UNIVERSITÉ D'AVIGNON
ET DES PAYS DE VAUCLUSE
MINISTÈRE DE L'ENSEIGNEMENT
SUPÉRIEUR ET DE LA RECHERCHE

ACADÉMIE D'AIX-MARSEILLE
UNIVERSITÉ D'AVIGNON ET DES PAYS DE VAUCLUSE

THÈSE

présentée à l'Université d'Avignon et des Pays de Vaucluse
pour obtenir le diplôme de DOCTORAT

SPÉCIALITÉ : Informatique

École Doctorale 536 « *Sciences et Agrosciences* »
Laboratoire Informatique d'Avignon (EA 931)

Compression automatique de phrases : une étude vers la génération de résumés

par
Alejandro MOLINA VILLEGAS

Soutenue publiquement le 30 septembre 2013 devant un jury composé de :

Horacio SAGGION	Universitat Pompeu Fabra	Rapporteur
Eric GAUSSIER	Université de Grenoble	Rapporteur
Guy LAPALME	Université de Montréal	Examinateur
Josiane MOTHE	Institut de Recherche en Informatique de Toulouse	Examinateur
Juan-Manuel TORRES-MORENO	Université d'Avignon et des Pays de Vaucluse	Directeur
Eric SANJUAN	Université d'Avignon et des Pays de Vaucluse	Co-directeur
Gerardo Eugenio SIERRA	Universidad Nacional Autónoma de México	Co-encadrant

---

Laboratoire Informatique d'Avignon
Consejo Nacional de Ciencia y Tecnología

Résumé

Cette étude présente une nouvelle approche pour la génération automatique de résumés, un des principaux défis du Traitement de la Langue Naturelle. Ce sujet, traité pendant un demi-siècle par la recherche, reste encore actuel car personne n'a encore réussi à créer automatiquement des résumés comparables, en qualité, avec ceux produits par des humains. C'est dans ce contexte que la recherche en résumé automatique s'est divisée en deux grandes catégories : le résumé par extraction et le résumé par abstraction. Dans le premier, les phrases sont triées de façon à ce que les meilleures conforment le résumé final. Or, les phrases sélectionnées pour le résumé portent souvent des informations secondaires, une analyse plus fine s'avère nécessaire.

Nous proposons une méthode de compression automatique de phrases basée sur l'élimination de fragments à l'intérieur de celles-ci. À partir d'un corpus annoté, nous avons créé un modèle linéaire pour prédire la suppression de ces fragments en fonction de caractéristiques simples. Notre méthode prend en compte trois principes : celui de la pertinence du contenu, l'informativité ; celui de la qualité du contenu, la grammaticalité, et la longueur, le taux de compression. Pour mesurer l'informativité des fragments, nous utilisons une technique inspirée de la physique statistique : l'énergie textuelle. Quant à la grammaticalité, nous proposons d'utiliser des modèles de langage probabilistes. La méthode proposée est capable de générer des résumés corrects en espagnol.

Les résultats de cette étude soulèvent divers aspects intéressants vis-à- vis du résumé de textes par compression de phrases. On a observé qu'en général il y a un haut degré de subjectivité de la tâche. Il n'y a pas de compression optimale unique mais plusieurs compressions correctes possibles. Nous considérons donc que les résultats de cette étude ouvrent la discussion par rapport à la subjectivité de l'informativité et son influence pour le résumé automatique.

Table des matières

Liste des illustrations

Liste des tableaux

Nomenclature

Conventions typographiques utilisées dans cette thèse.

φ	Une phrase.
$\widetilde{\varphi}$	Une phrase compressée.
$\widetilde{\varphi}_i$	Le candidat i à la compression.
$\widetilde{\varphi^\star}_i$	Le meilleur candidat à la compression.
τ	Le taux de compression d'une phrase.
w_i	Un mot.
$(w_1, w_2, \ldots, w_n)$	Une séquence de n mots.
$w_i^j = (w_i, \ldots, w_j)$	Une sous-séquence du mot w_i à w_j.
s_i	Un segment.
$(s_1, s_2, \ldots, s_n)$	Une séquence de n segments.
$s_i^j = (s_i, \ldots, s_j)$	Une sous-séquence du segment s_i à s_j.
$()$	Une séquence vide.
T	Le nombre de termes uniques d'un document (vocabulaire).
Φ	Le nombre de phrases dans un document.
$\mathbb{A} = (a_{i,j})$	Une matrice $\mathbb{A}$.
$\mathbb{A}^{\mathsf{t}}$	La transposée de $\mathbb{A}$.
$\mathbb{E}$	La matrice d'énergie textuelle.
$\mathbf{tf}(\varphi_i, w_j)$	Le nombre d'occurrences du mot w_j dans la phrase φ_i.
$\mathbf{P}(\bullet)$	La probabilité d'un événement.
$\mathbf{P}(A\|B)$	La probabilité conditionnelle d'un événement A, sachant B.
$\mathbf{log}(\bullet)$	La fonction logarithme.
$\mathbf{Seg}(\bullet)$	Une fonction qui compte le nombre de segments.
$\mathbf{Lon}(\bullet)$	Une fonction qui compte le nombre de mots.
$\mathbf{Info}(\bullet)$	Une fonction qui mesure l'informativité.
$\mathbf{Gram}(\bullet)$	Une fonction qui mesure la grammaticalité.
$\mathbf{Ener}(\bullet)$	Une fonction qui mesure l'énergie textuelle.
$\mathbf{LP}(w_1^n)$	Une fonction qui mesure la probabilité d'une séquence.
u_i	Une unité dans un réseau de Hopfield.
$\mathbf{poids}(u_i, u_j)$	Degré d'importance entre u_i et u_j.
α	Une valeur de seuil de probabilité.
β_i	Un paramètre dans un modèle de régression linéaire.

Introduction

C'est dans le cadre de la guerre froide qu'il faut chercher les origines du domaine du résumé automatique de documents. En effet, la compétition entre les États-Unis et l'URSS s'étendait à tous les domaines scientifiques et techniques. Les systèmes de traitement d'information n'auraient donc pas été considérés suffisamment importants sans ce phénomène. Après le lancement du «*Sputnik*», les financements dédiés aux sciences, y compris celles de l'information, sont devenus une affaire importante pour les deux superpuissances. C'est à la fin des années 50, que Malcom Dyson, directeur de recherche au *Chemical Abstracts Service* a demandé à Hans Peter Luhn, ingénieur de recherche à IBM, de collaborer dans diverses problématiques concernant la transformation de textes, l'auto-indexation et le résumé automatique de documents liés à la chimie. Le domaine de résumé automatique de documents venait de naître [1]. Un des premiers systèmes de résumé automatique de documents est décrit dans (Luhn, 1958). Dans ce travail, l'auteur montre que l'importance d'une phrase est liée aux fréquences des mots qu'elle contient. Plus la phrase possède de mots importants, plus elle est importante pour le document. Les mots sont donc classés selon leur fréquence : plus un mot est répété, plus il est important, à condition d'ignorer des mots d'usage très commun et ceux d'usage très rare. Dans la figure 1, tirée de l'article (Luhn, 1958), les mots à gauche de *C* qui appartiennent à la région de plus haute fréquence sont des mots d'usage très commun ; tandis que les mots à droite de *D* sont d'usage très rare. La région entre *C* et *D* est considérée comme celle qui contient les mots significatifs. De plus, l'auteur soutient aussi que la position relative d'un mot dans une phrase détermine son importance. Selon l'hypothèse de Luhn, des idées proches, matérialisées en mots, ont tendance à se retrouver proches aussi dans un document (Luhn, 1957). Aussi chez IBM, Phyllis Baxendale a fait des remarques essentielles pour le domaine du résumé automatique. Elle constate que la position relative d'une phrase dans un document est importante (Baxendale, 1958). Dans 85% des cas, la phrase-sujet (*topic sentence*) est justement la première d'un document et seulement dans 7% des cas elle correspond à la dernière. Baxendale a par ailleurs mené des expériences en comparant la performance entre les humains et les ordinateurs pour l'extraction de mots-clès, un des protocoles d'évaluation utilisés dans le domaine. Edmunson a utilisé les caractéristiques explorées par Baxendale et Luhn ainsi que d'autres caractéristiques, dans une approche d'ap-

1. Pour en savoir plus par rapport à l'histoire du résumé automatique, l'étude de (Rayward et Bowden, 2004) est très extensive et celle de (Mani et Maybury, 1999) lui consacre une partie importante dans son introduction.

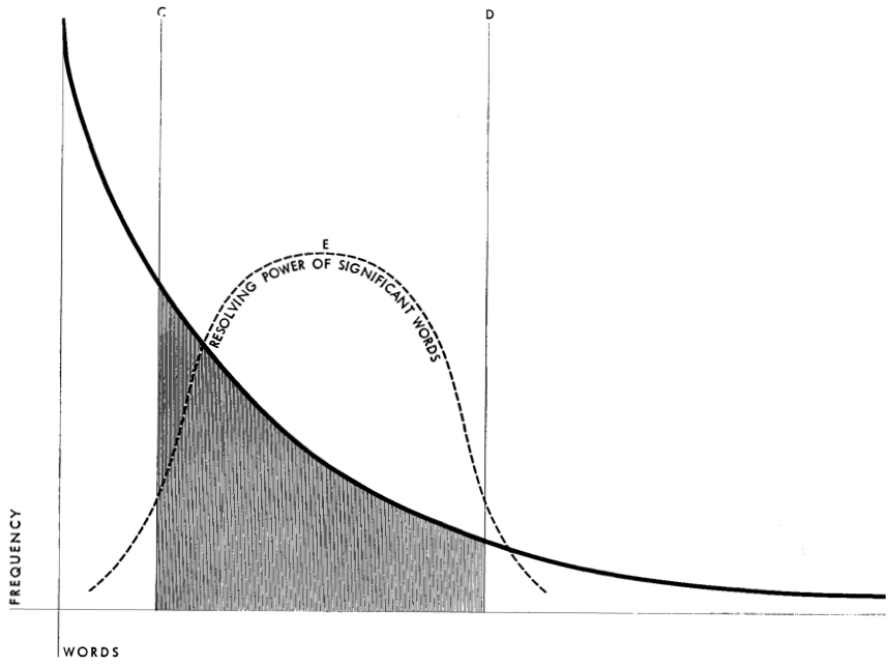


FIGURE 1 – *Diagramme de fréquences des mots dans un document : les abscisses sont des mots triés par leur fréquences.*

prentissage semi-supervisé sur un corpus de textes scientifiques (Edmundson, 1969). La façon d'estimer le score d'une phrase s est donc basée sur une fonction linéaire qui combine des caractéristiques des phrases et dont les poids ont été ajustés manuellement selon les meilleurs résultats.

Il est remarquable que les idées de ces trois chercheurs fassent encore partie des fondements du résumé automatique actuel. En faisant référence aux travaux de ces trois piliers du domaine, nous avons déjà mentionné certains concepts utilisés dans la présente étude. La façon de mesurer l'importance des mots dépend partiellement de leurs fréquences. Nous utilisons aussi une fonction linéaire qui combine plusieurs caractéristiques afin de savoir s'il est possible d'éliminer un fragment d'une phrase. Finalement, nous avons réalisé des évaluations en comparant les résumés produits manuellement et automatiquement.

Mais, comment est-ce qu'un problème datant des années cinquante est devenu un domaine de recherche pointu ? Ce demi-siècle de recherche a conçu une grande variété de méthodes de résumé automatique afin d'améliorer les résultats des méthodes de base. Néanmoins, en raison de la complexité implicite dans la tâche, personne n'a encore réussi à produire des résumés « *parfaits* » de façon automatique. Ces résumés seraient comparables à ceux produits par des professionnels dans n'importe quelles conditions : indépendants du nombre de documents sources, indépendants de la longueur des documents, indépendants de la langue, indépendants de la thématique, indépendants du registre de la langue et de qualité irréprochable. C'est dans ce contexte que la recherche en résumé automatique s'est divisée en plusieurs directions : selon le

but pour lequel le résumé est produit (résumé indicatif et résumé informatif) ; selon le récepteur du résumé final (résumé générique et résumé orienté) ; selon le nombre de documents qui produisent le résumé (résumé monodocument et résumé multidocuments) ; selon la ou les langues des documents sources (résumé monolingue et résumé multilingue) ; selon le genre du document source (scientifique, journalistique, encyclopédique). Chaque direction se consacre à des aspects ou des objectifs plus précis, afin d'arriver à une compréhension graduelle du domaine.

Les méthodes de résumé peuvent donc être classées selon le type de résumé qu'elles produisent en deux grandes catégories : le résumé par extraction et le résumé par abstraction. Dans le premier, le document source est divisé en phrases. Puis, chacune est scorée de façon à ce que les meilleures composent le résumé final. Un des désavantages majeurs de cette méthode est que les phrases qui sont incluses dans le résumé final ne sont pas modifiées. Cela implique la possibilité d'avoir des informations non essentielles dans le résumé simplement parce qu'elles font partie d'une phrase avec un bon score. Dans le résumé par abstraction, on ne se contente pas de sélectionner des phrases, on cherche plutôt à produire une représentation abrégée et exacte du contenu d'un document comme s'il avait été crée par son auteur (Torres-Moreno, 2012). Contrairement au résumé par extraction, cette représentation peut contenir des figures de style telles que la métaphore, la paraphrase et la gradation. Bref, c'est ce que tout le monde comprend par le terme *résumé* en français mais nous sommes obligés de faire cette distinction pour le différencier de résumé par extraction[2]. Néanmoins, produire un résumé par abstraction, de manière automatique, reste encore un défi très ambitieux. Selon (Boudin, 2008), la construction de la représentation sémantique d'un document est un travail qui nécessite des modèles conceptuels, des ressources linguistiques et des outils informatiques qui, pour la plupart d'entre eux, n'ont pas atteint la maturité nécessaire à une utilisation robuste. C'est la raison pour laquelle, de nos jours, le résumé par extraction est l'approche la plus largement utilisée. L'article de Radev et al. (2002) donne un panorama exhaustif des toutes les méthodes de l'état de l'art en résumé automatique jusqu'à 2002 et, plus récemment, les travaux de (Spärck-Jones, 2007; Das et Martins, 2007; Ježek et Steinberger, 2008).

Cette étude vise concrètement à résoudre une des majeures problématiques du résumé par extraction et, en même temps, établit un pont vers le résumé par abstraction. La problématique provient du fait qu'en sélectionnant une phrase par extraction, les informations superflues contenues dans la phrase seront forcément ajoutées au résumé. Ainsi, le but de la méthode de compression est d'éliminer ce qui n'est pas important. Notre méthode segmente les phrases « *finement* » afin d'éliminer des « *segments* » à condition que la version compressée soit grammaticalement correcte. Le tableau 2 présente un résumé produit par notre méthode à partir d'un document de 375 mots dont la deuxième phrase était à l'origine : « *Leurs expériences, relatées dans l'édition du 9 avril du magazine Science, ouvrent la voie à des composants capables de fonctionner à des fréquences 10 à 100 fois plus élevées que celles des puces actuelles qui sont bridées par des problèmes de dissipation de chaleur.* »

2. En anglais, cette différence est explicite grâce aux termes *summary* et *abstract*.

Le pont que nous trouvons entre le résumé par extraction et le résumé par abstraction s'établit moyennant l'abrègement du texte implicite dans la méthode par compression. En effet, si les phrases du texte d'origine sont modifiées, la méthode est au-delà des méthodes classiques par extraction. Les auteurs de (Jing et McKeown, 1999) ont montré que les deux opérations les plus importantes réalisées par les humains lors de la génération de résumés sont la compression de phrases et la combinaison de phrases. Cette étude s'est focalisée sur la compression de phrases et le résumé automatique.

Nous avons adopté le terme «*compression de phrases*» en faisant référence à la tâche connue, dans de nombreux articles, comme *sentence compression* car c'était le plus utilisé au début de ce projet. Il est possible de trouver, dans la littérature, d'autres termes tels que *sentence simplification*, *sentence summarization*, *sentence reduction*, *subtitle generation*, *text compression* et *text reduction*, pour des travaux plus ou moins liés à la même thématique.

Hormis quelques exemples, nous basons nos expériences sur des textes en espagnol. Cette langue est au deuxième rang mondial en termes de locuteurs natifs (juste après le mandarin) et aussi la deuxième langue parlée aux États-Unis. Selon Wikipédia, 406 millions de personnes ont l'espagnol comme langue maternelle ; 466 millions si l'on compte les pratiquants secondaires et 500 millions avec les pratiquants comme langue étrangère. Néanmoins, les méthodes ici décrites sont majoritairement statistiques, ce qui rend possible la reproduction des résultats dans d'autres langues avec des résultats similaires attendus.

Et si l'ordinateur pouvait fonctionner un jour, sans électricité ? (résumé)

La démarche de chercheurs américains de l'université de Notre Dame, montre que l'on peut manipuler des électrons pour construire des circuits élémentaires avec des quantités d'énergie infimes. Leurs expériences, ouvrent la voie à des composants capables de fonctionner à des fréquences 10 à 100 fois plus élevées.

TABLE 2 – *Exemple de résumé produit par notre méthode à partir d'un document avec 375 mots.*

Plan de la thèse

La structure de cette thèse est liée à l'évolution de notre recherche. Il y a six chapitres, chacun correspondant aux défis soulevés dans les résultats de nos expériences. Chaque chapitre contient des sections dédiées à l'état de l'art selon la pertinence.

Le chapitre 1 est consacré à la relation entre la compression de phrases et le résumé automatique par extraction. Nous citons d'abord les études qui ont donné naissance au concept de la compression de phrases. Ensuite, nous présentons nos premières recherches à ce sujet qui ont orienté la suite de notre travail.

Dans le chapitre 2 nous étudions l'apport de la segmentation discursive à notre méthodes de compression de phrases. Nous réalisons les premières expériences avec un segmenteur discursif pour l'espagnol, ce qui nous a permis de l'adapter pour la tâche qui nous intéresse. Grâce à ces expériences nous envisageons la possibilité d'un segmenteur discursif multilangue.

Dans le chapitre 3 nous abordons la problématique liée à la détermination pertinente de la grammaticalité d'une phrase compressée. Pour faire face à ce défi, nous proposons l'utilisation d'un modèle de langage probabiliste.

Le chapitre 4, explore l'évaluation de l'informativité de phrases compressées. Pour mesurer l'importance du contenu des segments discursifs, nous utilisons le modèle de l'énergie textuelle. Nous proposons de même une amélioration de celui-ci.

Dans le chapitre 5, nous présentons notre modèle de résumé automatique par compression de phrases. Celui-ci consiste en plusieurs variables liées à l'énergie textuelle, au modèle de langage probabiliste et à la segmentation discursive. Grâce à notre modèle, nous sommes capables de générer des résumés corrects mais nous constatons aussi que la tâche est fortement subjective.

Finalement, dans le chapitre 6 nous évaluons nos résumés avec un test de Turing. Nous remarquons que parmi 54 juges, seulement une personne présente des résultats statistiquement significatifs pour distinguer les résumés manuels et les résumés automatiques.

Après nos conclusions, nous ajoutons plusieurs annexes contenant des données expérimentales issues de nos recherches.

Chapitre 1

La compression de phrases : le résumé automatique au-delà de l'extraction

Sommaire

Dans ce chapitre nous abordons la compression de phrases. Pour ce faire, nous commençons par survoler les études antérieures à la définition formelle de cette tâche mais qui sont liées à celle-ci. La réflexion sur cette relation nous permet de prendre connaissance des problématiques inhérentes à la compression de phrases et d'imaginer ses possibles applications. Par la suite, nous présentons le travail pionnier de Knight et Marcu qui ont élaboré la première définition formelle de la tâche qui nous intéresse. D'après cette définition, pour compresser une phrase il faudrait éliminer les mots qui ne sont pas nécessaires. Dans nos premières expériences, nous avons pourtant demandé aux annotateurs d'éliminer librement des fragments textuels plutôt que des mots isolés. Nous détaillons la manière dont les résultats de ce premier travail ont orienté la suite de nos recherches.

1.1 Premières études liées à la compression de phrases

Les études de (Grefenstette, 1998) représentent les premiers essais pour l'automatisation de la compression textuelle au niveau de la phrase plutôt qu'au niveau du document entier. Ainsi, il a tenté d'identifier les passages les plus importants à l'intérieur de la phrase au lieu de trier les phrases selon leur degré d'importance. Grefenstette a présenté une méthode de réduction de textes (des « *versions télégraphiques* ») qui avait pour objectif de réduire le temps de lecture d'un système de synthèse texte-parole pour les mal-voyants. La méthode cherchait à repérer les éléments les plus importants, sans prendre en considération la grammaticalité. Pour ce faire, il a suivi les principes proposés par (Jing et Croft, 1994), selon lesquels :

- les noms propres sont plus importants que les noms communs ;
- les substantifs sont plus importants que les adjectifs ;
- les adjectifs sont plus importants que les articles ;
- la négation est toujours importante.

Le texte a été annoté avec un analyseur syntaxique commercial et par la suite un niveau de compression a été appliqué. Huit niveaux de compression ont été déterminés par l'ordre d'application des principes cités. Chaque niveau filtre les parties qui doivent être supprimées. Il n'y a pas eu d'évaluation, mais le système et ses applications ont été innovateurs à leur époque.

Plus tard (Witbrock et Mittal, 1999), ont présenté une méthode de génération non-extractive de résumé capable de générer des titres de taille variable. Les auteurs mentionnent que la compression de phrases peut aider à la génération automatique de titres. Par exemple, les agences de nouvelles reçoivent quotidiennement une grande quantité de textes qui doivent être titrés rapidement et de façon adéquate. Le système proposé dispose d'une étape d'apprentissage supervisé qui essaie de capturer, à la fois les règles de sélection des contenus importants ainsi que les règles de la réalisation du titre. Le modèle des règles de sélection a été basé sur la probabilité d'un mot de figurer dans le résumé étant donné son apparition dans le texte. La production du résumé final a consisté en une chaîne de Markov de premier ordre.

Dans un travail plus complet, (Jing et McKeown, 1999) ont montré que les deux opérations les plus importantes réalisées par les humains lors de la génération de résumés sont la réduction et la combinaison de phrases. En effet, les annotateurs choisissent d'abord les parties les plus pertinentes du texte puis, grâce à un travail d'édition, ils les incorporent sous de nouvelles formes dans le résumé. Les auteurs appellent cette technique couper-et-coller (*cut-and-paste*). Ils ont visé à reconstituer, en quelque sorte, le processus cognitif effectué par l'être humain. De plus, ils ont mené des expériences pour répondre aux questions suivantes : étant donnée une phrase dans un résumé, a-t-elle été construite avec la technique couper-et-coller à partir du texte original ? Si oui, quels éléments de la phrase d'origine sont repris à l'identique ? Et de quelles parties du document d'origine proviennent-ils ? Les auteurs ont analysé 120 phrases provenant de 15 résumés et ils ont identifié six opérations couper-et-coller :

– la réduction de phrases ;
– la combinaison de phrases ;
– la transformation syntaxique ;
– la paraphrase lexicale ;
– la généralisation et/ou la spécification ;
– le reclassement.

