resnik, P. (1995). Using information content to evaluate semantic similarity in a taxonomy. *Proceedings of the 14th international joint conference on artificial intelligence*, pp. 448–453.

Richardson, J. E. (2006). *Analysing newspapers : An approach from critical discourse analysis*. Palgrave MacMillan.

Rosen, S. T. (1999). The syntactic representation of linguistic events. *Glot international*, 4(2), 3–11.

Roth, M. (2014). *Inducing implicit arguments via cross-document alignment : A framework and its applications*. (Thèse de doctorat, Institut fur Computerlinguistik, Ruprecht-Karls-Universitat Heidelberg).

Ruppenhofer, J., Ellsworth, M., Petruck, M. R., Johnson, C. R. & Scheffczyk, J. (2006). FrameNet II : Extended theory and practice.

Saeed, I. J. (2013). *Semantics* (éd. Third). Wiley-Blackwell.

Saez-Trumper, D., Castillo, C. & Lalmas, M. (2013). Social media news communities : gatekeeping, coverage, and statement bias. *Proceedings of the 22nd acm international conference on conference on information & knowledge management*, pp. 1679–1684.

Saurí, R. & Pustejovsky, J. (2012). Are you sure that this happened ? assessing the factuality degree of events in text. *Computational linguistics*, 38(2), 261–299.

Scheihing, E., Vernier, M., Born, J., Guerra, J. & Carcamo, L. (2016). Classifying discourse in a cscl platform to evaluate correlations with teacher participation and progress. *arxiv preprint arxiv :1605.07268*, 1–9.

Schiffman, B. (2002). Experiments in novelty detection at columbia university. *Trec*.

Schleppegrell, M. J. (2007). The linguistic challenges of mathematics teaching and learning : A research review. *Reading & writing quarterly*, 23(2), 139–159.

Shiffrin, R. M. & Nosofsky, R. M. (1994). Seven plus or minus two : A commentary on capacity limitations. *Phychcological review*, 101(2), 357–361.

Shivakumar, N. & Garcia-Molina, H. (1995). Scam : A copy detection mechanism for digital documents. *2nd international conference in theory and practice of digital libraries (dl 1995)*.

Shmueli, G. (2010). To explain or to predict. *Statistical science*, 25(3), 289–310.

Soboroff, I. (2004). Overview of the trec 2004 novelty track. *Proceedings of the 12th text retrieval conference (trec 2002)*.

Soboroff, I. & Harman, D. (2003). Overview of the trec 2003 novelty track. *Proceedings of the 11th text retrieval conference (trec 2003)*, pp. 38–53.

Soboroff, I. & Harman, D. (2005). Novelty detection : the trec experience. *Proceedings of the conference on human language technology and empirical methods in natural language processing*, pp. 105-112.

Sorour, S. E., Mine, T., Goda, K. & Hirokawa, S. (2014). Predicting students' grades based on free style comments data by artificial neural network. *Ieee frontiers in education conference*, pp. 1–9.

Speed, J. G. (1893). Do newspaper now give the news. *Forum*, 15, 705–711.

Steinberger, R. (2012). A survey of methods to ease the development of highly multilingual text mining applications. *Language resources and evaluation*, 46(2), 155–176. Springer.

Tenny, C. & Pustejovsky, J. (2000). A history of events in linguistic theory. *Event as grammatical objects*, 3–37.

Tenny, C. & Pustejovsky, J. (2001). *Events as grammatical objects the converging perspectives of lexical semantics and syntax*. Center for the Study of Language and Inf (April 1 2001).

Tenny, C. L. (2000). Core events and adverbial modification. *Event as grammatical objects : The covering perspectives of lexical semantics and syntax*, 285–329.

Tobin, R. (1975). The canon of polykleitos. *American journal of archaeology*, 79(4), 307–321.

Tsai, M.-F. & Chen, H.-H. (2002). Some similarity computation methods in novelty detection. *Trec*.

Turney, P. D. (2012). Domain and function : A dual-space model of semantic relations and compositions. *Journal of artificial intelligence research*, 44, 533–585.

Tversky, A. (1977). Features of similarity. *Phychcological review*, 84(4), 327–352.

Tversky, A. & Gati, I. (1978). Studies of similarity. *Cognition and categorization*, 1(1978), 79–98.

Van Hage, W. R., Malaisé, V., Segers, R., Hollink, L. & Schreiber, G. (2011). Design and use of the simple event model (sem). *Web semantics : Science, services and agents on the world wide web*, 9(2), 128–136.

Velazquez Godinez, E. (2012). *Des techniuqes d'interaction bimanuelles pour la manipulation de réseaux*. (Mémoire de maîtrise, École de technologie supérieure).

Vitruvius, P. (2009). *De architectura : Ten books on architecture*. Digireads.com.

Von Luxburg, U. (2007). A tutorial on spectral clustering. *Statistics and computing*, 17(4), 395–416.

Warschauer, M. & Ware, P. (2006). Automated writing evaluation : Defining the classroom research agenda. *Language teaching research*, 10(2), 157–180.

Weber, J. (2006). Strassburg, 1605 : The origins of the newspaper in europe. *German history*, 24(3), 387–412.

Wu, Z. & Palmer, M. (1994). Verbs semantics and lexical selection. *Proceedings of the 32nd annual meeting on association for computational linguistics*.

Zhang, Y., Callan, J. & Minka, T. (2002). Novelty and redundancy detection in adaptative filtering. *roceedings of the 25th annual international acm sigir conference on research and development in information retrieval*, pp. 81–88.