Les résultats montrent qu'une proportion significative des phrases des résumés (78%) est générée par au moins l'une de ces opérations. Cependant, les deux opérations principales sont la réduction de phrases et leur combinaison. Les deux conclusions les plus intéressantes sont que les humains ont tendance à réduire les phrases en effaçant des passages textuels et rarement des mots isolés et qu'ils ont tendance à combiner des phrases plutôt proches.

(Barzilay et al., 1999) ont présenté une méthode de résumé multidocument où des phrases sont sélectionnées grâce à un résumeur extractif élémentaire. Puis, celles-ci sont divisées en morceaux (*chunks*) qui sont traités à leur tour par divers processus :

1. Sélection d'éléments.
2. Détection de paraphrases.
3. Ordre chronologique.
4. Génération de phrases.

La dernière étape combine les morceaux pour générer une seule phrase. L'objectif principal de cette méthode est de générer des résumés concis, tout en synthétisant les éléments similaires d'un ensemble de phrases extraites. Certaines caractéristiques des phrases sont envoyées à l'entrée d'un générateur de phrases. Ensuite, un système sémantique très sophistiqué révise les morceaux afin de repérer leur position dans la phrase finale. La méthode a besoin de plusieurs outils et de ressources linguistiques variées :

– analyse syntaxique de surface (Collins, 1996) ;
– analyse de structures prédicat–arguments (Kittredge et Mel'cuk, 1983) ;
– règles de paraphrase (Iordanskaja et al., 1991) ;
– étiquetage sémantique (Elhadad, 1992) ;
– utilisation de WordNet [1].

Un aspect très intéressant de ce travail est que l'évaluation a du se faire manuellement. En effet, les auteurs mentionnent le manque d'étalon pour d'autres méthodes que l'extraction. D'ailleurs, ils trouvent que les méthodes d'évaluation automatique classiques sont uniquement applicables pour des résumés par extraction. L'algorithme est donc évalué par comparaison de l'intersection entre les structures prédicat–arguments produites automatiquement et celles produites manuellement. L'évaluation a identifié correctement 74% des structures prédicat–arguments.

1. http://wordnet.princeton.edu/

1.2 La tâche de compression de phrases

Dans la section 1.1 nous avons décrit des travaux liés à la tâche de compression de phrases ayant des applications concrètes. D'autres applications possibles ont été proposées par d'autres auteurs au cours de la même décennie : la génération de sous-titres (Linke-Ellis, 1999; Robert-Ribes et al., 1999; Daelemans et al., 2004) ; la compression textuelle pour l'indexation par des moteurs de recherche (Corston-Oliver et Dolan, 1999) et les systèmes de lecture pour les mal–voyants (Grefenstette, 1998).

Néanmoins, les travaux (Knight et Marcu, 2000, 2002) sont considérés comme les pionniers de la tâche de compression de phrases pour diverses raisons. Pour la première fois, le terme *Sentence Compression* a surgi. Un degré de formalité mathématique a été employé pour décrire la tâche. Un corpus en anglais a été constitué pour permettre à d'autres chercheurs de comparer plusieurs approches en utilisant les mêmes données. Knight et Marcu ont évité de créer un corpus de phrases compressées manuellement. Au lieu de cela, des paires (phrase, phrase-compressée) ont été extraites automatiquement à partir d'articles de produits technologiques et leurs résumés. Une paire a été ajoutée au corpus si la phrase compressée, provenant du résumé, était une sous-séquence d'une phrase dans l'article. Sur les plus de 4 000 articles du corpus Ziff–Davis, seulement 1 067 paires ont été extraites. Cette faible proportion est probablement due au fait que même si un auteur réutilise leurs phrases, il ne se limite pas à éliminer des mots. En effet, les humains utilisent des synonymes, la paraphrase et des constructions syntaxiques alternatives. En conséquence, toute phrase apparue dans le résumé et qui n'a pas été générée par stricte élimination de mots :

- n'apparaît pas dans l'ensemble de paires (phrase, phrase compressée) ;
- ne pourrait pas être générée par la méthode de Knight et Marcu (ni par d'autres méthodes basées sur celle-ci).

D'après (Knight et Marcu, 2002), la tâche de compression de phrases se définit ainsi : soit une phrase φ une séquence de n mots $\varphi = (w_1, w_2, \ldots, w_n)$. Un algorithme peut annuler tout sous-ensemble de ces mots. Les mots qui restent (sans en modifier l'ordre) constituent une compression.

Nous introduisons une définition plus orientée vers le résultat attendu lors de la compression automatique. Etant donnée la phrase d'origine φ, un algorithme doit trouver une nouvelle version $\widetilde{\varphi^\star}$ devant à la fois :

- être plus courte, ($\mathbf{Lon}(\widetilde{\varphi^\star}) \leq \mathbf{Lon}(\varphi)$) ;
- contenir les informations les plus importantes ($\mathbf{Info}(\widetilde{\varphi^\star}) \simeq \mathbf{Info}(\varphi)$) ;
- être grammaticalement correcte ($\mathbf{Gram}(\widetilde{\varphi^\star}) \simeq \mathbf{Gram}(\varphi)$).

Où $\mathbf{Lon}(\bullet)$ est une fonction qui compte le nombre de mots dans une séquence ; $\mathbf{Info}(\bullet)$ est une fonction qui mesure l'informativité d'une phrase et $\mathbf{Gram}(\bullet)$ est une fonction qui mesure sa grammaticalité. Ainsi, $\widetilde{\varphi^\star}$ est un résumé de la phrase φ à condition de ne pas modifier l'ordre des mots ni de les substituer. Afin de calculer le profit, on définit le taux de compression (équation 1.1) comme le rapport du volume de mots après la compression sur le volume initial de φ.

$$\tau = \frac{\mathbf{Lon}(\widetilde{\varphi^{\star}})}{\mathbf{Lon}(\varphi)} \tag{1.1}$$

Sous ces restrictions, chaque mot de la séquence $\varphi = (w_1, w_2, \ldots, w_n)$ peut apparaître ou non dans $\widetilde{\varphi^{\star}}$. Ceci implique que l'algorithme doit choisir la sous-séquence qui optimise les trois critères parmi toutes les sous-séquences possibles. On en déduit le nombre de candidats à la compression possibles ; soit un espace de recherche de 2^n.

Des trois attributs mentionnés, uniquement la longueur peut être mesurée de façon directe. En effet, la possibilité de mesurer automatiquement le degré d'information importante ainsi que la grammaticalité est encore un défi important du *Traitement Automatique de la Langue Naturelle*. Notre étude vise donc à quantifier et à optimiser de façon simultanée, l'informativité et la grammaticalité des phrases tout en diminuant leur longueur.

Dans (Knight et Marcu, 2000), les auteurs ont présenté deux méthodes d'apprentissage supervisé pour générer des phrases compressées : le modèle du canal bruité (*Noisy Channel Model*) usuellement employé dans le domaine de la traduction automatique et le modèle par arbres de décision (*Decision Tree Learning*). Tous les deux traitent l'arbre syntaxique. Dans le modèle du canal bruité, on suppose qu'il existe une phrase compressée $\widetilde{\varphi}$ qui a été modifiée, de façon à ce qu'elle contienne des informations complémentaires dans sa version longue φ. Un algorithme doit trouver $\widetilde{\varphi}$ à partir de φ en supprimant des mots de φ. $\mathbf{P}(\widetilde{\varphi})$ est le modèle de la source, c'est-à-dire, la probabilité que la phrase $\widetilde{\varphi}$ existe. Et $\mathbf{P}(\varphi|\widetilde{\varphi})$ est le modèle du canal, soit l'estimation de probabilité de transformer $\widetilde{\varphi}$ en φ. La meilleure compression est trouvée en maximisant l'équation (1.2).

$$\mathbf{P}(\varphi, \widetilde{\varphi}) = \mathbf{P}(\widetilde{\varphi})\mathbf{P}(\varphi|\widetilde{\varphi}) \tag{1.2}$$

Dans la pratique, $\mathbf{P}(\widetilde{\varphi})$ est calculée selon le score de l'arbre syntaxique de $\widetilde{\varphi}$ (*Standard Probabilistic Context-Free Grammar score, SPCFG*) (Collins, 1996) et les probabilités des bigrammes de $\widetilde{\varphi}$ selon les fréquences du corpus *Penn Treebank*[2]. Pour estimer $\mathbf{P}(\varphi|\widetilde{\varphi})$ on calcule la probabilité de l'ensemble des opérations nécessaires pour transformer l'arbre syntaxique de $\widetilde{\varphi}$ en l'arbre syntaxique de φ. La probabilité de ces opérations est estimée sur la base des fréquences dans le corpus d'apprentissage.

Dans le modèle d'arbres de décision l'algorithme apprend comment transformer l'arbre syntaxique de la phrase d'origine en l'arbre syntaxique de la phrase compressée. Les mots de la phrase d'origine sont placés dans une pile et le modèle indique à quel moment réaliser une des opérations suivantes : se déplacer, réduire ou supprimer, en se basant sur l'information de ce qui reste dans la pile et d'une portion de l'arbre syntaxique de la phrase compressée.

En général, il y a deux problèmes majeurs liés à l'apprentissage de règles avec des grammaires non contextuelles (*Context-Free Grammar*, CFG) : a) on ne peut pas dériver tous les arbres des phrases compressées avec des productions CFG, et b) plusieurs

2. http://www.cis.upenn.edu/~treebank/

règles sont vues une seule fois dans le corpus. (Galley et McKeown, 2007) concluent que le manque d'exemples du corpus Ziff-Davis (1 067 paires), est la cause de la pauvreté des compressions générées.

(Clarke et Lapata, 2006), ont proposé de compresser les phrases en tenant compte de leur contexte. Ils n'ont pas utilisé le corpus Ziff-Davis car dans leur méthode il y avait des contraintes discursives ayant besoin d'annotations des caractéristiques du contexte. Ils ont indiqué aux annotateurs d'effacer des mots isolés dans chaque phrase. Pour la première fois, un corpus de compression de phrases a été annoté manuellement en considérant le contexte (Cohn et Lapata, 2009).

(McDonald, 2006) a obtenu des phrases compressées en utilisant une méthode glouton qui prend en compte les caractéristiques de la phrase. McDonald a analysé également les arbres syntaxiques, mais avec l'intention de caractériser des attributs dans un modèle d'apprentissage et non pas pour produire les phrases compressées à partir de son arbre. Il reprend également des informations à partir de l'arbre de dépendances et des mots. Il utilise l'apprentissage discriminant afin de pondérer ces attributs (qui ne sont pas nécessairement indépendants).

Nous souhaitons insister sur le fait que jusqu'ici toutes les expériences mentionnées ont été effectuées en anglais. Or, nous considérons impératif de franchir cette barrière langagière notamment pour l'espagnol et le français. (Waszak et Torres-Moreno, 2008; Yousfi-Monod et Prince, 2006) présentent des résultats intéressants de compression de phrases sur des documents en français. (Daelemans et al., 2004) développent un système pour le résumé de sous-titres en hollandais. (Aluísio et al., 2008) étudient les phénomènes de la simplification de phrases en portugais. Au cours des dernières années, les chercheurs ont exploré la compression de phrases. Il existe donc des connaissances solides, mais qui restent des bases mobilisables pour poursuivre les recherches sur ce sujet.

1.3 La compression de phrases et les systèmes de résumé automatique

L'article (Molina et al., 2010a) est notre première publication concernant la compression de phrases et, à notre connaissance, le premier travail pour compression de phrases en espagnol. L'intuition qui a motivé ce travail est que la compression de phrases peut être une voie vers la génération automatique de résumés par abstraction. Nous avons considéré la compression comme une forme élémentaire de paraphrase. Nous avons fait des expériences avec des compressions manuelles et automatiques, en suivant des stratégies plutôt simples, avec l'intention de trouver des pistes qui pourraient s'avérer utiles ultérieurement.

La méthodologie utilisée est présentée dans la figure 1.1. Dans un premier temps, nous avons sélectionné un corpus de 40 articles de recherche en médecine de la re-

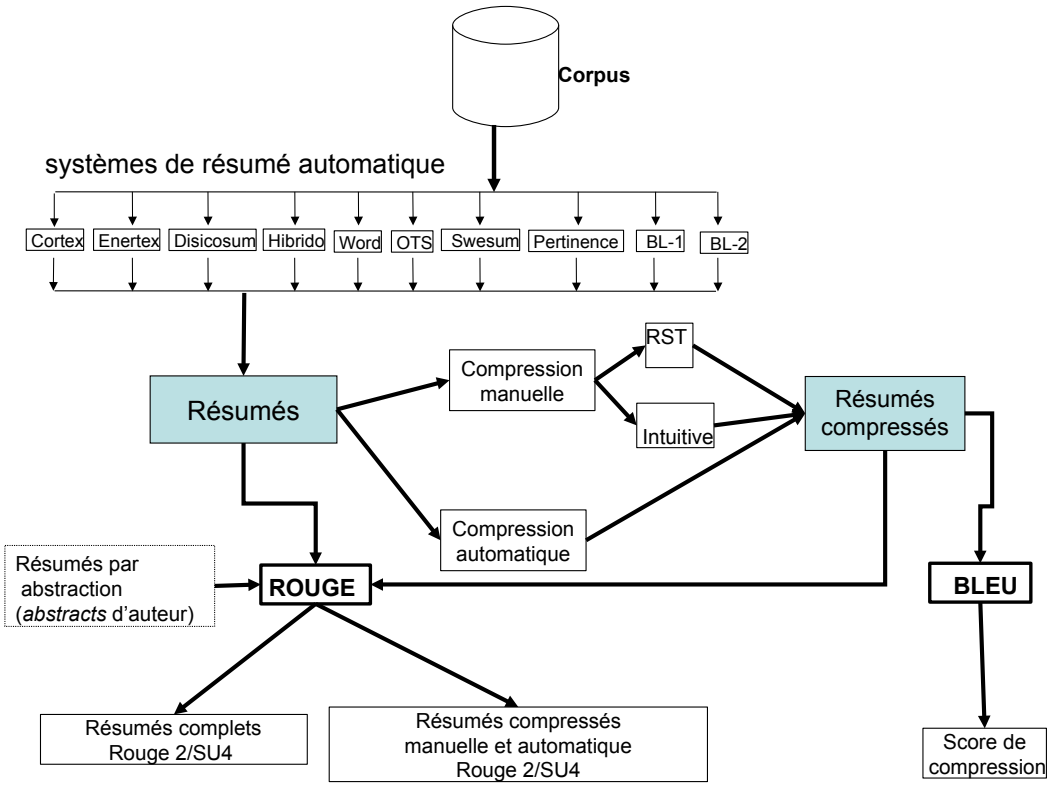


FIGURE 1.1 – *Protocole expérimental pour l'évaluation de l'impact de la compression de phrases appliqué sur dix systèmes de résumé par extraction.*

vue *Medicina Clínica* [3]. Puis, nous avons généré des résumés par extraction en utilisant plusieurs systèmes de résumé automatiques existants de l'état de l'art. Ensuite, nous avons compressé les phrases résultantes des résumés extraits en utilisant des compressions manuelles et automatiques. Nous avons évalué les résultats avec des métriques automatiques.

L'idée est d'éliminer des parties inutiles des phrases extraites par les systèmes extractifs afin de vérifier si la compression améliore la qualité des résumés. Les systèmes de résumé automatique par extraction [4] utilisés sont : Cortex (Torres-Moreno et al., 2001; Boudin et Torres-Moreno, 2007), Enertex (Fernández et al., 2008), Disicosum (da Cunha et Wanner, 2005; da Cunha et al., 2007), un système hybride (da Cunha et al., 2007; Vivaldi et Rodríguez, 2001), Swesum (Dalianis, 2001), Microsoft Word-Office Summarizer [5], Pertinence Summarizer [6] et Open Text Summarizer [7].

Nous avons utilisé trois stratégies pour compresser les phrases produites par les

3. www.elsevier.es/es/revistas/medicina-clinica-2
4. Nous utilisons aussi le terme « *résumeur* ».
5. www.microsoft.com/education/en-us/teachers/how-to/Pages/autosummarize-document.aspx
6. www.pertinence.net
7. libots.sourceforge.net

résumeurs : une compression manuelle intuitive, une compression manuelle par élimination de satellites basée sur la théorie de la structure rhétorique [8] (*Rhetorical Structure Theory* ou RST) (Mann et Thompson, 1988) et une compression par élimination automatique de parenthèses, d'adjectifs et d'adverbes. Le tableau 1.2 montre l'exemple d'une phrase compressée par chacune de ces stratégies. Les fragments barrés correspondent aux parties éliminées.

La RST est une théorie de l'organisation du document où il existe deux types d'éléments discursifs : les noyaux et les satellites. Les noyaux gardent les informations essentielles de la phrase, tandis que les satellites, les informations complémentaires. Deux annotateurs experts en RST ont éliminé systématiquement tous les éléments identifiés comme des satellites. La compression par élimination de satellites a été réalisée manuellement car à ce moment là il n'existait pas d'analyseur discursif pour l'espagnol capable de distinguer automatiquement les noyaux des satellites.

Pour la compression manuelle intuitive, nous avons employé la même méthodologie que celle de la construction du corpus de phrases compressées en français du projet ANR-RPM2 (de Loupy et al., 2010). Cette stratégie implique un haut degré de subjectivité car les annotateurs sont libres d'éliminer ce qu'ils veulent. Une personne peut considérer qu'un élément est pertinent et décider de le conserver, tandis qu'une autre peut décider de l'éliminer.

Pour la compression par élimination automatique de parenthèses, d'adjectifs et d'adverbes, nous avons développé quatre systèmes de compression, en suivant quelques idées de (Yousfi-Monod et Prince, 2006). Ces auteurs identifient essentiellement deux classes de constituants syntaxiques potentiellement effaçables : les modifieurs et les compléments. Ils montrent que les adjectifs et les adverbes sont des éléments non essentiels de la phrase quand ils fonctionnent comme modifieurs. Ils mettent également en évidence d'autres éléments non essentiels tels que certains déterminants, parenthétiques et locutions. Nous avons donc proposé un système de compression de phrases qui utilise les catégories grammaticales obtenues avec TreeTagger (Schmid, 1995) et des règles simples d'élimination de mots. À titre d'exemple, nous montrons dans le tableau 1.2, une phrase extraite de la « *Convention de Bâle sur le Contrôle des mouvements Transfrontaliers de Déchets Dangereux et de leur Elimination* » [9]. Les fragments barrés représentent des parties éliminées selon chaque stratégie.

Les résultats de cette première étude exploratoire nous ont donné des pistes très intéressantes qui ont orienté la suite de nos recherches. La lecture directe des phrases produites a révélé qu'en de nombreux cas la grammaticalité des phrases compressées a été dégradée. Le système d'élimination de parenthèses et d'adverbes a compressé très peu et quand il l'a fait, il a produit souvent des phrases agrammaticales. Cependant, l'élimination systématique de tous les satellites d'une phrase n'est pas non plus une bonne stratégie. Il faut donc une méthode pour décider de la pertinence d'éliminer un segment discursif. Nous remarquons aussi que 13% des séquences éliminées com-

8. Cette théorie est expliquée dans le chapitre 2.

9. http://europa.eu/legislation_summaries/environment/waste_management/l28043_fr.htm

mencent par une virgule, ce qui nous à donné l'idée d'utiliser ce signe de ponctuation pour d'obtenir une segmentation encore plus fine.

En ce qui concerne l'évaluation des phrases compressées, à notre connaissance il n'y a pas de métrique consacrée spécifiquement à la tâche de compression de phrases. Nous avons essayé d'évaluer avec les métriques les plus répandues pour l'évaluation de résumés par extraction : ROUGE (*Recall-Oriented Understudy for Gisting Evaluation*) (Lin, 2004) et BLEU (Papineni et al., 2002), développées par IBM pour la traduction automatique. Ces deux métriques ne sont pas adaptées pour évaluer des phrases compressées. Nous dédierons un chapitre à cette question. Il est assez étonnant, par exemple, que les scores ROUGE obtenus pour les résumés compressés soient très proches de ceux obtenus sans compression. En effet, il a pénalisé les résumés compressés pour avoir moins de *n*-grammes que les résumés sans compression de phrases.

1.4 Conclusions du chapitre

Dans ce premier chapitre, nous avons fait un état de l'art des premières études liées à la compression de phrases avant le travail pionnier de Knight et Marcu. Puis, nous avons défini la tâche de compression de phrases. Pour ce faire, nous avons non seulement repris la définition proposée par Knight et Marcu, mais proposé une nouvelle définition plus orientée vers le résultat attendu lors de la tâche de compression de phrases.

Nous avons présenté la méthodologie et les résultats de nos premières expériences.

Phrase d'origine :
Considéré comme moralement contraignant par les signataires, l'amendement interdit l'exportation de déchets dangereux de certains pays développés (pour la plupart membres de l'OCDE).
Phrase compressée par élimination de satellites RST :
~~Considéré comme moralement contraignant par les signataires,~~$_{satellite}$ l'amendement interdit l'exportation de déchets dangereux de certains pays développés$_{noyau}$ ~~(pour la plupart membres de l'OCDE).~~$_{satellite}$
Phrase compressée par élimination automatique de parenthèses, d'adjectifs et d'adverbes :
Considéré comme ~~moralement~~$_{adv}$ contraignant par les signataires, l'amendement interdit l'exportation de déchets ~~dangereux~~$_{adj}$ de certains pays ~~développés~~$_{adj}$ ~~(pour la plupart membres de l'OCDE)~~$_{par}$.
Phrase compressée par élimination intuitive :
Considéré ~~comme moralement~~ contraignant par les signataires, l'amendement interdit l'exportation de déchets dangereux de certains pays ~~développés (pour la plupart membres de l'OCDE)~~.

TABLE 1.2 – *Exemple d'une phrase compressée selon trois stratégies différentes : élimination manuelle de satellites de la RST ; élimination manuelle intuitive de mots ; élimination automatique de parenthèses, d'adjectifs et d'adverbes.*

Grâce à celles-ci, nous avons vérifié que les résultats des résumeurs par extraction peuvent améliorer en compressant les phrases résultantes.

Bien que la majorité des systèmes de résumé par extraction ont amélioré significativement, d'autres ont à peine évolué. Ces résultats ont engendré beaucoup de réflexion autour de l'adaptation d'un module de compression à la sortie d'un système de résumé automatique par extraction. Nous sommes arrivés à la conclusion qu'il faut plutôt intégrer la compression *dans* le système et non pas uniquement l'appliquer à la sortie d'un système classique d'extraction.

L'élimination de segments discursifs est la stratégie la plus intéressante pour la compression. Nous avons vu que l'élimination de parenthèses, d'adjectifs et d'adverbes a produit des taux de compression très bas et souvent des phrases agrammaticales. Ceci nous a mené à proposer une méthode pour pallier la grammaticalité.

De plus, le fait d'avoir trouvé que 13% des séquences éliminées commencent par une virgule, nous à donné l'idée d'utiliser les signes de ponctuation pour obtenir une segmentation encore plus fine. Dans le chapitre 2, nous expliquerons de manière détaillée les aspects d'un nouvelle segmentation discursive orientée vers la compression de phrases.

L'évaluation automatique des résumés avec des phrases compressées est encore un problème non résolu car elle doit à la fois considérer la grammaticalité, le taux de compression et la pertinence des contenus gardés. Nous ferons ultérieurement une analyse plus détaillée des raisons pour lesquelles ni ROUGE, ni BLEU ne sont bien adaptés à la compression de phrases.

Chapitre 2

Segmentation discursive pour la compression de phrases

Sommaire

Dans ce chapitre nous présentons des travaux sur la compression de phrases plus récents qui portent pour l'élimination de fragments plutôt que des mots. Après une étude des fragments éliminés d'un corpus, nous montrons qu'il serait très utile de segmenter la phrase en unités discursives élémentaires grâce à une technique de segmentation discursive. Finalement, nous expliquons comment le segmenteur discursif utilisé lors de nos expériences a été optimisé pour la tâche de compression de phrases.

2.1 Études récentes de la compression de phrases

Dans le chapitre précédent (1), nous avons discuté sur plusieurs approches pour l'élimination de mots à l'intérieur des phrases. Cependant, selon (Jing et McKeown, 1999), les humains ont tendance à réduire les phrases en effaçant des passages textuels et rarement des mots isolés. De plus, la suppression de mots individuels n'a parfois qu'un faible impact sur le taux de compression. L'élimination d'un seul mot d'une phrase d'une quarantaine ne réduit pas significativement sa taille originale mais peut être très risquée par rapport à la grammaticalité (Molina et al., 2010a). Par exemple, éliminer un verbe ou une négation peut modifier fortement le sens de la phrase et dans le pire des cas, la rendre grammaticalement incorrecte. Tout cela a motivé les chercheurs à éliminer des passages textuels plutôt que des mots isolés lors la tâche de compression de phrases.

Des études récentes (toutes pour l'anglais) ont obtenu de bons résultats en se concentrant sur l'élimination de propositions à l'intérieur de la phrase. Par exemple, l'algorithme proposé par (Steinberger et Jezek, 2006) découpe d'abord les phrases en propositions simples (*clauses*). Ensuite, les propositions simples du niveau inférieur (les feuilles dans l'arbre de dépendances) sont éliminées de la phrase. Les phrases ainsi réduites sont évaluées en utilisant *Latent Semantic Analysis* (LSA) (Deerwester et al., 1990; Landauer et al., 1998). En effet, dans cette approche, l'informativité est mesurée grâce au LSA mais aucune composante pour pallier la génération de phrases agrammaticales n'a été incluse dans cet algorithme. Le problème crucial de cette étude est que, si le sujet principal de la phrase est éliminé, la phrase peut être gravement dégradée. Les auteurs ont réussi à améliorer cette méthode avec une approche d'apprentissage supervisé (Steinberger et Tesar, 2007). Ceci demande de disposer d'un corpus préalablement annoté de taille considérable.

Comme alternatives à l'élimination de propositions simples, d'autres études ont exploré les structures discursives (Sporleder et Lapata, 2005). Les auteurs argumentent que l'analyse discursive au niveau du document représente toujours un défi important, et qu'appliquée au niveau de la phrase, elle pourrait représenter une alternative réaliste à la tâche de compression de phrases. De plus, des modèles basés sur l'analyse discursive au niveau de la phrase ont montré une précision comparable à la performance humaine (Soricut et Marcu, 2003). Une technique de mémorisation par bloc du discours a été présentée par (Sporleder et Lapata, 2005), comme alternative à l'analyse discursive, montrant ainsi une application directe à la compression de phrases.

Nous nous sommes intéressés par l'élimination d'unités discursives pour la compression de phrases (Molina et al., 2011). Nous sommes partis de l'idée que si une phrase est déjà assez simple, elle peut être considérée comme une unité discursive élémentaire. Elle n'aura donc pas besoin d'être compressée. Si, au contraire, elle est longue et complexe, elle contiendra plusieurs unités discursives et probablement quelques-unes sans importance (par rapport au contexte) qui pourraient être éliminées.

Dans les sections suivantes, nous décrivons le cadre théorique dans lequel nous nous basons pour la segmentation discursive appliquée à la compression de phrases.

2.2 La théorie de la structure rhétorique

La théorie de la structure rhétorique (*Rhetorical Structure Theory*, RST) est une théorie d'organisation textuelle basée sur l'idée qu'un document peut être segmenté en unités discursives qui sont reliées entre elles (Mann et Thompson, 1988). Ceci donne lieu à un arbre rhétorique hiérarchique qui représente le document dans son intégralité. Les unités discursives élémentaires (en anglais *Elementary Discourse Units ou EDUs*) correspondent aux feuilles de l'arbre RST. Ces unités discursives peuvent être des noyaux ou des satellites. Les noyaux offrent une information pertinente par rapport aux propos de l'auteur du texte. Les satellites apportent une information additionnelle aux noyaux, dont ils dépendent. Dans le cadre de la RST, il y a des relations discursives de différents types. Les listes de ces relations et leurs définitions peuvent être consultées sur le site Web de la RST pour l'anglais [1], le français [2] et l'espagnol [3].

Les relations RST peuvent être noyau-satellite ou multinucléaires. Dans le cas des relations noyau-satellite, un satellite dépend d'un noyau. Dans le cas de relations multinucléaires, divers noyaux (au moins deux) sont en relation au même niveau.

Un exemple d'arbre RST est présenté dans la figure 2.1. Le texte correspond au résumé d'un article intitulé « *Darwin : un géologue* » extrait du site Web de la RST pour le français. Le résumé a été divisé en cinq unités (le titre étant considéré comme une unité) :

[Darwin : un géologue]_s_1 [Aujourd'hui, on a tendance à le considérer comme un biologiste,]_s_2 [mais durant ses cinq années à bord du Beagle, ses travaux concernaient essentiellement la géologie]_s_3 [et il se considérait lui-même comme géologue.]_s_4 [Ses travaux constituent une contribution significative à ce domaine.]_s_5

Dans la figure 2.1, la flèche qui va de l'unité s_2 vers l'unité s_1, indique que l'unité s_2 est le satellite de l'unité s_1, le noyau. Il s'agit d'une relation de type CONCESSION. À leur tour, les unités s_1 et s_2 comprennent le noyau de trois relations de type DÉMONSTRATION.

L'analyse discursive avec la RST inclut trois étapes consécutives :

1. La segmentation discursive ;
2. La détection de relations discursives ;
3. La construction d'arbres rhétoriques hiérarchiques.

La segmentation discursive vise à trouver les frontières des unités discursives d'un texte. La détection de relations discursives détecte quelles sont les unités qui sont reliées entre elles et quel est le type de la relation. Enfin, la construction de l'arbre rhétorique hiérarchique cherche à identifier la nucléarité de chaque unité et éventuellement l'importance vis-à-vis du texte en entier (la hiérarchie). La figure 2.2 illustre avec un exemple, le processus d'analyse RST intra-phrase.

1. http://www.sfu.ca/rst/01intro/definitions.html
2. http://www.sfu.ca/rst/07french/definitions.html
3. http://www.sfu.ca/rst/08spanish/definiciones.html

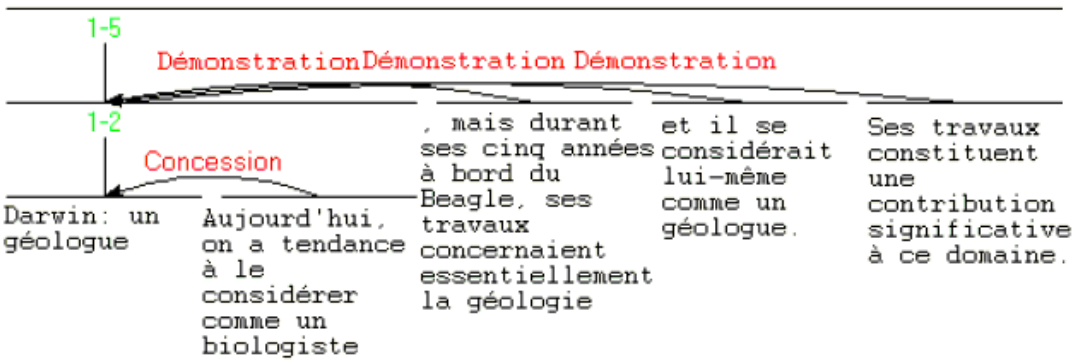


FIGURE 2.1 – *Exemple d'un arbre rhétorique de la RST appliqué à l'article « Darwin : un géologue ».*

Dans cette étude, nous utilisons exclusivement la première phase de l'analyse, la segmentation discursive, pour délimiter les phrases avant de décider s'il est possible d'effacer certains segments discursifs. Concrètement, nous partons de la conception de segmentation discursive de (Tofiloski et al., 2009) : « *La segmentation discursive est le processus pour décomposer le discours en Unités Discursives Élémentaires (EDUs), qui peuvent être des phrases simples ou des propositions dans une phrase complexe, et à partir desquelles des arbres discursifs sont élaborés* ». Cependant, notre objectif n'est pas de construire l'arbre discursif d'un document mais de segmenter les phrases en ses unités discursives afin d'éliminer celles qui ne sont pas importantes dans la phrase, du point de vue de l'informativité.

Dans les sections suivantes nous verrons qu'en effet la segmentation discursive au niveau de la phrase peut être utile à l'identification des EDUs éliminables lors de la tâche de compression.

2.3 DiSeg : un segmenteur discursif pour l'espagnol

Un segmenteur discursif est un système qui sert à détecter les EDUs d'un texte. Nous nous sommes intéressés particulièrement à l'analyse discursive intra-phrase. Le tableau 2.1 montre le résultat de la segmentation discursive de la phrase « *Juliette prépare un gâteau pour le manger bien qu'elle n'ait pas faim.* ».

Notre méthode de résumé par compression de phrases vise à décider quels segments discursifs doivent être éliminés à partir d'une phrase qui a été segmentée en EDUs. Dans le tableau 2.1, les parties en gras sont des éléments très particuliers de la phrase. Ce sont des marqueurs discursifs explicites, grâce auxquels il est possible de détecter les frontières entre les segments discursifs. Quelques exemples de marqueurs discursifs en français sont : « *afin de* », « *pour que* », « *donc* », « *quand bien même que* »,

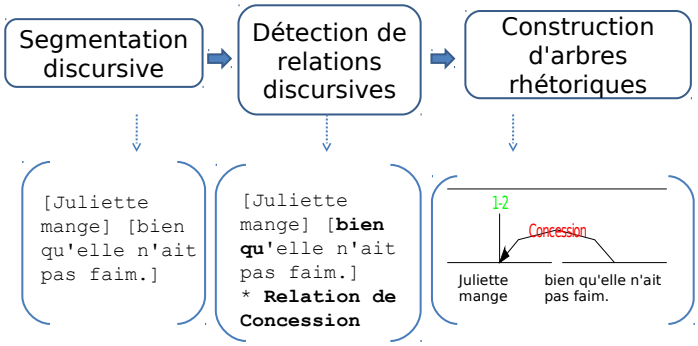


FIGURE 2.2 – *Exemple des trois étapes de l'analyse discursive intra-phrase.*

« *ensuite* », « *de fois que* », « *globalement* », « *par contre* », « *sinon* », « *à ce moment-là* », « *cependant* », « *subséquemment* », « *puisque* », « *au fur et à mesure que* », « *si* », « *finalement* », etc. Quant à l'espagnol nous pouvons citer :« *con la finalidad de* », « *para* », « *entonces* », « *aún cuando* », « *luego* », « *a veces* », « *globalemente* », « *al contrario* », « *si no* », « *en ese momento* », « *sin embargo* », « *subsecuentemente* », « *dado que* », « *en la medida que* », « *si* », « *a fin cuentas* », etc.

(Roze et al., 2012) présentent une étude approfondie des marqueurs discursifs en français ainsi qu'un dictionnaire disponible sur le Web [4]. Pour l'espagnol, (La Rocca, 2012) présente une compilation exhaustive de marqueurs discursifs. Toutes les langues possèdent des marqueurs discursifs, ce qui permet, en théorie, le développement de segmenteurs discursifs en différentes langues. Néanmoins, la segmentation discursive et complexe car toutes les EDUs ne portent pas de marqueurs explicites ou alors ces marqueurs-ci sont ambigus. Pour ces raisons, la segmentation discursive a besoin de

4. http://www.linguist.univ-paris-diderot.fr/~croze/D/Lexconn.xml

Entrée :

<φ>Juliette prépare un gâteau pour le manger bien qu'elle n'ait pas faim.</φ>

Sortie :

<φ>
<s_1>Juliette prépare un gâteau</s_1>
<s_2>**pour** le manger</s_2>
<s_3>**bien qu'**elle n'ait pas faim.</s_3>
</φ>

TABLE 2.1 – *Exemple du résultat de la segmentation discursive.*

ressources linguistiques supplémentaires telles que analyse syntaxique de surface et d'autres heuristiques spécifiques.

Il existe des segmenteurs discursifs pour le français (Afantenos et al., 2010), l'anglais (Tofiloski et al., 2009), le portugais (Maziero et al., 2007), le Thaï (Ketui et al., 2012) et l'espagnol (da Cunha et al., 2010). Cela est important car on peut ainsi projeter d'étendre nos études de compression à d'autres langues. Dans cette thèse nous avons utilisé DiSeg, un segmenteur discursif RST pour l'espagnol (da Cunha et al., 2010), disponible sur le Web [5].

DiSeg a été créé par modification de la grammaire d'un analyseur syntaxique de surface (Castellón et al., 1998). Ces modifications recatégorisent certaines expressions considérées comme des « *candidats à marqueurs discursifs* ». On parle de « *candidats* » parce qu'il existe des marqueurs ambigus. En effet, les marqueurs discursifs peuvent avoir des formes diverses. Elles peuvent être simples : « *comme* » (« *como* ») et « *alors* » (« *entonces* »). Elles peuvent être composées : « *par exemple* » (« *por ejemplo* ») et « *au contraire* » (« *al contrario* »). Elle peuvent aussi être plus complexes : « *d'abord ... ensuite* » (« *primero ... luego* ») et « *d'un côté ... d'un autre côté* » (« *por un lado ... por otro lado* »). Le processus de segmentation de DiSeg se fait en deux étapes (da Cunha et al., 2012). Dans un premier temps, des frontières candidates sont détectées grâce à l'apparition de différents éléments tels que :

- les formes verbales (participe présent, gérondif et infinitif) ;
- les conjonctions ;
- les propositions subordonnées ;
- d'autres éléments classés dans la grammaire comme des *candidats à marqueurs discursifs* : *por ejemplo* (par exemple), *al contrario* (au contraire).

Ensuite, les EDUs sont définies en utilisant un analyseur renversé qui lit de droite à gauche (d'où son nom) la sortie de la première étape et insère une balise de frontière à condition d'avoir un verbe conjugué dans chaque sous-segment. DiSeg a été évalué en comparant ses résultats avec un corpus de référence (*gold standard*) segmenté manuellement. Le système initial (da Cunha et al., 2010) a obtenu un F-Score de 80% avec un corpus issu de textes médicaux. Le système optimisé (da Cunha et al., 2012) a obtenu un F-Score de 96% avec un corpus de textes médicaux et 91% avec un corpus de textes de terminologie.

Nous avons étudié l'adaptation de DiSeg pour la tâche de compression de phrases (Molina et al., 2011). Nous avons constaté qu'un bon nombre de passages éliminés par les annotateurs correspondent à des segments discursifs. Cependant, nous avons repéré quelques inconvénients lors de la détection des segments qui peuvent être éliminés. Nous avons observé certains fragments de texte que les annotateurs ont éliminés mais que le système est incapable de détecter comme EDUs. Ce sont les cas des :

- fragments qui commencent par un participe ;
- propositions relatives ;

5. http://daniel.iut.univ-metz.fr/~iula/WebDiSeg/

– fragments qui commencent par un marqueur discursif mais qui n'incluent pas un verbe.

À partir des résultats décrits dans les sections 2.4 et 2.5, nous avons modifié DiSeg pour créer un nouveau segmenteur. Il vise notamment à détecter des fragments textuels que les annotateurs ont éliminés mais que DiSeg n'a pas reconnus. Dans la section 2.6 nous expliquons comment nous avons adapté DiSeg à la compression de phrases.

2.4 Analyse quantitative des fragments éliminés

Dans (Molina et al., 2011), nous avons choisi un corpus de quatre genres : sections Wikipédia, nouvelles journalistiques brèves, résumés d'articles scientifiques et littéraire (contes et nouvelles courtes). Ces textes ont été distribués entre deux annotateurs linguistes experts, auxquels nous avons demandé de les lire, puis de les compresser, phrase par phrase en suivant ces instructions :

– ne pas réécrire les phrases ;
– ne pas modifier l'ordre des mots ;
– ne pas substituer des mots ;
– s'assurer que les phrases compressées soient grammaticales ;
– s'assurer que les phrases compressées contiennent le même sens que celui de la phrase d'origine ;
– s'assurer que le document conserve son sens d'origine.

Nous avons segmenté les phrases de notre corpus avec DiSeg. Par la suite, nous avons extrait tous les fragments éliminés des phrases d'origine. Nous les avons classifiés en trois classes :

1. fragments éliminés correspondants à des EDUs détectées par DiSeg ;
2. fragments éliminés non détectés par DiSeg comme étant des EDUs :
   (a) fragments avec sens discursif ;
   (b) fragments sans sens discursif.

La classe 1 contient des fragments textuels éliminés par les annotateurs qui ont été reconnus par DiSeg. Comme nous l'avons déjà mentionné, ces EDUs peuvent être détectées partiellement en raison de la présence des marqueurs discursifs. Comme prévu, il y a des fragments qui contiennent un marqueur discursif mais qui ne coïncident pas exactement aux EDUs identifiées par DiSeg, bien que la majorité le fasse. En effet, les annotateurs ont tendance à éliminer des passages discursifs en entier plutôt que juste les marqueurs. Le tableau 2.2 montre les moyennes du taux de coïncidences entre les fragments dans la classe 1 et des EDUs DiSeg. Il présente aussi le pourcentage des fragments dans la classe 1 qui coïncident exactement avec les EDUs DiSeg. Le taux de coïncidences est défini par :

$$\text{taux de coïncidences} = \frac{\text{Longueur du fragment éliminé}}{\text{Longueur de l'EDU}} \tag{2.1}$$

La classe 2 inclut les fragments éliminés que DiSeg n'a pas reconnu comme étant des EDUs. Cette classe est divisée en deux sous-classes : la classe 2a inclut les fragments qui devraient être considérés comme EDUs mais qui n'ont pas été détectés par DiSeg car ils ne respectent pas les critères du système. La classe 2b est composée d'éléments sans sens discursif, c'est-à-dire, d'unités courtes, comme des phrases adverbiales, des adjectifs et des expressions d'usage. Dans la section 2.5, nous montrons quelques exemples extraits de notre corpus.

Genre	Moyenne du taux de coïncidences (%)	Volume de fragments coïncidents complètement
Wikipédia	91%	73%
Nouvelles	92 %	73%
Scientifique	100%	100%
Littéraire	81 %	57%

TABLE 2.2 – *Proportion de coïncidences pour les fragments éliminés correspondant à des EDUs détectés par DiSeg.*

Le tableau 2.3 montre la distribution du contenu éliminé, appartenant à chacune des classe décrites auparavant. Les passages avec sens discursif (classe 1 ∪ classe 2a) correspondent à la moitié de la totalité du volume que les annotateurs ont effacé. Comme prévu, les genres Wikipédia et nouvelles contiennent plus de passages discursifs éliminables, à cause de leurs explications ou informations supplémentaires. À l'opposé, le genre littéraire exprime l'information d'une manière plus élaborée. Il semble donc plus convenable de préférer d'éliminer des éléments isolés, notamment des adjectifs ou adverbes. En ce qui concerne les textes scientifiques, nous considérons que les résumés contiennent déjà des phrases simplifiées et des informations essentielles. Ceci est dû au fait que les textes scientifiques d'origine sont eux-même des résumés.

Genre	classe 1	classe 2a	classe 2b
Wikipédia	31.55	29.57	38.88
Nouvelles	34.95	16.47	48.58
Scientifique	30.26	17.26	52.48
Littéraire	20.68	09.06	70.26

TABLE 2.3 – *Proportions du contenu (en pourcentage de mots) éliminé dans trois classes : fragments éliminés correspondant à des EDUs détectées par DiSeg ; fragments avec sens discursif ; fragments sans sens discursif.*

Afin de savoir si tous les satellites ont été éliminés systématiquement, nous avons séparé les noyaux et les satellites de la classe 1. Le tableau 2.4 montre la proportion de noyaux et de satellites pour la classe 1. Nous observons que bien que la majorité des EDUs éliminées étaient des satellites, quelques noyaux ont aussi été éliminés.

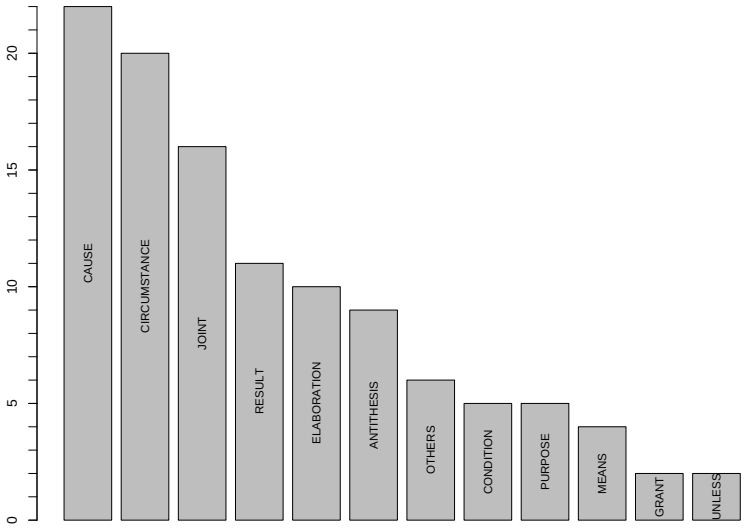


FIGURE 2.3 – *Fréquences des relations RST identifiées par DiSeg.*

Genre	Noyaux	Satellites
Wikipédia	8 (29 %)	20 (71 %)
Nouvelles	9 (13 %)	58 (87 %)
Scientifique	0 (0 %)	8 (100 %)
Littéraire	4 (40 %)	6 (60 %)

TABLE 2.4 – *Proportion des EDUs éliminées correspondant à des noyaux ou à des satellites.*

Les fréquences des EDUs dans la classe 1 divisées par rapport au type de relation RST sont présentées dans la figure 2.3. Nous observons qu'il y a des relations plus susceptibles à élimination que d'autres. Dans nos expériences, nous avons identifié les relations les plus éliminées : les satellites de CAUSE (27.85%), les satellites de CIRCONSTANCE (25.32%), les noyaux de JONCTION (20.25%), les satellites de RÉSULTAT (13.92%) et les satellites d'ÉLABORATION (12.66%).

2.5 Analyse qualitative des fragments éliminés

Après l'analyse quantitative des fragments éliminés, nous avons réalisé une analyse qualitative afin de comprendre quels éléments ont eu tendance à être éliminés durant la compression de phrases. Nous présentons ici les résultats de cette analyse. Les exemples que nous utilisons sont extraits de notre corpus en espagnol. Nous proposons donc une traduction mot-à-mot en français, parfois forcée afin de mettre en évidence

les phénomènes en question [6].

Nous commençons par analyser les différentes régularités que nous avons trouvées pour les segments DiSeg (classe 1). Tous les fragments éliminés que nous montrons par la suite ont été efficacement détectés par DiSeg.

La première remarque intéressante provient du fait qu'il est possible de déterminer la nucléarité, c'est-à-dire, si une EDUs est un noyau ou un satellite, uniquement en regardant le marqueur discursif explicite qu'elle contient. Autrement dit, il n'est pas nécessaire d'analyser la phrase entière pour ce faire. L'exemple 1 illustre cette situation. Nous voyons clairement qu'avec une seule EDUs, il est possible d'identifier un satellite de CAUSE.

> Exemple 1. [**ya que** se reducirían las interacciones entre fármacos, sus efectos adversos, y favorecería el cumplimiento de unos tratamientos que cada vez incluyen más pastillas.]
> Exemple 1. [**étant donné que** les interactions entre les médicaments et leurs effets indésirables seraient réduits en favorisant l'accomplissement des traitements qui incluent de plus en plus de pilules.]

Dans certains cas, le marqueur discursif est ambigu. Il pourrait indiquer plusieurs types de relations rhétoriques. Dans l'exemple 2, le marqueur discursif *cuando* (quand) peut autant indiquer une relation de CIRCONSTANCE qu'une relation de CONDITION. Dans ce cas, il est nécessaire de lire la phrase entière afin de déterminer le type de relation rhétorique correct.

> Exemple 2. [Sin embargo, el uso de Internet a edades cada vez más tempranas representa no solamente una herramienta educativa útil,][sino también puede constituir grandes peligros] [**cuando** su uso está relacionado con contenidos inapropiados para su adecuado desarrollo.]
> Exemple 2. [Cependant, l'utilisation d'Internet par les plus jeunes représente non seulement un outil éducatif important,] [mais aussi, elle peut constituer de grands risques] [**quand** son emploi est lié à des contenus inappropriés.]

Dans d'autres cas, le marqueur est un gérondif, situation également ambiguë. Les exemples 3, 4 et 5 montrent des EDUs contenant un gérondif mais qui indiquent trois différents types de relations rhétoriques : RÉSULTAT (ex. 3), ÉLABORATION (ex. 4) et MÉTHODE (ex. 5).

> Exemple 3. [**limitándose** a reducir el factor de comportamiento sísmico que controla las resistencias de diseño.]
> Exemple 3. [**en se limitant** à réduire le facteur du comportement sismique qui contrôle les résistances du design.]
> Exemple 4. [**diseñando** mejoras para el equipo eléctrico traído del otro lado del océano gracias a las ideas de Edison.]
> Exemple 4. [**en prévoyant** des améliorations pour l'équipement électronique importé de l'autre côté de l'océan grâce aux idées d'Edison.]

6. Ceci est valable pour toutes les traductions que nous utilisons dorénavant.

Exemple 5. [**hablando** acerca de la prevención necesaria.]
Exemple 5. [**en parlant** sur la prévention nécessaire.]

La majorité des EDUs éliminées ont un marqueur discursif explicite, tels que *ya que* (étant donné que) dans l'exemple 1 ou *cuando* (quand) dans l'exemple 2. Cependant, il y a aussi quelques EDUs qui ne contiennent pas de marqueurs. Dans ces cas là, il est plus difficile de leur assigner un type de relation rhétorique. L'exemple 6 illustre cette situation.

Exemple 6. [se incluyeron además corredores entre las plantas hechos con tepujal, un material que ayuda a conservar la humedad en la tierra]
Exemple 6. [de plus, des couloirs faits en tepujal ont été inclus, ce matériau aide à conserver l'humidité de la terre]

Dans la majorité des cas, les EDUs éliminées correspondent à des satellites (exemples 1 à 5), mais parfois elles correspondent à des noyaux (exemple 6). Ceci veut dire que le satellite ne peut pas toujours être éliminé sans une perte importante d'information. De plus, quelquefois le noyau n'est pas essentiel à la compréhension du texte, comme il a été argumenté dans d'autres travaux (Marcu, 2000).

Nous analysons à présent les éléments de la classe 2, c'est-à-dire les paragraphes éliminés par les annotateurs experts qui ne correspondent pas aux EDUs détectées par DiSeg. Nous identifions deux cas : (a) unités avec sens discursif et (b) unités sans sens discursif.

Pour les unités avec sens discursif (a), nous détectons trois régularités : des fragments qui commencent par un participe ; des fragments qui correspondent à des propositions relatives et des fragments qui ne contiennent pas de verbe. L'exemple 7 présente le cas d'un fragment éliminé qui commence par un participe.

Exemple 7. [valorado en 40 000 dólares.]
Exemple 7. [estimée en 40 000 dollars .]

L'exemple 8 montre un cas où le fragment éliminé correspond à une proposition relative.

Exemple 8. [que agrupaba los vídeos más vendidos.]
Exemple 8. [qui regroupait les vidéos les plus vendues.]

L'exemple 9 montre un fragment éliminé qui ne contient pas de verbe.

Exemple 9. [a causa de la malnutrición durante la ocupación alemana.]
Exemple 9. [à cause de la malnutrition durant l'occupation allemande.]

Le critère de segmentation de DiSeg ne détecte pas les fragments exposés dans le cas (a) comme étant des EDUs. Néanmoins, beaucoup de ces fragments ont été éliminés. Nous considérons que la détection de ces unités serait utile à la tâche de compression automatique des phrases. Pour le cas (b), dans la section suivante nous montrons une analyse plus approfondie. Ainsi, en considérant les observations de ces deux cas, nous présentons une adaptation de DiSeg.

2.6 CoSeg : un segmenteur pour la compression de phrases

Nous avons créé CoSeg, (*Compression Segmenter*) un segmenteur orienté pour la compression de phrases. Celui-ci est basé sur DiSeg, mais est plus flexible par rapport à la production de segments discursifs. CoSeg a été adapté pour segmenter les trois cas dont DiSeg ne reconnaissait pas les EDUs : des segments sans verbe, des segments qui commencent par un participe passé et des propositions relatives.

Nous avons constaté que certains segments discursifs sont sujets à être éliminés lors de la compression manuelle de phrases. Si l'on considère la totalité des fragments textuels correspondant à des segments discursifs, elle représente environ la moitié du corpus[7]. Cependant, l'autre moitié correspond à des segments discursifs que DiSeg a été incapable d'identifier ou ce sont des segments non discursifs.

Nous avons modifié les règles de DiSeg afin d'inclure les segments discursifs que DiSeg ne reconnaît pas. Nous avons alors adapté la définition de EDU implicite dans le segmenteur DiSeg. Concrètement, nous avons modifié le module d'application de règles (le dernier dans la chaîne de traitement) pour la détection des EDUs du système DiSeg. Les deux modifications réalisées ont été :

1. La révocation des restrictions verbales.
2. L'addition de marqueurs explicites.
3. L'addition de signes de ponctuation.

Les marqueurs explicites ajoutés correspondent à deux types : les pronoms relatifs et les signes de ponctuation. Ces modifications ont permis l'identification des propositions relatives introduites normalement par des pronoms relatifs et parfois par des signes de ponctuation. On ajoute les marqueurs suivants à la liste pour l'espagnol : « *el que* », « *la que* », « *los que* », « *las que* », « *que* », « *quien* », « *quienes* », « *cuyo* », « *cuya* », « *cuyos* », « *cuyas* », « *el cual* », « *la cual* », « *los cuales* », « *las cuales* », « *cuanto* », « *cuanta* », « *cuantos* », « *cuantas* », « *todo* », « *toda* », « *todos* », « *todas* », « *donde* ». Ces termes correspondent aux marqueurs suivants en français : « *qui* », « *que* », « *quoi* », « *dont* », « *où* », « *lequel* », « *quiconque* » et leurs composés. Les signes de ponctuation ajoutés sont : « *,* », « *;* », « *:* », « *. . .* », « *( )* », « *[ ]* », « « » » et « – ».

Nous avons supprimé la restriction verbale de DiSeg afin de pouvoir inclure des segments sans un verbe explicite. En effet, les critères de ce segmenteur exigent la présence d'un verbe principal qui représente l'action d'un sujet explicite ou implicite.

Le diagramme 2.4 montre l'architecture de DiSeg, d'après (da Cunha et al., 2010). La couche où les règles de segmentation ont été modifiées est indiquée. Avant l'application des règles de détection des EDUs, les étapes antérieures restent intactes.

Afin de mesurer la couverture de CoSeg, nous présentons la figure 2.5. En nous basant sur les 675 fragments que DiSeg n'a pas reconnus, nous avons ajouté les nouvelles

7. L'annexe A montre les parties éliminées par les annotateurs qui correspondent aux segments discursifs identifiés par DiSeg.

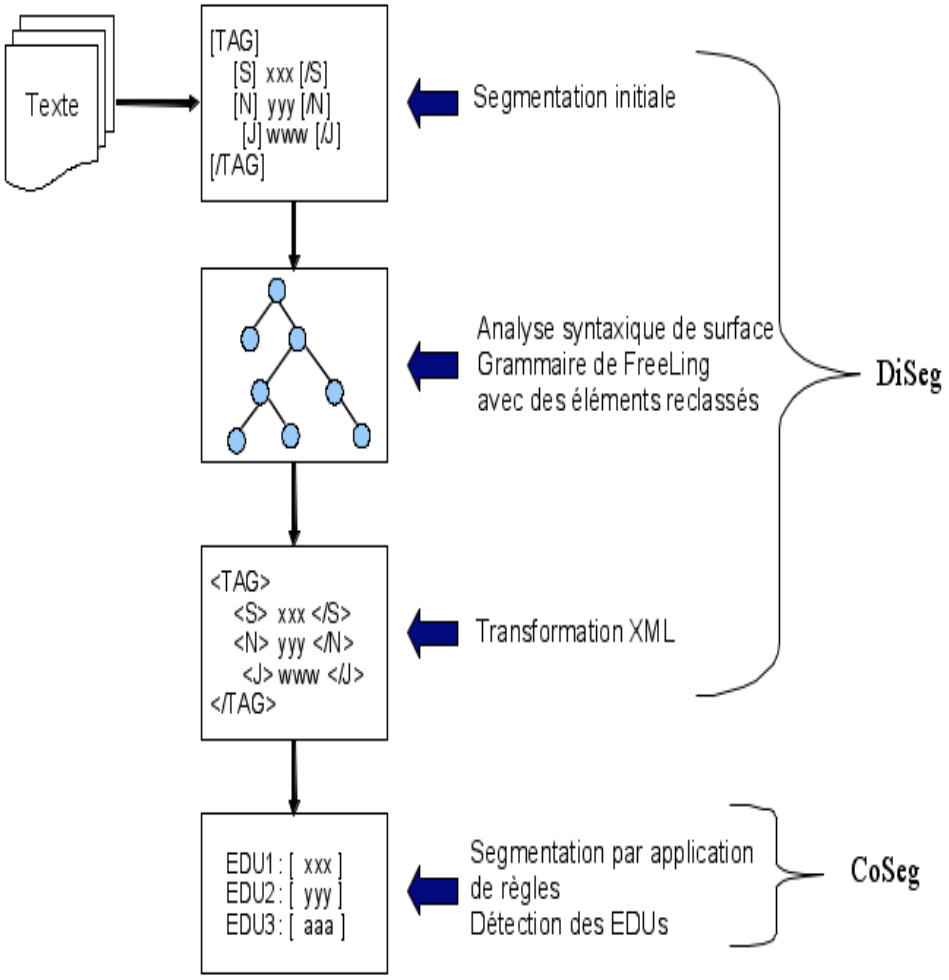


FIGURE 2.4 – *Architecture d'un segmenteur discursif pour des phrases compressées en espagnol : CoSeg.*

règles de manière progressive. La première colonne présente en noir la proportion des EDUs reconnues en supprimant la restriction verbale. La deuxième présente la proportion conjointe de la suppression de la restriction verbale et l'ajout des nouveaux marqueurs. La dernière colonne correspond à la totalité des modifications qui conforment CoSeg : la suppression de la restriction verbale, l'ajout des nouveaux marqueurs et l'ajout des signes de ponctuation. On peut observer que le volume des fragments éliminés couverts par CoSeg augmente jusqu'à 80%.

2.7 Vers la segmentation automatique multilingue

Après la transformation du segmenteur DiSeg en CoSeg, nous nous sommes interrogés sur la possibilité de créer un segmenteur utilisant peu de ressources linguistiques, basé uniquement sur une liste de marqueurs discursifs et un étiqueteur grammatical

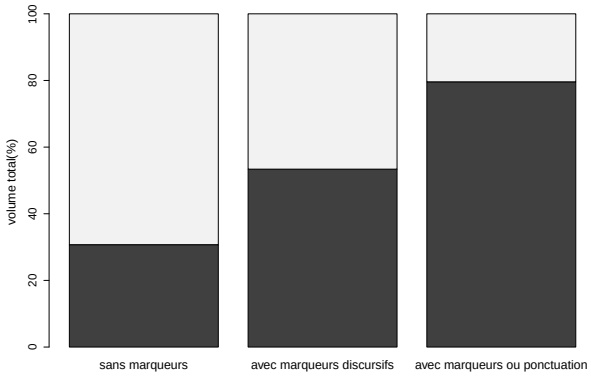


FIGURE 2.5 – *Couverture du segmenteur CoSeg pour 675 fragments (2 651 mots) non reconnus par DiSeg.*

(*part-of-speech tagger*) (Saksik et al., 2013). Ces deux ressources linguistiques existent pour de nombreuses langues. Nous avons conçu une architecture où les listes de marqueurs, en différentes langues, sont des ressources externes (du type *plug-in*) au système de segmentation.

L'architecture d'un segmenteur discursif multi-langue est présentée dans la figure 2.6. Afin d'améliorer la version de base, qui utilise uniquement ces listes, nous avons décidé de combiner la liste avec les étiquettes grammaticales. Pour l'étiquettage, nous avons utilisé l'outil TreeTagger [8], disponible en plusieurs langues : allemand, anglais, français, italien, hollandais, espagnol, bulgare, russe, grec, portugais, galicien, chinois, swahili, latin, estonien et vieux français.

Dans une expérience pilote, nous avons réalisé des tests en Français en utilisant le corpus Annodis (projet ANR ANNOtation DIScursive) issu de la collaboration de trois laboratoires français CLLE-ERSS, IRIT et GREYC. Ce corpus est un ensemble de documents en français segmentés manuellement en unités discursives. Les documents de ce corpus proviennent de quatre sources : l'Est Républicain (39 articles, 10 000 mots) ; Wikipédia (30 articles + 30 extraits, 242 000 mots) ; Actes du Congrès Mondial de Linguistique Française 2008 (25 articles, 169 000 mots) ; Rapports de l'Institut Français de Relations Internationales (32 rapports, 266 000 mots).

Afin de déterminer les capacités de la segmentation discursive en utilisant uniquement la liste des marqueurs et accessoirement l'étiquetage grammatical, nous avons développé trois stratégies décrites ci-dessous.

Le SEGMENTEUR$_\mu$: segmentation par marqueur explicite. Un système de base qui s'appuie uniquement sur une liste de marqueurs discursifs pour réaliser la segmentation. Il remplace l'apparition d'un marqueur dans la liste pour un symbole spécial, par exemple μ, qui indique une frontière entre le segment droit et le segment gauche.

8. http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/

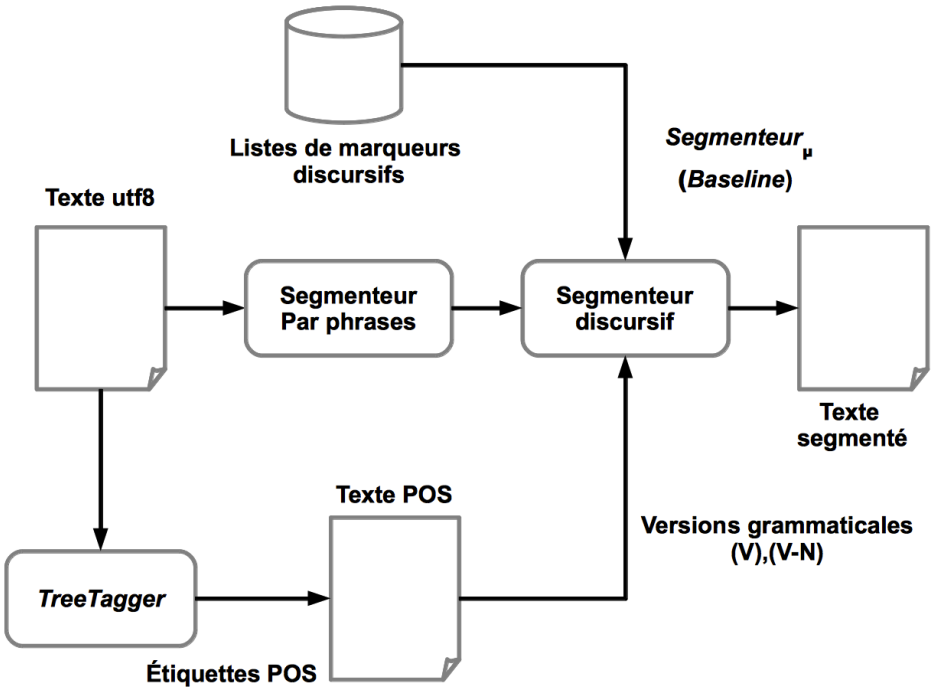


FIGURE 2.6 – *Architecture d'un segmenteur discursif multilingue utilisant peu de ressources linguistiques.*

Le segmenteur de base SEGMENTEUR$_{\mu}$ a été amélioré en suivant deux stratégies différentes.

- SEGMENTEUR$_{\mu+V}$: il s'appuie uniquement sur la présence de verbes à droite et à gauche du marqueur discursif. Les deux règles de cette stratégie sont : S'il n'y a pas de verbes dans les segments droits ou gauches, alors ils sont regroupés en un seul segment. S'il y a au moins un verbe à droite ou à gauche du marqueur, les segments sont séparés.
- SEGMENTEUR$_{\mu+(V-N)}$: il s'appuie sur la présence de verbes et de substantifs. On utilise les mêmes règles que pour le segmenteur précédent mais en considérant la présence des verbes et des substantifs.

En ce qui concerne les évaluations des systèmes, nous utilisons les 78 documents du sous-corpus expert pris comme référence. Nous obtenons pour nos trois systèmes les résultats du tableau 2.5 :

Système	F-score	P	R
SEGMENTEUR$_{\mu}$	0.515	0.649	0.435
SEGMENTEUR$_{\mu+V}$	0.511	0.609	0.448
SEGMENTEUR$_{\mu+(V-N)}$	0.504	0.611	0.437

TABLE 2.5 – *Performances des segmenteurs automatiques*

Le SEGMENTEUR$_{\mu+V}$ présente des performances très proches du système de base (SEGMENTEUR$_{\mu}$) en précision et F-score qui dépassent la *baseline* en rappel. Tout en étant aussi proche de la version verbale, le SEGMENTEUR$_{\mu+(V-N)}$ arrive à obtenir une meilleure précision.

Les résultats montrent que l'on peut construire une version de base très simple, qui emploie uniquement une liste de marqueurs, tout en ayant des performances considérables. La qualité de la liste de marqueurs est un facteur prépondérant pour une segmentation correcte. Nous avons vérifié que même la version de base a donné des résultats intéressants en rappel et F-score.

Les résultats pour le deux sous-corpus d'Annodis (expert et naïf) sont présentés dans le tableau 2.6. Nous constatons qu'ils sont sensiblement les mêmes. Ceci est intéressant, car nous nous attendions à des écarts plus importants. Quoi qu'il en soit, nous pouvons en déduire qu'au moins dans ce corpus il n'est pas nécessaire d'être un expert en linguistique pour segmenter discursivement les documents.

Référence	F-score	*P*	*R*
Experts	0.960	0.983	0.941
Naïfs	0.961	0.971	0.952

TABLE 2.6 – *Performances des segmentations manuelles.*

Ces résultats sont intéressants pour la tâche de compression de phrases car ils nous offrent la possibilité de reproduire les expériences dans d'autres langues. Dans tous les cas, il suffirait uniquement d'avoir à disposition une liste de marqueurs dans chaque langue.

2.8 Conclusions du chapitre

Dans ce chapitre nous avons présenté des travaux récents sur la compression de phrases qui proposent l'élimination de fragments au lieu de mots. Grâce à cet état de l'art et aux résultats du chapitre 1, nous avons remarqué l'utilité de la théorie de la structure rhétorique. Nous avons exploré cette théorie notamment dans la première étape, dans la segmentation discursive.

Nous avons détaillé l'utilité de la segmentation intra-phrase pour la détection des segments éliminables. Nous avons vérifié expérimentalement, en faisant des analyses qualitatives et quantitatives, que les segments discursifs sont très propices à l'élimination. Dans cette perspective, nous avons fait appel au segmenteur discursif pour l'espagnol DiSeg. Pour tester l'efficacité de DiSeg, nous avons fourni un corpus à des annotateurs en leur demandant d'éliminer des fragments textuels. Les résultats ont montré que la moitié des fragments étaient des segments discursifs identifiés automatiquement par DiSeg.

Pour l'autre moitié, celle composée de fragments éliminés mais non identifiés par

DiSeg, nous avons conclu qu'elle se composait essentiellement des fragments ayant les caractéristiques suivantes :

- ils commencent par un participe ;
- ils commencent par un pronom relatif ;
- ils contiennent un marqueur discursif sans inclure un verbe ;
- ils sont entourés par des signes de ponctuation.

Après une analyse approfondie, nous avons proposé une modification du système DiSeg pour couvrir les EDUs qu'il ne reconnaît pas. Nous avons donc créé un segmenteur basé sur DiSeg mais orienté vers la compression de phrases : CoSeg.

Chapitre 3

Pondération de la grammaticalité des phrases compressées

Sommaire

Dans ce chapitre nous abordons la grammaticalité et la problématique qu'elle représente pour la tâche de compression de phrases. Nous cherchons notamment à savoir si une phrase dont quelques segments ont été éliminés reste grammaticale ou pas. Ceci nous mène à nous demander s'il est possible de mesurer le degré de correction grammaticale d'une phrase. Nous expliquons qu'il existe des analyseurs produisant des arbres syntaxiques mais qu'ils ne mesurent pas le degré de grammaticalité. Ainsi, nous montrons l'utilité des modèles de langage probabilistes pour calculer la probabilité d'existence d'une phrase compressée.

3.1 Génération de phrases compressées par élimination de segments discursifs

Dans le chapitre 2 nous avons vu qu'il est intéressant que certains segments discursifs soient éliminés lors de la tâche de compression de phrases. Nous proposons que les phrases compressées soient générées en effaçant les segments discursifs non essentiels de la phrase d'origine φ. Ainsi, l'élimination de segments discursifs donne lieu à un ensemble de candidats à la compression $\{\widetilde{\varphi}_1, \ldots, \widetilde{\varphi}_{2^n}\}$. L'objectif est donc de trouver le meilleur candidat selon les critères d'informativité, de grammaticalité et du taux de compression.

Nous définissons le concept de candidat à la compression de la manière suivante :

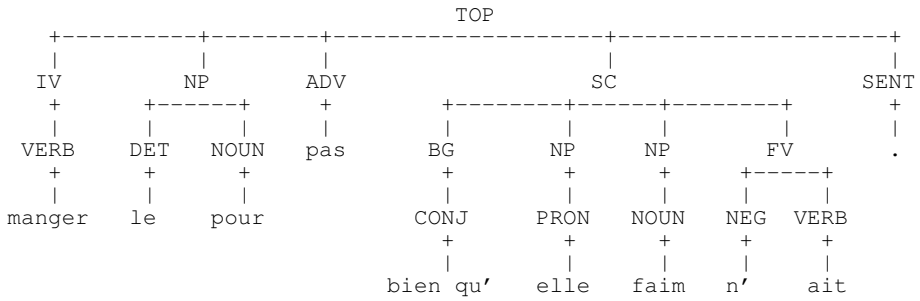
Soit φ une phrase avec n segments : $\varphi = (s_1, s_2, \ldots, s_n)$. Un candidat à la compression $\widetilde{\varphi}$ est une sous-séquence de φ qui respecte l'ordre original des segments. Nous avons deux cas spéciaux : la sous-séquence complète contenant tous les segments et la sous-séquence vide. Dans le premier cas, la séquence correspond à la phrase d'origine φ. Souvent il n'y a aucune version grammaticale plus courte que la phrase elle-même. Ceci est particulièrement vrai pour les phrases courtes ne possédant qu'un seul segment. Le deuxième cas spécial correspond à la séquence vide, qui signifie que la phrase peut être supprimée. Ainsi, rien ne doit être écrit à la place de la phrase d'origine.

La première étape pour générer des phrases compressées est la segmentation de la phrase d'origine. Ensuite, on doit décider pour chaque segment s'il fera partie ou pas du candidat à la compression. C'est-à-dire qu'il faudra générer de façon exhaustive toutes les permutations possibles des segments. Le tableau 3.2 montre un exemple de tous les candidats à la compression pour une phrase avec trois segments : [Juliette prépare un gâteau,]$_{s_1}$ [pour le manger,]$_{s_2}$ [bien qu'elle n'ait pas faim.]$_{s_3}$.

Si l'on devait décider lequel de ces candidats à la compression représente le mieux la phrase d'origine, probablement personne ne choisirait le candidat $\widetilde{\varphi}_4$, puisqu'il s'agit d'une phrase agrammaticale. Mais comment déterminer automatiquement si une phrase est grammaticalement correcte ? C'est justement l'un des défis les plus difficiles du *Trai-*

$\widetilde{\varphi}_1 = (s_1, s_2, s_3)$ =	Juliette prépare un gâteau, pour le manger, bien qu'elle n'ait pas faim.
$\widetilde{\varphi}_2 = (s_1, s_3)$ =	Juliette prépare un gâteau, bien qu'elle n'ait pas faim.
$\widetilde{\varphi}_3 = (s_1, s_2)$ =	Juliette prépare un gâteau, pour le manger.
$\widetilde{\varphi}_4 = (s_2, s_3)$ =	Pour le manger, bien qu'elle n'ait pas faim.
$\widetilde{\varphi}_5 = (s_1)$ =	Juliette prépare un gâteau.
$\widetilde{\varphi}_6 = (s_2)$ =	Pour le manger.
$\widetilde{\varphi}_7 = (s_3)$ =	Bien qu'elle n'ait pas faim.
$\widetilde{\varphi}_8 = ()$ =	

TABLE 3.2 – *Exemple de candidats à la compression pour la phrase « Juliette prépare un gâteau, pour le manger, bien qu'elle n'ait pas faim. ».*




```
SUBJ(ait,elle)
OBJ(manger,pour)
OBJ(ait,faim)
VMOD_POSIT1(manger,pas)
VMOD_POSIT1_SUBORD(manger,ait)
VMOD(ait,manger)
DETERM_DEF(pour,le)
CONNECT(ait,bien qu')
NEGAT(ait)

0>TOP{IV{manger} NP{le pour} pas SC{BG{bien qu'} NP{elle} NP{faim} FV{n' ait}} .}
```

FIGURE 3.1 – *Arbre syntaxique correspondant à une phrase agrammaticale.*

tement Automatique de la Langue Naturelle. Il suffit, par exemple, d'analyser les traductions des systèmes automatiques pour vérifier qu'effectivement traiter la grammaticalité au niveau de la phrase n'est pas trivial. Cependant, des avancées ont été faites pour répondre à cette problématique. Dans la section suivante (3.2) nous décrivons une méthode basée sur des modèles probabilistes qui permet de mesurer la grammaticalité des phrases.

3.2 Les modèles de langage probabilistes

Actuellement, il existe beaucoup d'analyseurs syntaxiques en diverses langues (*parsers*). La plupart d'entre eux produisent des arbres syntaxiques à partir d'une séquence textuelle donnée. Cependant, ces outils ne quantifient pas la grammaticalité. En conséquence, on obtient un arbre syntaxique à partir de n'importe quelle phrase, même si celle-ci est grammaticalement incorrecte. Calculer une probabilité pour mesurer la validité de cet arbre n'est pas faisable par ces systèmes. Par exemple, l'arbre syntaxique montré dans la figure 3.1 a été obtenu avec le *Xerox Incremental Parser*[1] (réputé d'être très performant) et correspond à la phrase agrammaticale « *manger le pour pas bien qu'elle faim n'ait* » :

Ainsi, même si la phrase est agrammaticale, nous avons pu produire son arbre syntaxique. Une expérience similaire est décrite dans (Jonnalagadda et Gonzalez, 2010) pour la phrase agrammaticale « *He is an amazing* ». En utilisant un analyseur syntaxique statistique (Klein et Manning, 2003) et un analyseur syntaxique relationnel (*link*

1. http://open.xerox.com/Services/XIPParser/Consume/64

parser) (Sleator et Temperley, 1995), les auteurs se heurtent au même obstacle. Un analyseur syntaxique relationnel a correctement trouvé des relations incomplètes, laissant de côté le déterminant « *an* » dans la phrase d'exemple. Néanmoins, cet analyseur est incapable d'identifier tous les cas corrects d'un corpus avec 1 100 phrases. Les résultats ont donné plus de 33% de faux négatifs (des phrases agrammaticales prises par grammaticales) car ces analyseurs favorisent la robustesse face aux phrases grammaticalement pauvres. Les auteurs étudient également l'évaluation de la performance de ces analyseurs en utilisant des corpora avec de phrases grammaticalement correctes. Ces résultats ne sont pas favorables dans le contexte de la compression de phrases ou de la simplification syntaxique où la distinction entre phrases grammaticales et phrases agrammaticales est un point critique. (Siddharthan, 2006) étudie la difficulté à déterminer automatiquement le degré de grammaticalité des phrases simplifiées, mais il finit par faire une évaluation manuelle.

Une alternative pour quantifier la grammaticalité d'une phrase sont les modèles de langage probabilistes (Chen et Goodman, 1999; Manning et Schütze, 1999). Ces modèles permettent de calculer la probabilité d'une séquence de mots à partir des occurrences de n-grammes dans un corpus. En général, pour une phrase $\varphi = (w_1, w_2, \ldots, w_n)$, un modèle de langage probabiliste calcule la probabilité de φ en tant que séquence de mots : $\mathbf{P}(\varphi) = \mathbf{P}(w_1, w_2, \ldots, w_n)$.

Supposons, par exemple, que nous voulons calculer la probabilité de la phrase « *Juliette prépare un gâteau pour le manger* ». Nous pouvons le faire en utilisant des probabilités conditionnelles comme suit :

$$\begin{aligned}\mathbf{P}(\text{Juliette, prépare, un, gâteau, pour, le, manger}) = \mathbf{P}(\text{Juliette}) \times \\ \mathbf{P}(\text{prépare} \mid \text{Juliette}) \times \\ \mathbf{P}(\text{un} \mid \text{Juliette prépare}) \times \\ \mathbf{P}(\text{gâteau} \mid \text{Juliette prépare un}) \times \\ \mathbf{P}(\text{pour} \mid \text{Juliette prépare un gâteau}) \times \\ \mathbf{P}(\text{le} \mid \text{Juliette prépare un gâteau pour}) \times \\ \mathbf{P}(\text{manger} \mid \text{Juliette prépare un gâteau pour le})\end{aligned}$$

La probabilité de φ est estimée par l'équation (3.1) où $w_i^j = (w_i, \ldots, w_j)$ est la sous-séquence de mots de w_i à w_j.

$$\mathbf{P}(w_1^n) = \mathbf{P}(w_1) \times \mathbf{P}(w_2|w_1) \times \mathbf{P}(w_3|w_1^2) \times \ldots \times \mathbf{P}(w_n|w_1^{n-1}). \tag{3.1}$$

Nous avons utilisé un modèle de langage probabiliste pour quantifier la grammaticalité des phrases compressées par élimination de segments discursifs (Molina et al., 2012). L'idée principale est d'identifier les candidats dont la grammaire a été perturbée.

En général, les phrases compressées grammaticalement correctes produisent des valeurs de probabilité supérieures que les phrases agrammaticales. Dans nos expériences, nous avons utilisé le corpus Google Web de $1T$ mots [2] pour calculer les probabilités, ainsi qu'une interpolation du modèle basée sur le lissage de Jelinek-Mercer (Chen et Goodman, 1999) en suivant les recommandations de configuration optimale décrites sur le site du logiciel *Language Modeling Toolkit SRILM* [3](Stolcke, 2002). Dans l'équation (3.2), le maximum de vraisemblance estimé pour une phrase est interpolé avec la distribution lissée afin de considérer les n-grammes non existants dans le corpus. Dans la pratique, les estimations du logiciel SRILM sont efectuées dans l'espace log afin d'eviter des erreurs numériques (*u*nderflow) et parce que la somme est normalement plus rapide que la multiplication. Pour une séquence w_1^n, on obtient donc la valeur de l'équation (3.3) à la sortie du système.

$$\begin{aligned}\mathbf{P}_{\text{interp}}(w_i|w_{i-n+1}^{i-1}) &= \lambda_{w_{i-n+1}^{i-1}}\mathbf{P}_{\text{max likelihood}}(w_i|w_{i-n+1}^{i-1}) + \\ &\quad (1-\lambda_{w_{i-n+1}^{i-1}})\mathbf{P}_{\text{interp}}(w_i|w_{i-n+2}^{i-1})\end{aligned} \tag{3.2}$$

$$\mathbf{LP}(w_1^n) = \mathbf{log}(\mathbf{P}(w_1)) + \mathbf{log}(\mathbf{P}(w_2|w_1)) + \ldots + \mathbf{log}(\mathbf{P}(w_n|w_1^{n-1})) \tag{3.3}$$

L'avantage principal d'utiliser un modèle de langage avec un corpus multilingue est que la vérification d'une phrase dans différentes langues se fait en modifiant uniquement un argument. Ceci rend les résultats de cette étude transposables à d'autres langues que l'espagnol. Néanmoins, en termes pratiques, le traitement du corpus Google peut avoir besoin de beaucoup de ressources informatiques et la qualité de ses n-grammes est très faible car ils ont été extraits à partir du Web (y compris la publicité, le texte des hyper-liens, les entêtes, les pieds de page et d'autres contenus indésirables). Nous avons decidé de faire des expériences en utilisant un corpus plus représentatif de l'espagnol écrit (Lara et al., 1979). En utilisant 15 000 phrases de ce corpus nous obtenons l'équation (3.4) qui décrit la relation entre le nombre de mots et la probabilité des phrases dans le modèle de langage obtenu.

$$\mathbf{LP}(w_1^n) = -2.6n - 2.4 \tag{3.4}$$

L'équation (3.4) indique, grosso modo, que pour l'espagnol, le logarithme de la probabilité d'une phrase avec n mots diminue en raison de 2.6 n. Ainsi, le score (3.5) prend en compte la distance entre la courbe empirique de l'équation (3.4) et la probabilité d'une phrase dans le corpus cité. Ce qui permet de considérer la longueur de façon implicite. Toutefois, le score (3.5) est adapté uniquement à l'espagnol. Afin que les résultats de notre travail soient généraux et exploitables pour d'autres langues, nous avons

2. Le corpus Google Web 1T 5-gram, 10 European Languages Version 1. LDC Catalog No. : LDC2009T25

3. www.speech.sri.com/projects/srilm/download.html

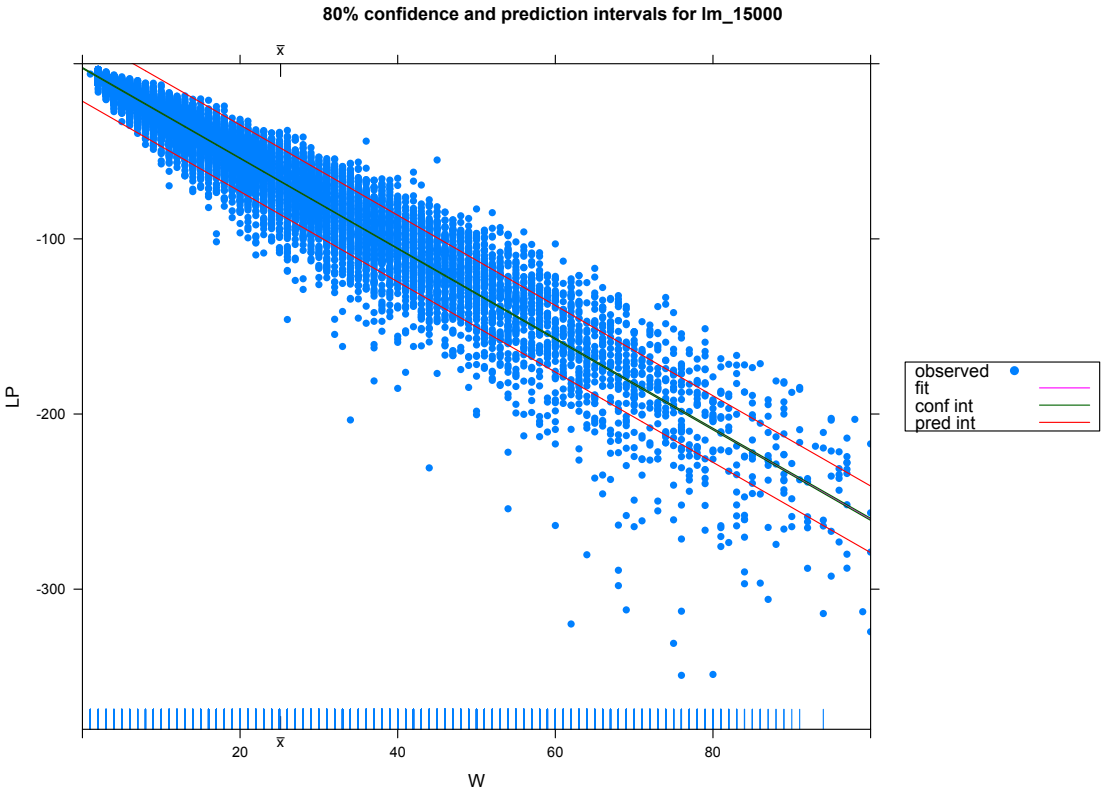


FIGURE 3.2 – *Relation entre le nombre de mots (W) et la probabilité (LP où Log Prob) dans un modèle de langue avec 15 000 phrases en espagnol.*

utilisé plutôt l'équation 3.3 dans nos expériences et nous avons ajouté les composantes de longueur de façon explicite dans le modèle final.

$$\textbf{Score}_{\text{implicit dist}}(w_1^n) = \frac{-2.6 \times \textbf{Lon}(w_1^n) - \textbf{LP}(w_1^n) - 2.4}{\sqrt{-2.6^2 + 1}} \tag{3.5}$$

3.3 Évaluation de la grammaticalité de phrases compressées basée sur des modèles de langage probabilistes

Afin de vérifier l'utilité des modèles de langage probabilistes, nous avons compressé les phrases du corpus (Molina et al., 2011) avec trois systèmes différents et nous avons demandé à un annotateur expert de réaliser cette même action. Nous avons réutilisé ce corpus pour comparer la grammaticalité des phrases compressées automatiquement et celle des phrases compressées manuellement. Celui-ci se compose de quatre genres : sections Wikipédia, nouvelles journalistiques brèves, résumés d'articles scientifiques et littéraire (contes et nouvelles courtes). Ces textes ont été fournis à l'annotateur, auquel

nous avons demandé de les lire, puis de les compresser, phrase par phrase en suivant les même instructions de la section 2.4 :

- ne pas réécrire les phrases ;
- ne pas modifier l'ordre des mots ;
- ne pas substituer des mots ;
- s'assurer que les phrases compressées soient grammaticales ;
- s'assurer que les phrases compressées contiennent le même sens que celui de la phrase d'origine ;
- s'assurer que le document conserve son sens d'origine.

Les trois systèmes que nous avons développés se basent sur des valeurs de probabilité obtenues à partir du modèle de langage probabiliste. Ces systèmes sont :

- **Tous en compétition** (*SysTous*) : $\widetilde{\varphi^{\star}}_i$ correspond au $\widetilde{\varphi}_i$ ayant obtenu le plus haut score de grammaticalité parmi tous les $\widetilde{\varphi}_i$ de la phrase φ.
- **Contenant le premier segment en compétition** (*SysPremier*) : $\widetilde{\varphi^{\star}}_i$ correspond au $\widetilde{\varphi}_i$ ayant obtenu le score de grammaticalité le plus haut parmi les $\widetilde{\varphi}_i$ qui contiennent le premier segment.
- **Aléatoire** (*SysAleatoire*) : Un système où l'on élimine des mots dans la phrase de façon aléatoire tout en respectant le même taux de compression que les humains.

Dans le système *SysTous*, on considère que la position des segments n'est pas déterminante pour la compression de phrases. Le système *SysPremier* a été utilisé car nos résultats dans (Molina et al., 2011) montrent que souvent, les annotateurs gardent le premier segment lors de la compression. Nous avons développé le système de compression aléatoire *SysAleatoire* afin de comparer les résultats des deux premiers contre ceux d'un système sans stratégie.

Nous avons généré les phrases compressées de façon automatique. D'abord, nous avons segmenté les phrases en utilisant le segmenteur discursif DiSeg (chapitre 2). Puis nous avons généré tous les candidats à la compression $(\widetilde{\varphi}_1, \ldots, \widetilde{\varphi}_{2^n})$ (section 3.1) pour chaque phrase de chaque document. Finalement, nous avons substitué la phrase d'origine par la compression avec le meilleur score de grammaticalité en suivant les stratégies mentionnées auparavant. Le texte obtenu qui contient les phrases compressées est ce que nous nommons un résumé par compression de phrases.

Les résultats d'une analyse qualitative (réalisée par trois linguistes) des séquences obtenues par les différentes méthodes sont montrés dans le tableau 3.3. La première colonne correspond au taux de compression moyen, le volume restant après la compression. Il est important de noter que les juges ne connaissaient pas le texte avant d'avoir été compressé. Ils ont évalué la cohérence globale du résumé (deuxième colonne), en assignant la valeur -1 aux résumés sans cohérence, 0 aux résumés où ils trouvent quelques problèmes de cohérence et +1 aux productions cohérentes. De plus, ils ont annoté pour les phrases contenues dans le résumé final si elles sont grammaticalement correctes en considérant leur contexte (troisième colonne). Ces analyses montrent qu'un modèle de langage probabiliste est capable de détecter les phrases compressées dont la grammaticalité a été dégradée. Si l'on considère l'annotateur (*Humain*) et le système *SysAleatoire*

comme points de repère, on observe que, même si les systèmes *SysPremier* et *SysTous* produisent parfois des phrases agrammaticales, leurs performances sont tout de même assez élevées. En effet, le système *SysTous* produit 8,12% de phrases agrammaticales tandis que le système *SysPremier* uniquement 6,98%. Les résultats confirment notre intuition initiale selon laquelle le système *SysPremier* a une tendance à préserver le sujet principal de la phrase dans la plupart des cas. Ainsi, l'introduction de cette simple heuristique améliore la qualité grammaticale des résumés par compression.

résumeur	Taux de compression moyenne (%)	Note de cohérence	% Compressions agrammaticales
SysTous	30.50	+0.37	8.12
SysPremier	18.80	+0.62	6.98
SysAleatoire	22.97	-0.50	76.60
Humain	22.34	+1.00	0.00

TABLE 3.3 – *Résultats de l'évaluation manuelle de la grammaticalité par trois juges.*

En raison des résultats d'évaluation de phrases compressées obtenus avec le score ROUGE (voir chapitre 1), nous avons utilisé le score automatique FRESA [4] (Torres-Moreno et al., 2010b; Saggion et al., 2010) pour vérifier si celui-ci est sensible à la dégradation de la grammaticalité. FRESA calcule la divergence entre deux distributions de fréquences de n-grammes (F_1 pour 1-grammes, F_2 pour 2-grammes, F_4 pour 4-grammes avec trous et F_M pour la moyenne) entre le résumé à évaluer et le document d'origine. Les divergences utilisées dans le logiciel correspondent à celle de Kullback-Leibler (KL) Jensen-Shannon (JS) (Louis et Nenkova, 2008). Soit T l'ensemble de termes contenus dans le document d'origine. Pour chaque $t \in T$, C_t^T régressent le nombre d'apparitions de t dans le document d'origine et C_t^S dans le résumé à évaluer.

$$\mathcal{D}(T||S) = \sum_{t \in T} \left| \log\left(\frac{C_t^T}{|T|} + 1\right) - \log\left(\frac{C_t^S}{|S|} + 1\right) \right| \tag{3.6}$$

L'équation 6.4 calcule la différence absolue entre les divergences des distributions (dans l'espace *log*). Les hautes valeurs de FRESA (basse divergence) sont associées à une grande similitude entre le résumé et le texte d'origine tandis que les valeurs basses (haute divergence) impliquent peu de similitude. Le tableau 3.4 montre les valeurs du score FRESA pour nos systèmes et l'annotateur. Dans tous les cas, le texte de référence est celui d'origine. On constate qu'il n'y a pas de tendance claire pour lier ces résultats en termes de la grammaticalité des phrases vérifiées manuellement (tableau 3.3). Comme attendu, nos résultats confirment que les scores automatiques ROUGE, BLUEU et FRESA ne considèrent pas la grammaticalité des résumés. Le développement d'un nouveau score d'évaluation qui prenne en compte la grammaticalité s'avère donc nécessaire.

4. http://lia.univ-avignon.fr/fileadmin/axes/TALNE

résumeur	F_1	F_2	F_4	F_M
SysTous	0.9197	0.9124	0.9078	0.9133
SysPremier	0.9461	0.9472	0.94512	0.9461
SysAleatoire	0.9593	0.9536	0.9535	0.9555
Human	0.9460	0.9427	0.9372	0.9420

TABLE 3.4 – *Résultats de l'évaluation avec le système FRESA en utilisant le texte d'origine comme référence.*

3.4 Conclusions du chapitre

Nous avons abordé le concept de grammaticalité ainsi que la problématique qu'elle pose pour la compression de phrases. Afin de savoir si une phrase compressée par élimination de fragments reste grammaticale ou pas, nous avons montré comment mesurer le degré de grammaticalité en utilisant les modèles de langage probabilistes. En effet, les résultats de la grammaticalité des phrases compressées ont montré que ces modèles sont sensibles à la perte de grammaticalité de la phrase lors de sa compression.

Nous avons aussi constaté qu'une des caractéristiques décisives pour l'élimination de segments à intraphrase est justement sa position relative à l'intérieur de celle-ci. En comparant des compressions manuelles avec des compression automatiques, nous avons observé que la simple heuristique de garder systématiquement le premier segment discursif d'une phrase réduit significativement la probabilité de générer des phrases agrammaticales.

Quant à l'évaluation, nous avons remarqué que deux métriques automatiques utilisées pour évaluer la qualité des résumés automatiques par extraction s'avèrent inefficaces pour mesurer la grammaticalité des phrases compressées. En comparant les valeurs du score FRESA avec des juges humains, nous avons vérifié expérimentalement qu'aucun de ces scores peut-être utilisé pour évaluer des phrases compressées.

Chapitre 4

Pondération de l'informativité des phrases compressées basée sur l'énergie textuelle

Sommaire

Dans ce chapitre nous abordons l'évaluation de l'informativité des phrases compressées. Nous présentons le modèle de l'énergie textuelle qui permet de repérer quelles informations sont importantes dans une phrase tout en considérant le contexte du document dans son intégralité. Nous expliquons de manière détaillée comment calculer les valeurs d'énergie textuelle à partir d'un document. Nous démontrons que lorsqu'on utilise la façon traditionnelle de calculer l'énergie textuelle, on obtient une distribution très asymétrique de ses valeurs déterminée par l'étendue du vocabulaire. Nous terminons donc en expliquant comment corriger cette distribution grâce aux logarithmes des valeurs d'énergie.

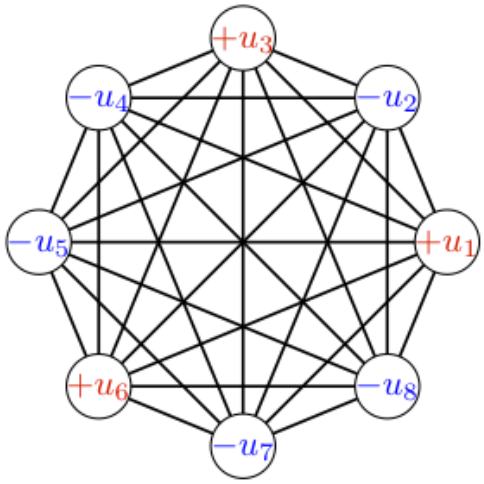


FIGURE 4.1 – *Un exemple de graphe complet avec cinq sommets (K_5) vu comme réseau de Hopfield.*

4.1 Du modèle magnétique d'Ising à l'énergie textuelle

L'énergie textuelle est un modèle inspiré de la physique statistique dans lequel les mots d'un document sont codés comme des spins à deux états : (mots présents = +1, mots absents = -1) qui interagissent les uns avec les autres (Fernández et al., 2007a,b; Fernández, 2009). Selon cette analogie, chaque unité est reliée aux autres de façon directe ou à travers ses voisins. Ceci modélise les caractéristiques d'un système physique. Ainsi, tout comme les électrons dans une barre d'acier, les mots dans un document cohérent peuvent être si fortement liés que le texte résultant est très solide. En revanche, il est possible de trouver des documents où la liaison des concepts ou des mots n'est pas claire et le document finit par nous donner l'impression d'être fragile comme du cristal.

Pour mieux comprendre l'énergie textuelle, nous présentons les idées de base du modèle d'Ising. Dans ce modèle, le moment magnétique d'un matériau est codé par des variables discrètes qui représentent les spins avec les deux états possibles. Les spins sont arrangés dans un graphe de façon à ce que chacun soit lié au reste, permettant l'interaction de tous les sommets (Amit et al., 1987).

En 1982, Hopfield (Hopfield et Tank, 1985) réunit plusieurs idées concernant des réseaux de neurones associatifs inspirées du modèle d'Ising. Dans le réseau de Hopfield, chaque unité a un état d'activation binaire et il existe une connexion entre chaque paire d'unités. La connexion entre deux unités u_i et u_j est décrite par une valeur de $\mathbf{poids}(u_i, u_j)$. Un réseau de Hopfield peut être formellement décrit comme un graphe complet non orienté dont les connexions (arêtes) ont les deux restrictions suivantes :

- il n'y a pas d'auto-connexions (boucles) : $\mathbf{poids}(u_i, u_i) = 0 \; \forall i$;
- les connexions sont symétriques : $\mathbf{poids}(u_i, u_j) = \mathbf{poids}(u_j, u_i) \; \forall i, j$.

Un réseau de Hopfield possédant huit unités est présenté dans la figure 4.1. Dans cette figure, chaque sommet correspond à un neurone et les signes ($+$ ou $-$) repré-

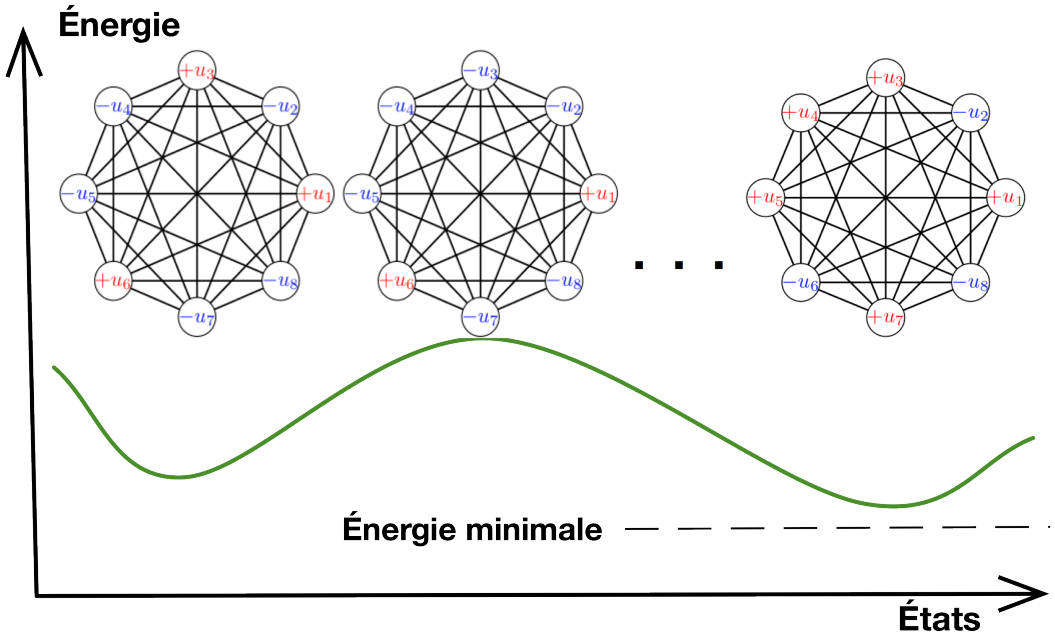


FIGURE 4.2 – *Énergie d'un réseau de Hopfield, mettant en évidence l'état actuel du réseau et un état d'attraction vers lequel il finit par converger avec un niveau d'énergie minimal et un bassin d'attraction.*

sentent les états possibles (spins) de chaque unité. Une application majeure du réseau de Hopfield est la «*mémoire associative*» dont l'idée principale est de fournir des exemples au réseau (soit une partie de l'information, soit de l'information bruitée) afin qu'elle mémorise les régularités et puisse converger vers un état d'équilibre qui permettra de récupérer les motifs appris. Dans l'étape d'apprentissage, les poids des connexions entre les neurones sont mis à jour selon le motif présenté. Les états du réseau sont en fait des attracteurs du système qui essaiera de mémoriser le motif. Dans l'étape d'apprentissage, les poids des connexions entre les neurones sont mis à jour jusqu'à ce que l'état du réseau soit stable. Lors de l'introduction de nouveaux motifs, l'état de la mémoire change de manière drastique mais essaie de se rétablir rapidement. Ces changements soudains peuvent être considérés comme des «*décharges*» énergétiques d'où le nom de la valeur associée à l'état de la mémoire : l'énergie. La figure 4.2 présente la convergence d'un réseau de Hopfield et des changements d'énergie associés à chaque état du réseau.

Les capacités et limitations de la mémoire associative de Hopfield sont établies dans (Hertz et al., 1991). Ce dernier montre que le système sature rapidement et seulement une partie des patrons peut être mémorisée correctement (14%). Cette situation restreint fortement leurs applications pratiques. Cependant, dans le cas du *Traitement Automatique de la Langue Naturelle* (TALN), (Fernández, 2009; Fernández et al., 2008) adaptent le concept de mémoire associative de Hopfield et surtout celui de l'énergie pour proposer «*l'énergie textuelle*» qui tire profit des faiblesses du modèle pour le traitement de textes, notamment dans le cas du résumé automatique et la segmentation thématique.

4.2 L'énergie textuelle et le TALN

Si l'on définit T comme la taille du vocabulaire, où T est le nombre de termes uniques d'un document, on peut représenter une phrase comme une chaîne de T spins. Le terme i est présent ou absent : $\varphi = (\pm 1, \ldots, \pm 1)$, ou de façon équivalente en notation binaire $\varphi = (1, 0, 1, \ldots, 1, 0, 1)$. Ainsi, un document de Φ phrases, est composé de Φ vecteurs dans un espace vectoriel de dimensions $[\Phi \times \mathrm{T}]$ où les phrases sont représentées comme des vecteurs dans cet espace vectoriel. Ils seront donc plus ou moins corrélés, selon les termes qu'ils partagent. Si les thématiques sont proches, il est raisonnable de supposer que le degré de corrélation sera élevé.

Soit $\mathbb{A}$ (4.1) la représentation vectorielle d'un document de Φ phrases et T termes,

$$\mathbb{A} = (a_{i,j})_{;1 \leq i \leq \Phi, 1 \leq j \leq \mathrm{T}} \tag{4.1}$$

où $a_{i,j} = \mathbf{tf}(\varphi_i, w_j)$ est la fréquence du terme w_j dans la phrase φ_i.

Pour calculer les interactions entre les T termes du vocabulaire, la règle de Hebb est appliquée en forme matricielle, ce qui produit l'équation (4.2) comme indiqué dans (Fernández, 2009; Fernández et al., 2008). Cette règle suggère que les connexions sont proportionnelles à la corrélation entre les états des spins. La façon de calculer l'interaction d'échange entre deux unités u_i et u_j proposée par (Hertz et al., 1991) est :

$$\mathbb{J} = \mathbb{A}^{\mathrm{t}} \times \mathbb{A} \tag{4.2}$$

En effet, $j_{i,j} \in \mathbb{J}$ mesure l'interaction entre les termes w_i et w_j par le produit des fréquences d'occurrences simultanées dans les phrases. L'énergie textuelle (4.3) peut alors s'exprimer ainsi :

$$\mathbb{E} = \mathbb{A} \times \mathbb{J} \times \mathbb{A}^{\mathrm{t}} = (\mathbb{A} \times \mathbb{A}^{\mathrm{t}})^2 \tag{4.3}$$

La matrice d'énergie textuelle $\mathbb{E}$ correspondant à la matrice d'adjacence du graphe $G = (\mathbb{A} \times \mathbb{A}^{\mathrm{t}})^2$ qui relie à la fois des phrases ayant des termes communs. La matrice englobe le graphe d'intersection, ainsi que les phrases partageant un même voisinage sans forcément partager le même vocabulaire (Fernández et al., 2009). Cette matrice donne le nombre de chemins de longueur au plus deux entre deux sommets de G. Par rapport à TEXTRANK (Mihalcea, 2004) le calcul de l'énergie est plus simple puisqu'elle se limite aux deux premières itérations, alors que TEXTRANK décrit un processus itératif (de 30 pas approximativement) basé sur le calcul du premier vecteur propre de la matrice de liens entre phrases (Fernández et al., 2009). Il se trouve que d'après les résultats présentés dans (Fernández et al., 2007a), l'énergie textuelle atteint les mêmes performances que TEXTRANK.

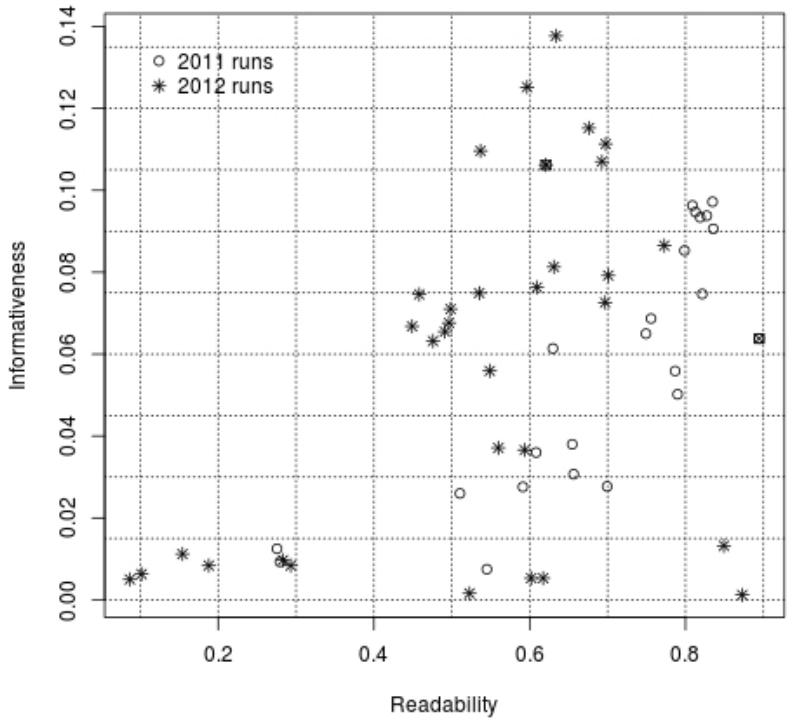


FIGURE 4.3 – *Résultats des évaluations INEX d'informativité et lisibilité dans le track de contextualization de tweets par génération de résumé pour les années 2011 et 2012.*

L'énergie textuelle a été largement utilisée dans le TALN pour mesurer l'informativité textuelle. Ses applications sont variées, elles incluent le regroupement (Molina et al., 2010b; Sierra et al., 2010) ; la détection des frontières thématiques (Labadié et al., 2008) ; le résume mono et multi-document (Fernández et al., 2009) et plus récemment la contextualisation de tweets (Bellot et al., 2013). En fait, cette mesure a été choisie comme référence pour évaluer l'informativité des systèmes participant dans les campagnes INEX (*Initiative for the Evaluation of XML retrieval*) [1]. Les résultats des évaluations d'informativité et de lisibilité des participants du *track* de contextualisation de tweets par génération de résumé pour les années 2011 et 2012 sont montrés dans la figure 4.3. Le carré dans cette figure correspond au système de base utilisant l'énergie textuelle. Par la suite, nous allons nous concentrer plutôt sur les applications de l'énergie textuelle dans le résumé automatique de documents, et plus spécifiquement dans la pondération des EDUs issues de la segmentation intra-phrase. Le lecteur intéressé peut consulter (Torres-Moreno, 2012) pour plus de détails sur les applications de l'énergie textuelle.

1. https://inex.mmci.uni-saarland.de/

4.3 Calcul de l'énergie textuelle pour la compression de phrases

Pour calculer l'énergie textuelle des segments, nous commençons par segmenter les phrases en suivant la méthodologie décrite dans le chapitre 2, puis nous générons deux matrices de représentation vectorielle. La première matrice correspond à la représentation du texte segmenté par phrases, avec une phrase par ligne. La deuxième correspond au texte segmenté par segments discursifs, avec une EDU par ligne.

Considérons, à titre d'exemple, le texte « *Le thermomètre* », extrait de notre corpus (section 5.2) et traduit en français [2]. Chaque phrase entre crochets a été numérotée pour son identification ultérieure :

Le thermomètre (non segmenté)

[Pour savoir la chaleur ou la froideur d'une chose, c'est-à-dire, si on désire connaître la température, il faut utiliser un instrument qui offre une donnée fiable, le thermomètre.]φ_1 [Cet instrument a beaucoup d'emplois dans les foyers, dans les usines et dans les unités de santé.]φ_2 [À la maison il est utile d'avoir un thermomètre pour savoir avec précision si quelqu'un dans la famille a de la fièvre.]φ_3 [Dans l'usine les thermomètres mesurent la température des fours et chaudrons, ainsi que de divers matériaux et substances qui changent durant un processus productif.]φ_4 [Comme on le voit, il est fréquemment nécessaire de mesurer la température de différentes choses, de l'air, du corps humain, d'un four ou de l'eau d'une piscine, ainsi il existe différents types de thermomètres.]φ_5 [Quelque soit le type de thermomètre, dans tous ceux-ci la température se mesure en unités nommées degrés.]φ_6 [Chaque marque de l'instrument est un degré et chaque thermomètre inclue une échelle de mesure qui, généralement, se donne en degrés centigrades.]φ_7

Si l'on considère cette segmentation par phrases, les dimensions de la matrice $\mathbb{A}$(4.1) seront $[7 \times \mathrm{T}]$. En revanche, lorsqu'on considère la segmentation produite par le système DiSeg, décrit dans la section 2.3, le texte de l'exemple est divisé de la façon suivante :

Le thermomètre (segmenté en EDUs DiSeg)

[Pour savoir la chaleur ou la froideur d'une chose, c'est-à-dire,]_s_1 [si on désire connaître la température, il faut utiliser un instrument qui offre une donnée fiable, le thermomètre.]_s_2 [Cet instrument a beaucoup d'emplois dans les foyers, dans les usines et dans les unités de santé.]_s_3 [À la maison il est utile d'avoir un thermomètre]_s_4 [pour savoir avec précision]_s_5 [si quelqu'un dans la famille a de la fièvre.]_s_6 [Dans l'usine les thermomètres mesurent la température des fours et chaudrons, ainsi que de divers matériaux]_s_7 [et substances qui changent durant un processus productif.]_s_8 [Comme on le voit,]_s_9 [il est fréquemment nécessaire de mesurer la température de différentes choses, de l'air, du corps humain, d'un four ou de l'eau d'une piscine,]_s_{10} [ainsi il existe différents types de thermomètres.]_s_{11} [Quelque soit le type de thermomètre, dans tous ceux-ci la température se mesure en unités nommées degrés.]_s_{12} [Chaque marque de l'instrument est un degré]_s_{13} [et chaque thermomètre inclue une échelle de mesure qui, généralement, se donne en degrés centigrades.]_s_{14}

La matrice $\mathbb{A}$ résultante après la segmentation avec DiSeg est de dimensions $[14 \times \mathrm{T}]$ et la matrice d'énergie textuelle $\mathbb{E}$ est donc de $[14 \times 14]$. Le tableau 4.1 présente les valeurs d'énergie textuelle des phrases et de segments DiSeg.

2. La traduction en français a été forcée afin d'avoir le même nombre de segments qu'en espagnol.

De la même manière, en utilisant le segmenteur CoSeg (section 2.6) le texte est segmenté plus finement. En conséquence, les dimensions de sa matrice $\mathbb{A}$ sont $[26 \times \mathrm{T}]$ et celles de sa matrice d'énergie $\mathbb{E}$ sont $[26 \times 26]$. Le tableau 4.2 présente les valeurs d'énergie textuelle des phrases et de segments en utilisant ce segmenteur.

Le thermomètre (segmenté en EDUs CoSeg)

[Pour savoir la chaleur ou la froideur d'une chose,]_s_1 [c'est-à-dire,]_s_2 [si on désire connaître la température,]_s_3 [il faut utiliser un instrument qui offre une donnée fiable,]_s_4 [le thermomètre.]_s_5 [Cet instrument a beaucoup d'emplois dans les foyers,]_s_6 [dans les usines et dans les unités de santé.]_s_7 [À la maison il est utile d'avoir un thermomètre]_s_8 [pour savoir avec précision]_s_9 [si quelqu'un dans la famille a de la fièvre.]_s_{10} [Dans l'usine les thermomètres mesurent la température des fours et chaudrons,]_s_{11} [ainsi que de divers matériaux]_s_{12} [et substances qui changent durant un processus productif.]_s_{13} [Comme on le voit,]_s_{14} [il est fréquemment nécessaire de mesurer la température de différentes choses,]_s_{15} [de l'air,]_s_{16} [du corps humain,]_s_{17} [d'un four]_s_{18} [ou de l'eau d'une piscine,]_s_{19} [ainsi il existe différents types de thermomètres.]_s_{20} [Quelque soit le type de thermomètre,]_s_{21} [dans tous ceux-ci la température se mesure en unités nommées degrés.]_s_{22} [Chaque marque de l'instrument est un degré]_s_{23} [et chaque thermomètre inclue une échelle de mesure qui,]_s_{24} [généralement,]_s_{25} [s'indique en degrés centigrades.]_s_{26}

Nous remarquons que l'énergie textuelle nous permet d'identifier les segments intra-phrase qui ont des informations complémentaires. À titre d'exemple, considérons la phrase la plus énergétique du document (φ_5). Dans le tableau 4.1, nous observons que

Ener(φ)	**Ener**(s_i)	
78	9	[Pour savoir la chaleur ou la froideur d'une chose, c'est-à-dire,]_s_1
78	29	[si on désire connaître la température, il faut utiliser un instrument qui offre une donnée fiable, le thermomètre.]_s_2
36	21	[Cet instrument a beaucoup d'emplois dans les foyers, dans les usines et dans les unités de santé.]_s_3
60	13	[À la maison il est utile d'avoir un thermomètre]_s_4
60	0	[pour savoir avec précision]_s_5
60	14	[si quelqu'un dans la famille a de la fièvre.]_s_6
65	29	[Dans l'usine les thermomètres mesurent la température des fours et chaudrons, ainsi que de divers matériaux]_s_7
65	4	[et substances qui changent durant un processus productif.]_s_8
*119	4	[Comme on le voit,]_s_9
119	35	[il est fréquemment nécessaire de mesurer la température de différentes choses, de l'air, du corps humain, d'un four ou de l'eau d'une piscine,]_s_{10}
119	18	[ainsi il existe différents types de thermomètres.]_s_{11}
49	33	[Quelque soit le type de thermomètre, dans tous ceux-ci la température se mesure en unités nommées degrés.]_s_{12}
66	11	[Chaque marque de l'instrument est un degré]_s_{13}
66	26	[et chaque thermomètre inclue une échelle de mesure qui, généralement, se donne en degrés centigrades.]_s_{14}

TABLE 4.1 – *Exemple de valeurs d'énergie textuelle de segments DiSeg.*

même si la phrase φ_5 est la plus énergétique du document, l'énergie de son premier segment DiSeg est très basse tandis que les valeurs des deux segments suivants sont plus élevées, d'ailleurs le segment avec le plus d'énergie se trouve dans cette phrase.

Maintenant, si nous considérons les valeurs d'énergie de la même phrase avec la segmentation du CoSeg (tableau 4.2), nous constatons qu'il y a six segments dont l'énergie est zéro parmi lesquels deux se trouvent dans la phrase ayant le plus d'énergie.

Nous proposons donc d'utiliser ces valeurs d'énergie textuelle pour mesurer l'informativité des segments et éventuellement pour décider si un segment peut être éliminé ou non. En effet, il est possible de générer des phrases compressées en analysant les valeurs d'énergie textuelle. Concrètement, φ_5 peut être compressée de deux manières différentes selon la segmentation. On peut éliminer les segments de plus basse énergie dans cette phrase :

Ener(φ)	**Ener**(s_i)	
78	3	[Pour savoir la chaleur ou la froideur d'une chose,]_s_1
78	0	[c'est-à-dire,]_s_2
78	2	[si on désire connaître la température,]_s_3
78	3	[il faut utiliser un instrument qui offre une donnée fiable,]_s_4
78	7	[le thermomètre.]_s_5
36	2	[Cet instrument a beaucoup d'emplois dans les foyers,]_s_6
36	10	[dans les usines et dans les unités de santé.]_s_7
60	11	[À la maison il est utile d'avoir un thermomètre]_s_8
60	0	[pour savoir avec précision]_s_9
60	7	[si quelqu'un dans la famille a de la fièvre.]_s_{10}
65	18	[Dans l'usine les thermomètres mesurent la température des fours et chaudrons,]_s_{11}
65	4	[ainsi que de divers matériaux]_s_{12}
65	2	[et substances qui changent durant un processus productif.]_s_{13}
*119	2	[Comme on le voit,]_s_{14}
119	13	[il est fréquemment nécessaire de mesurer la température de différentes choses,]_s_{15}
119	0	[de l'air,]_s_{16}
119	0	[du corps humain,]_s_{17}
119	1	[d'un four]_s_{18}
119	5	[ou de l'eau d'une piscine,]_s_{19}
119	16	[ainsi il existe différents types de thermomètres.]_s_{20}
49	15	[Quelque soit le type de thermomètre,]_s_{21}
19	3	[dans tous ceux-ci la température se mesure en unités nommées degrés.]_s_{22}
66	6	[Chaque marque de l'instrument est un degré]_s_{23}
66	14	[et chaque thermomètre inclue une échelle de mesure qui,]_s_{24}
66	0	[généralement,]_s_{25}
66	0	[s'indique en degrés centigrades.]_s_{26}

TABLE 4.2 – *Exemple de valeurs d'énergie textuelle de segments CoSeg.*

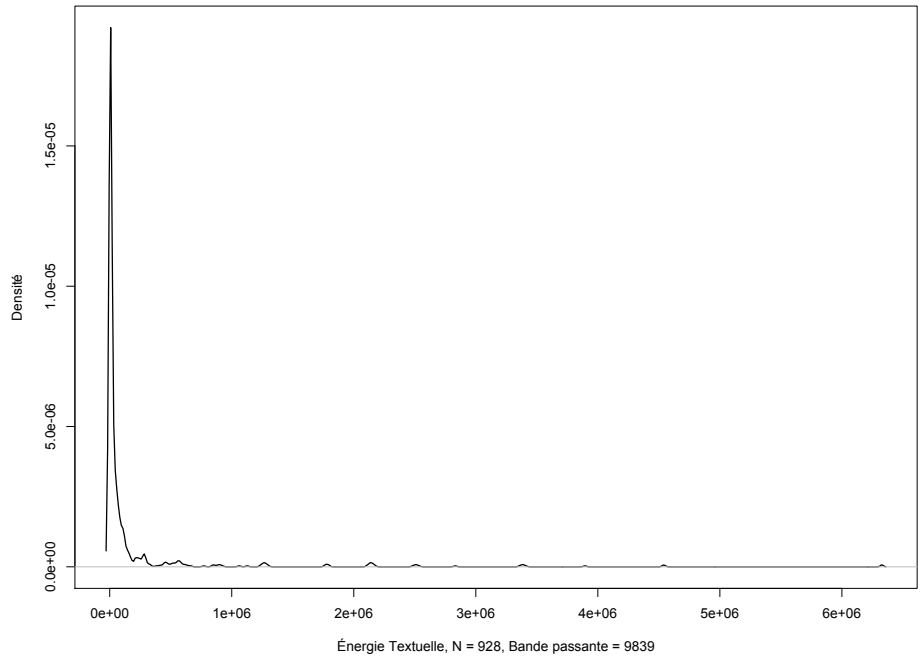


FIGURE 4.4 – *Densité de la distribution des valeur d'énergie textuelle pour des segments discursifs.*

φ_5 : Comme on le voit, il est fréquemment nécessaire de mesurer la température de différentes choses, de l'air, du corps humain, d'un four ou de l'eau d'une piscine, ainsi il existe différents types de thermomètres.

$\widetilde{\varphi^\star}$: Il est fréquemment nécessaire de mesurer la température de différentes choses, de l'air, du corps humain, d'un four ou de l'eau d'une piscine, ainsi il existe différents types de thermomètres.

$\widetilde{\varphi^\star}$:Il est fréquemment nécessaire de mesurer la température de différentes choses, ainsi il existe différents types de thermomètres.

Néanmoins, il est déconseillé d'éliminer un segment uniquement parce qu'il possède une valeur d'énergie négligeable. En effet, comme il a été précisé dans le chapitre 3, nous sommes obligés de préserver aussi la grammaticalité des phrases compressées.

4.4 L'énergie textuelle transformée

Dans la pratique, nous avons trouvé deux inconvénients à l'utilisation de l'énergie textuelle dans sa version originale :

- la distribution des valeurs, présentée dans la figure 4.4, est asymétrique avec une concentration sur les faibles valeurs. Cette distribution est caractéristique d'une loi de puissance ;
- en outre, l'énergie textuelle semble ne pas être bornée.

Afin de résoudre le premier inconvénient, nous avons décidé de faire une analyse en appliquant la transformation Box-Cox, typiquement utilisée pour corriger des distri-

butions asymétriques (Box et Cox, 1964). Cette transformation est définie comme une fonction continue qui varie par rapport à un seul paramètre lambda λ. L'idée est de trouver itérativement la valeur lambda qui corrige le mieux la distribution des valeurs d'énergie textuelle.

Pour appliquer la transformation Box-Cox à l'énergie textuelle, on peut transformer les valeurs $e_{i,j}$ en $e_{i,j}^{\lambda}$ avec l'équation 4.4.

$$e_i^{\lambda} = \begin{cases} K_1(e_i^{\lambda-1}) & \text{si } \lambda \neq 0, \\ K_2 \mathbf{log}(e_i) & \text{si } \lambda = 0 \end{cases} \tag{4.4}$$

où, K_2 correspond à la moyenne géométrique des valeurs $e_1, \cdots, e_n$:

$$K_2 = \left(\prod_{i=1}^{n} e_i \right)^{1/n} \tag{4.5}$$

et K_1 (équation 4.6) dépend du paramètre λ et de la moyenne géométrique :

$$K_1 = \frac{1}{\lambda K_2^{\lambda-1}}. \tag{4.6}$$

Après avoir fait varier la valeur du paramètre λ de 0 à 1 avec un pas de 0.001, nous avons trouvé que la distribution la plus proche de la distribution normale est obtenue avec la valeur $\lambda = 0.0151$. Ainsi, comme suggéré par (Box et Cox, 1964), étant donné que $\lambda \to 0$, on peut corriger la distribution asymétrique des valeurs d'énergie textuelle en utilisant simplement la fonction logarithme. La distribution résultante sera légèrement déviée mais on n'a pas besoin d'introduire d'autres paramètres pour mesurer l'informativité. L'évolution de la distribution de l'énergie textuelle en appliquant la transformation Box-Cox pour différentes valeurs de λ est présentée dans la figure 4.5. Pour souci de clarté nous montrons uniquement les itérations avec un pas de 0.1 à chaque fois.

Les distributions des valeurs d'énergie textuelle (E) et du logarithme de l'énergie (log(E)) sont comparés dans la figure 4.6. La première ligne, correspondant aux fréquences, permet d'observer déjà la grande différence entre les deux distributions. La deuxième ligne, la boîte à moustaches, s'agit d'un rectangle allant du premier au troisième quartile et coupé par la médiane. Les segments aux extrémités du rectangle correspondent aux valeurs extrêmes et les points à des valeurs aberrantes. La troisième ligne, le diagramme Quantile-Quantile, permet d'évaluer la pertinence de l'ajustement des distributions à une loi gaussienne. Si les points sont alignés sur la première bissectrice c'est que la distribution suit probablement ce type de loi. On observe que, même s'il y a des soucis aux extrêmes après la transformation, la distribution semble être bien ajustée à une loi gaussienne, alors que la distribution des valeurs d'énergie textuelle semble suivre une autre loi. Dans la section suivante nous déduirons qu'il s'agit d'une

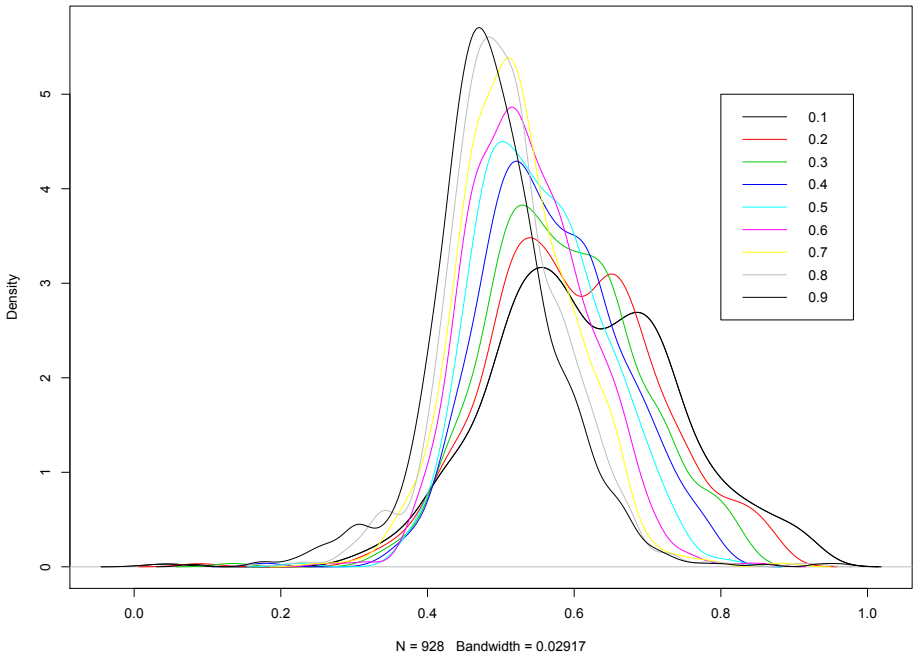


FIGURE 4.5 – *Densité de la distribution des valeurs d'énergie textuelle pour des segments discursifs corrigée par la transformation Box-Cox avec divers valeurs de λ.*

loi exponentielle par rapport à la taille du vocabulaire. En considérant ces résultats, dorénavant, au lieu d'utiliser la valeur d'énergie textuelle d'une phrase φ ou un segment s, $\mathbf{Ener}(\varphi)$ et $\mathbf{Ener}(s)$ respectivement, nous utiliserons les transformations $\mathbf{log}(\mathbf{Ener}(\varphi))$ et $\mathbf{log}(\mathbf{Ener}(s))$.

Les figures 4.7 et 4.8 (à la fin du chapitre) présentent les valeurs d'énergie textuelle des phrases segmentées avec DiSeg et CoSeg respectivement. L'énergie textuelle des phrases a été transformée grâce à la fonction logarithme, elle est montrée dans la première colonne. Celle des segments, transformée de la même manière, est présentée dans la deuxième. Le degré de tonalité de gris est associé aux valeurs de l'énergie transformée dans l'intention de repérer les phrases et les segments les plus énergétiques. Les carrés les plus clairs de la première colonne correspondent aux phrases importantes du document. De la même manière, les carrés clairs de la deuxième colonne correspondent aux segments importants. La densité de la distribution des valeurs d'énergie textuelle pour ce document en particulier est présentée en bas. Remarquons que les valeurs sont bornées grâce à la transformation. La ligne en pointillés indique la valeur d'énergie textuelle transformée (0,5). Cette ligne permet de voir quand est-ce que les valeurs des segments (la ligne continue) se trouvent en dessous ou au dessus de cette valeur.

4.5 Analyse des valeurs maximales de l'énergie textuelle

Pour expliquer la distribution des valeurs d'énergie textuelle dans sa version non transformée, nous allons d'abord étudier le cas le plus simple de la représentation vectorielle d'un document. Il s'agit du cas binaire où dans la matrice $\mathbb{A}$ un terme sera représenté par 1 ou 0 selon s'il est présent ou absent dans la phrase. Nous ne prenons pas en compte la fréquence d'apparition du terme.

Dans ce cas, les valeurs d'énergie textuelle sont maximales si tous les termes du vocabulaire apparaissent dans toutes les phrases du document. Soit $a_{i,j} = 1; \quad \forall a_{i,j} \in \mathbb{A}$, alors :

$$\mathbb{A}_{[\Phi \times \mathrm{T}]} = \begin{pmatrix} a_{1,1} & \cdots & a_{1,\mathrm{T}} \\ \vdots & \ddots & \vdots \\ a_{\Phi,1} & \cdots & a_{\Phi,\mathrm{T}} \end{pmatrix} = \begin{pmatrix} 1 & \cdots & 1 \\ \vdots & \ddots & \vdots \\ 1 & \cdots & 1 \end{pmatrix} \tag{4.7}$$

Chaque élément de $(\mathbb{A} \times \mathbb{A}^{\mathrm{t}})$ est donc de la forme $\underbrace{1^2 + \cdots + 1^2}_{\mathrm{T\ fois}} = \mathrm{T}$:

$$(\mathbb{A} \times \mathbb{A}^{\mathrm{t}})_{[\Phi \times \Phi]} = \begin{pmatrix} \underbrace{1^2 + \cdots + 1^2}_{\mathrm{T}} & \cdots & \underbrace{1^2 + \cdots + 1^2}_{\mathrm{T}} \\ \vdots & \ddots & \vdots \\ \underbrace{1^2 + \cdots + 1^2}_{\mathrm{T}} & \cdots & \underbrace{1^2 + \cdots + 1^2}_{\mathrm{T}} \end{pmatrix} = \begin{pmatrix} \mathrm{T} & \cdots & \mathrm{T} \\ \vdots & \ddots & \vdots \\ \mathrm{T} & \cdots & \mathrm{T} \end{pmatrix} \tag{4.8}$$

Enfin pour l'équation 4.3, les valeurs de l'énergie textuelle correspondent aux éléments de la matrice $(\mathbb{A} \times \mathbb{A}^{\mathrm{t}})^2$ où chaque élément a la forme $\underbrace{\mathrm{T}^2 + \cdots + \mathrm{T}^2}_{\Phi\ \mathrm{fois}} = \Phi\mathrm{T}^2$.

$$(\mathbb{A} \times \mathbb{A}^{\mathrm{t}})^2_{[\Phi \times \Phi]} = \begin{pmatrix} \underbrace{\mathrm{T}^2 + \cdots + \mathrm{T}^2}_{\Phi} & \cdots & \underbrace{\mathrm{T}^2 + \cdots + \mathrm{T}^2}_{\Phi} \\ \vdots & \ddots & \vdots \\ \underbrace{\mathrm{T}^2 + \cdots + \mathrm{T}^2}_{\Phi} & \cdots & \underbrace{\mathrm{T}^2 + \cdots + \mathrm{T}^2}_{\Phi} \end{pmatrix} = \begin{pmatrix} \Phi\mathrm{T}^2 & \cdots & \Phi\mathrm{T}^2 \\ \vdots & \ddots & \vdots \\ \Phi\mathrm{T}^2 & \cdots & \Phi\mathrm{T}^2 \end{pmatrix} \tag{4.9}$$

Ainsi, selon l'équation (4.9), les valeurs d'énergie textuelle $e_{i,j}$ sont de l'ordre $O(\Phi\mathrm{T}^2)$ pour le cas de représentation binaire.

Dans le cas général, soit $\mathbf{tf}(\varphi_i, w_j) = c; \quad \forall i \in [1, \Phi], j \in [1, \mathrm{T}]$; de façon à ce que la valeur c de fréquence associée est maximale pour toutes les phrases et tous les termes.

Par analogie, les éléments de $(\mathbb{A} \times \mathbb{A}^{\mathrm{t}})$ ont la forme $\underbrace{c^2 + \cdots + c^2}_{\mathrm{T\ fois}} = \mathrm{T}c^2$.

Les éléments de la matrice $(\mathbb{A} \times \mathbb{A}^{t})^2$ ont donc la forme $\underbrace{\mathrm{T}^2c^4 + \cdots + \mathrm{T}^2c^4}_{\Phi \text{ fois}}$.

D'où, chaque élément de la matrice d'énergie $e_{i,j}$ est de l'ordre $O(\Phi \mathrm{T}^2 c^4)$.

Bien que la fréquence maximale c soit à la puissance 4, c'est la taille du vocabulaire T qui domine la croissance des valeurs d'énergie textuelle. En effet, dans des textes réels, nous trouvons majoritairement que $0 \leq c \leq 3$, car un terme peut apparaître dans la même phrase une seule fois, rarement deux et quasiment jamais trois fois ou plus. En outre, le nombre de phrases est toujours inférieur ou égal au nombre de termes, c'est-à-dire $\phi \leq \mathrm{T}$.

En conclusion, c'est toujours la taille du vocabulaire T qui domine la distribution des valeurs d'énergie textuelle (figure 4.4). En effet, la richesse lexicale est la raison de la distribution très asymétrique de ces valeurs. Par exemple, pour un texte de 300 mots (la taille d'un résumé) nous pouvons avoir des valeurs d'énergie de l'ordre de $10e + 05$ et pour un de 1000 mots (une page et demie) de l'ordre de $10e + 06$ [3].

4.6 Conclusions du chapitre

Nous venons de détailler le modèle de l'énergie textuelle qui permet de mesurer l'informativité des phrases compressées. Nous avons aussi analysé ses capacité et limitations, ceci nous a permis de proposer une transformation orientée vers la tâche qui nous intéresse. Pour adapter l'énergie textuelle, nous avons appliqué la transformation Box-Cox, typiquement utilisée pour corriger des distributions asymétriques. Nous avons ainsi trouvé que la meilleure distribution des valeurs d'énergie textuelle transformées est obtenue avec une valeur du paramètre λ très proche à zéro. Nous devons donc utiliser la fonction logarithme de l'énergie textuelle dans nos expériences.

Grâce à cette nouvelle mesure d'énergie textuelle, nous avons pu identifier plus facilement les segments à l'intérieur de la phrase qui sont pauvres en informativité. De la même manière, ces valeurs nous ont aidé à distinguer les segments qui contiennent des informations essentielles, tout ceci en considérant le contexte du document en entier.

Ainsi, nous avons proposé d'éliminer les segments les moins énergétiques à condition que cette élimination ne dégrade pas la grammaticalité. En conséquence, la nouvelle phrase (la version compressée) contient uniquement les segments essentiels, soit au niveau de l'informativité, soit au niveau de la grammaticalité ou les deux.

3. En supposant une grande richesse lexicale dans les deux exemples.

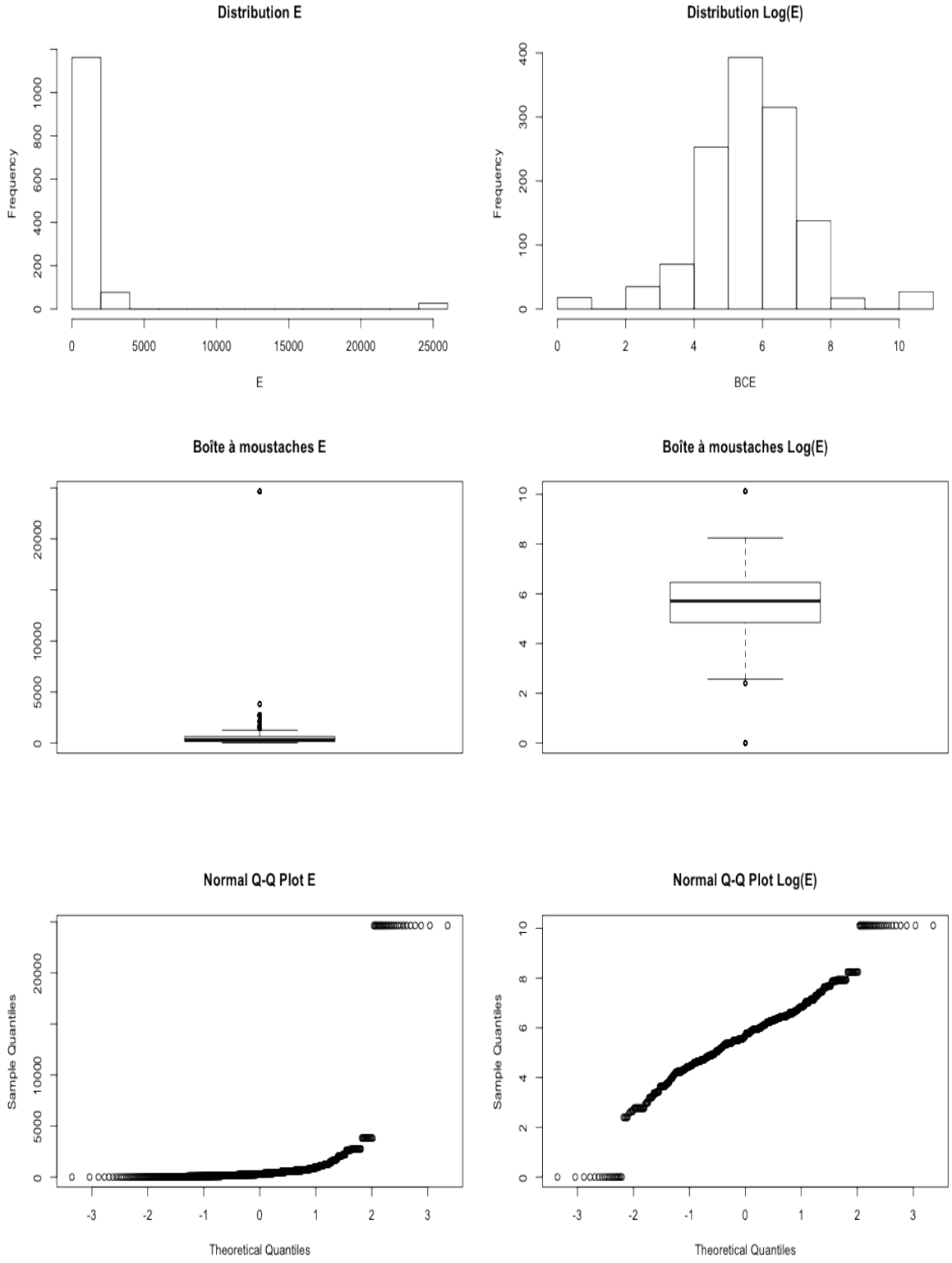


FIGURE 4.6 – *Comparaison entre l'énergie textuelle et l'énergie textuelle transformée.*

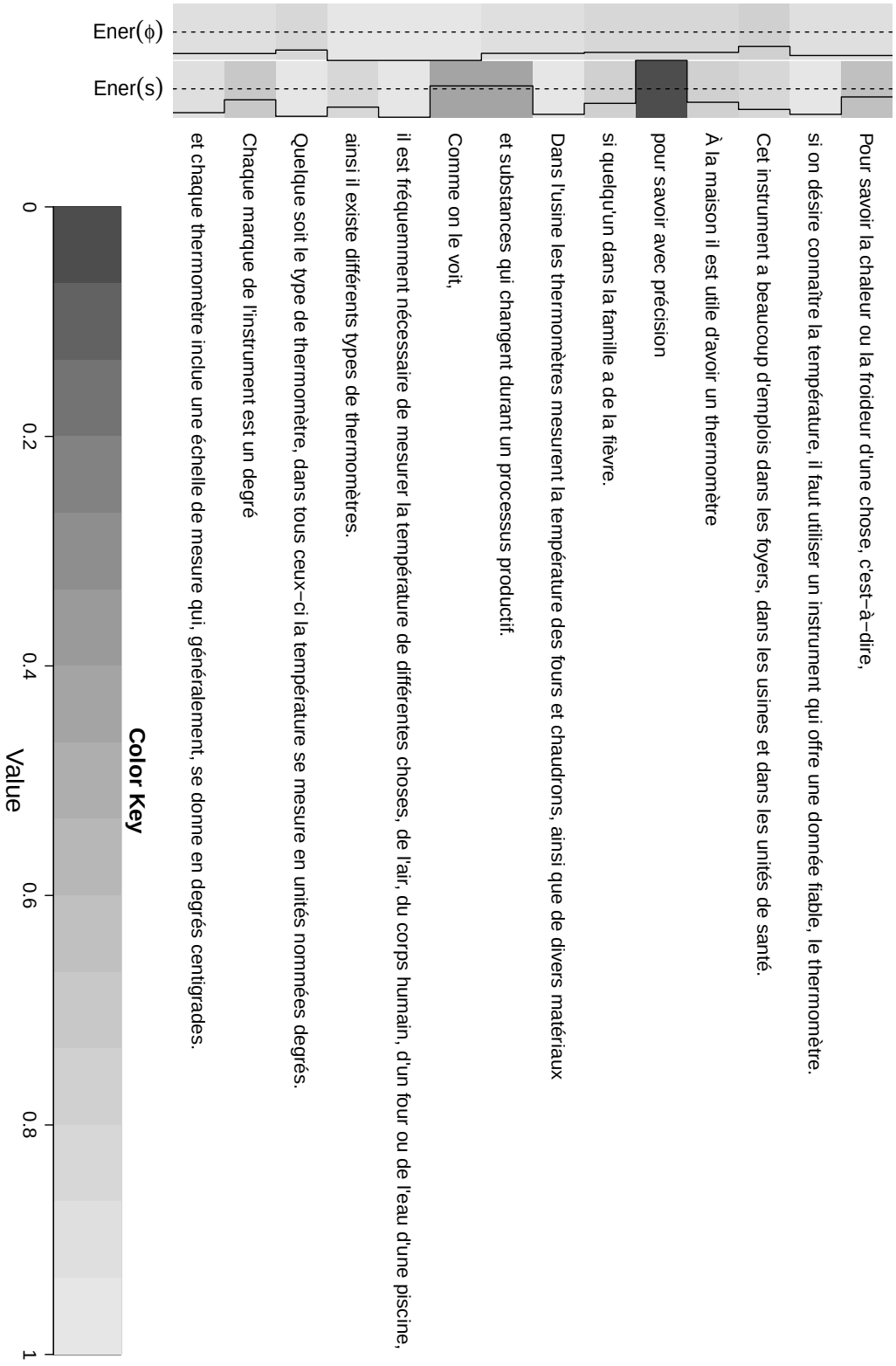


FIGURE 4.7 – *Exemple des valeurs d'énergie textuelle pour des segments DiSeg.*

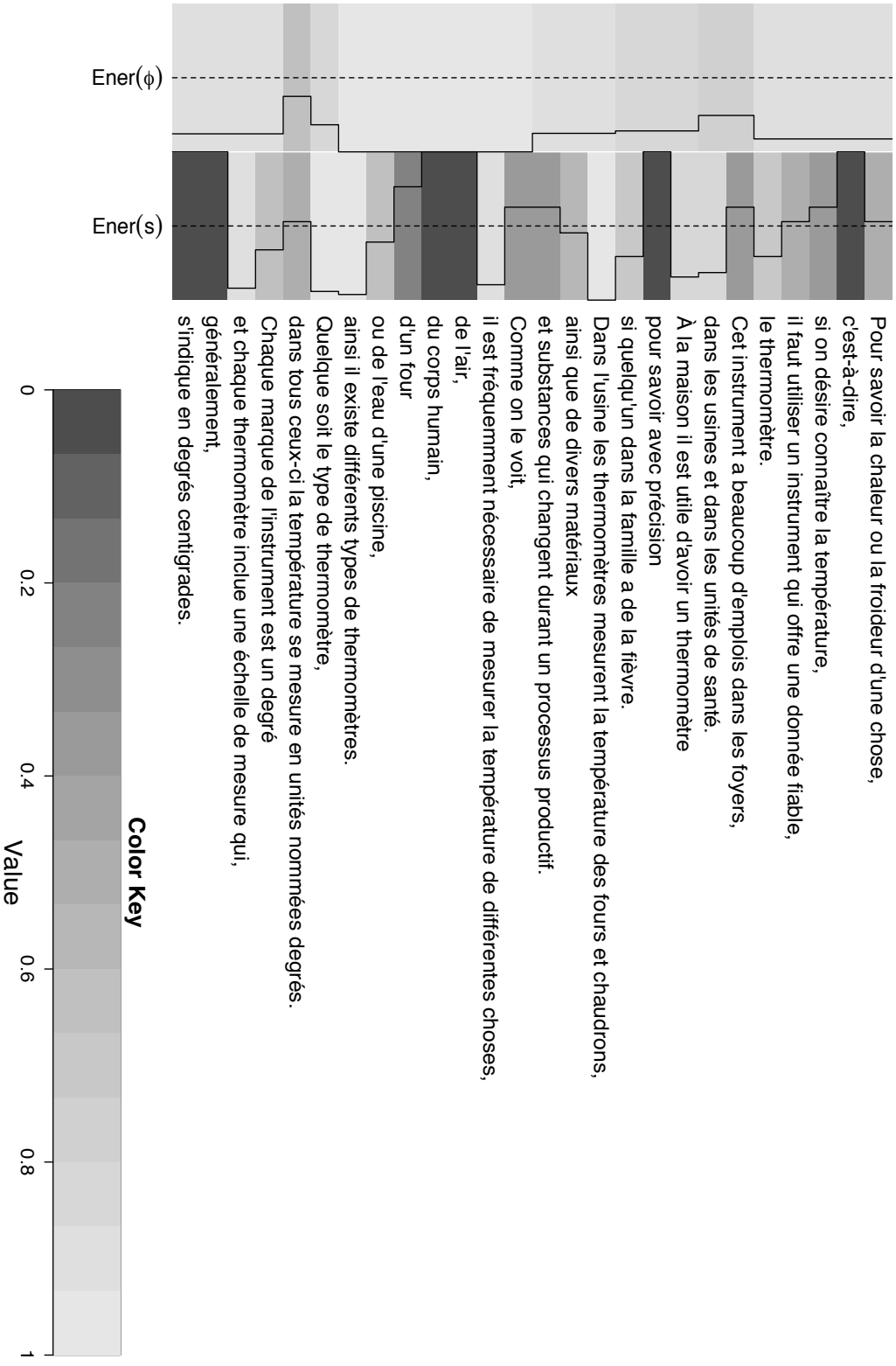


FIGURE 4.8 – *Exemple des valeurs d'énergie textuelle pour des segments CoSeg.*

Chapitre 5

Un modèle probabiliste d'élimination de segments intra-phrase

Sommaire

Dans ce chapitre, nous faisons appel à tous les concepts présentés tout le long de la thèse afin de proposer notre approche de compression de phrases. Nous présentons notre protocole d'annotation du corpus de phrases compressées avec l'aide d'une plate-forme Web. Nous analysons les résultats de l'annotation où nous trouvons un haut degré d'ambiguïté. Nous décrivons un modèle de régression linéaire et la manière dont il peut être utilisé pour la tâche de compression de phrases en nous inspirant du paradigme d'apprentissage supervisé d'Edmundson. Nous proposons une fonction linéaire pour prédire si un segment doit être éliminé sur la base d'environ soixante mille segments annotés. Finalement, nous présentons deux algorithmes de génération de résumé par élimination de segments. Dans le premier, on peut conditionner la probabilité déterminant si un segment doit être éliminé. Dans le deuxième, on peut contrôler le taux de compression souhaité pour le résumé final.

5.1 La compression de phrases est-elle un problème d'optimisation ?

Nous allons proposer une adaptation de notre définition de compression de phrases, plus en accord avec les concepts auparavant vus. Étant donnée la phrase d'origine φ, un algorithme doit trouver une nouvelle version $\widetilde{\varphi^\star}$ devant à la fois :

- être plus courte, ($\mathbf{Seg}(\widetilde{\varphi^\star}) \leq \mathbf{Seg}(\varphi)$) ;
- contenir les informations les plus importantes, ($\mathbf{Ener}(\widetilde{\varphi^\star}) \geq \mathbf{Ener}(\varphi)$) ;
- être grammaticalement correcte, ($\mathbf{LP}(\widetilde{\varphi^\star}) \geq \mathbf{LP}(\varphi)$).

Où $\mathbf{Seg}(\bullet)$ est une fonction qui compte le nombre de segments d'une phrase ; $\mathbf{Ener}(\bullet)$ est une fonction qui mesure l'énergie textuelle d'une phrase ou un segment et $\mathbf{LP}(\bullet)$ est une fonction qui mesure la probabilité d'une séquence dans un modèle de langage probabiliste. Cette définition est cohérente avec celle de la section 1.2 mais adaptée aux concepts utilisés dans notre approche. Elle établit deux suppositions sur les trois aspects : ils sont mesurables et ils sont comparables. Cependant, elle ne suppose pas que la solution $\widetilde{\varphi^\star}$ soit unique. Nous avons donné des exemples de phrases compressées de deux façons différentes. Dans le chapitre 3 nous utilisons la phrase « *Juliette prépare un gâteau pour le manger bien qu'elle n'ait pas faim.* » pour montrer qu'elle a au moins trois compressions grammaticalement correctes :

> φ : Juliette prépare un gâteau pour le manger bien qu'elle n'ait pas faim.
> $\widetilde{\varphi_1^\star}$: Juliette prépare un gâteau bien qu'elle n'ait pas faim.
> $\widetilde{\varphi_2^\star}$: Juliette prépare un gâteau pour le manger.
> $\widetilde{\varphi_3^\star}$: Juliette prépare un gâteau.

Dans le chapitre 4 nous utilisons une phrase de notre corpus pour expliquer qu'il est possible de générer des phrases compressées en analysant les valeurs d'énergie textuelle :

> φ : Comme on le voit, il est fréquemment nécessaire de mesurer la température de différentes choses, de l'air, du corps humain, d'un four ou de l'eau d'une piscine, ainsi il existe différents types de thermomètres.
> $\widetilde{\varphi_1^\star}$: Il est fréquemment nécessaire de mesurer la température de différentes choses, de l'air, du corps humain, d'un four ou de l'eau d'une piscine, ainsi il existe différents types de thermomètres.
> $\widetilde{\varphi_2^\star}$: Il est fréquemment nécessaire de mesurer la température de différentes choses, ainsi il existe différents types de thermomètres.

Est-il donc pertinent de parler d'une seule compression optimale ? En termes de la longueur oui puisque pour une phrase la compression vide est toujours la compression optimale. En termes de grammaticalité non car il peut y avoir plus d'une compression qui soit grammaticalement correcte. Ceci peut toujours être vérifié (même manuellement). En termes d'informativité on ne peut pas parler non plus de compression

optimale. Il peut y avoir plusieurs compressions avec des valeurs d'énergie textuelle très élevées. Par contre, la vérification manuelle de l'informativité reste un problème puisque le concept est subjectif : différentes personnes peuvent avoir différentes opinions.

Même si nous ne pouvons pas transformer la compression de phrases en un problème d'optimisation nous pouvons analyser ce que font les annotateurs en essayant de compresser les phrases d'un document pour former un résumé. Nous proposons de réaliser la tâche autant de fois nécessaires afin qu'il y ait des données suffisantes pour « *apprendre* », avec des méthodes d'apprentissage supervisé, ce que les humains considèrent comme une phrase courte, grammaticale et contenant l'information importante.

5.2 Les sciences citoyennes pour l'annotation de corpus

En raison de l'absence d'études de compression de phrases en espagnol [1] et afin de collecter un grand corpus, nous avons dirigé le protocole d'annotation sous le paradigme des « *sciences citoyennes* », appelées aussi « *sciences participatives* » (Irwin, 1995). Les sciences citoyennes proposent d'engager des volontaires non-experts dans des tâches scientifiques. Bien que ce concept ne soit pas nouveau, au cours de cette dernière décennie beaucoup de chercheurs ont décidé d'en tirer profit. Dans le *Traitement Automatique de la Langue Naturelle*, il représente une alternative efficace pour l'annotation massive de corpus (Molina, 2013). Par exemple, dans le projet décrit par (Chamberlain et al., 2009), les auteurs ont développé un jeu vidéo où l'utilisateur joue le rôle d'un détective pour découvrir un mystère en désambiguïsant les références anaphoriques d'un document.

(Snow et al., 2008) ont étudié la fiabilité des annotations produites par des non-experts. Ils arrivent à la conclusion que les données ainsi obtenues sont aussi fiables que celles produites par des experts, à condition que lors des expériences certains principes soient respectés :

- la description des tâches doit être précise et
- la participation des annotateurs doit se limiter à choisir parmi un nombre limité d'alternatives.

Dans notre protocole d'annotation nous avons suivi ces principes. Nous avons rédigé un « *manuel d'annotation pour le corpus de compression de phrases et le résumé automatique en espagnol* » [2] et nous avons conçu un système qui limitait les annotateurs selon les critères de compression de phrases par élimination de segments discursifs.

Nous avons donc lancé une campagne d'annotation citoyenne avec des volontaires non-experts qui devaient resumer 30 textes courts en éliminant des segments discursifs. Chaque document a été segmenté en utilisant les systèmes DiSeg (section 2.3) et

1. Nous avons construit le premier corpus en cette langue.
2. http://molina.talne.eu/compress4/man/

CoSeg (section 2.6). Nous avons recruté 66 volontaires, la majorité des étudiants de licence, tous des hispanophones natifs. Nous avons demandé aux annotateurs de choisir les segments à conserver pour produire un résumé par compression en respectant les critères suivants :

- **Couverture** : il faut préserver au moins un segment par phrase.
- **Importance** : l'idée principale du texte doit être conservée.
- **Grammaticalité** : les phrases compressées doivent être compréhensibles et ne pas avoir de problèmes de cohérence (par exemple, les phrases doivent avoir un verbe principal).
- **Breveté** : il faut compresser le plus possible. C'est-à-dire, il faut éliminer le maximum de segments afin que la phrase dise la même chose, mais en moins de mots.

Le manuel d'annotation et le système se trouvent à l'adresse `http://molina.talne.eu/compress4/man/`. L'interface d'annotation utilisée ainsi que ses composants est présentée dans la figure 5.1.

1. **Segments** : ce sont des fragments de texte qui peuvent être conservés ou éliminés en cliquant sur eux.
2. **Texte d'origine** : il contient le document initial (les segments peuvent être consultés même après leur élimination).
3. **Texte compressé** : il affiche le texte résultant après l'élimination des segments sélectionnés
4. **Bouton** « *Recommencer texte* » : il rétablit le texte d'origine avant toute suppression.
5. **Bouton** « *Envoyer texte* » : il sauvegarde le texte compressé dans la base de données et affiche le prochain texte à analyser.

L'architecture du système d'annotation possède trois couches : la couche de données, la couche de traitement et celle de présentation. Les aspects les plus avantageux de ce type d'architecture sont la division claire du travail entre le client (un navigateur Web) et le serveur, ainsi que le contrôle de l'annotation du corpus qui est validé par la couche intermédiaire.

Ces trois couches sont représentées par trois fichiers. Un premier fichier HTML, correspondant à la couche de présentation, contient le squelette de l'interface d'utilisateur ainsi que les boutons des commandes. Un script, correspondant à la couche de traitement, contient le code qui gère les événements du navigateur et du DOM (*Document Object Model*). La couche de présentation est la seule à laquelle ont accès les annotateurs. Ils ne peuvent pas modifier directement les textes du corpus, stockés préalablement dans la base de données, ni réaliser des actions non indiquées dans les spécifications d'annotation. Un dernier script, correspondant à la couche de données, contient les codes d'interaction avec une base de données MySQL. Les échanges entre les différentes couches se font grâce à des objets JSON.

Grâce au système d'annotation, nous avons annoté 60 844 fois la décision des annotateurs d'effacer ou conserver un segment. Nous avons collecté, tout le long de dix

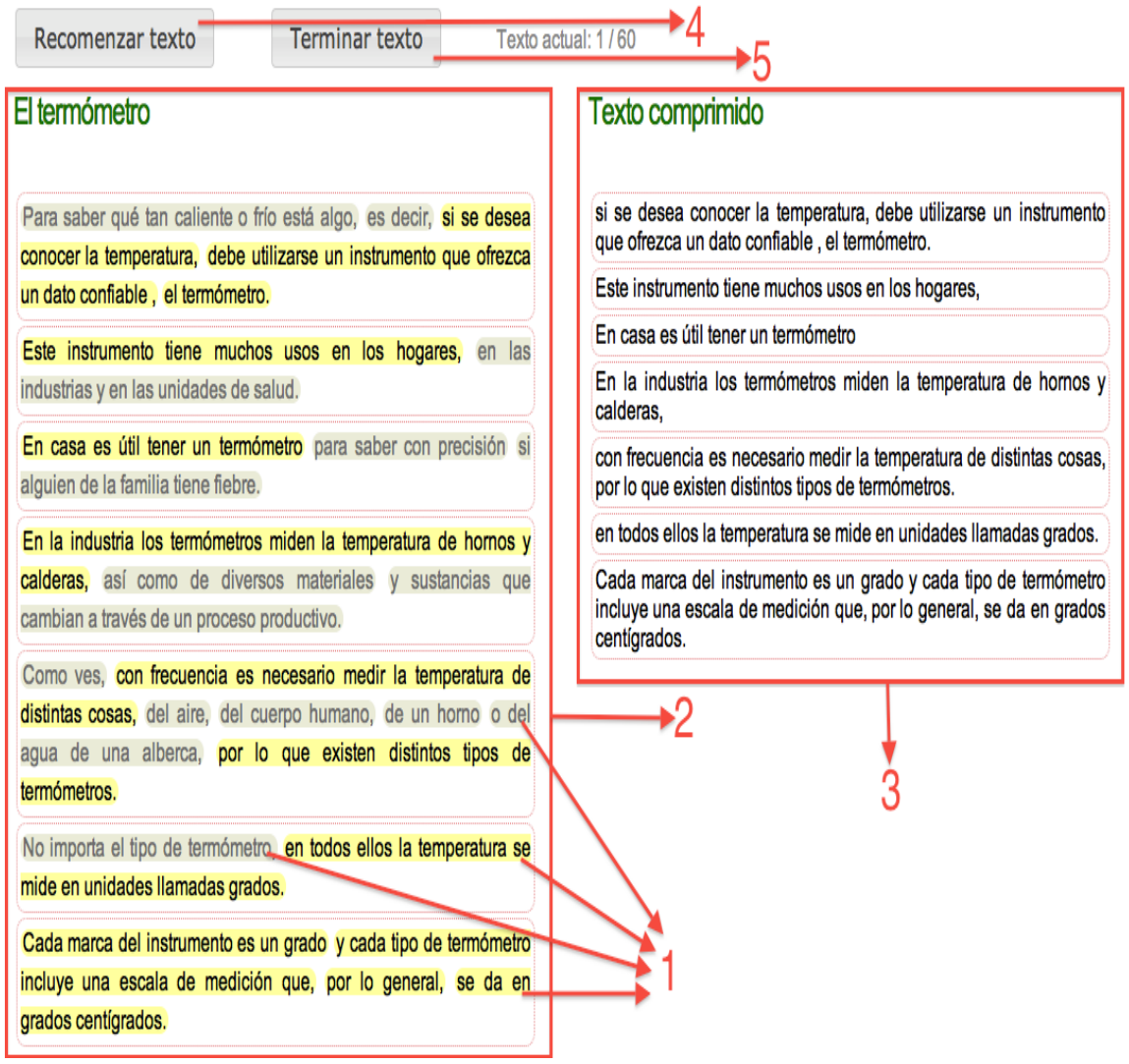


FIGURE 5.1 – *Interface du système d'annotation.*

semaines, 2 877 résumés (en moyenne 48 différents par document). Ces données ont été utilisées pour ajuster une régression linéaire et ainsi prédire l'élimination des segments.

5.3 Accord des annotateurs

La campagne d'annotation nous a permis de collecter un corpus important de textes en espagnol, annotés pour l'élaboration d'un modèle de compression de phrases par élimination d'EDUs. Dans ce corpus les textes ont été préalablement segmentés par DiSeg et CoSeg. Les annotateurs ont ensuite marqué les EDUs qui peuvent être éliminées pour obtenir un résumé du texte d'origine. Chaque phrase est donc associée à un ensemble de compressions proposées.

Les tableaux 5.1 et 5.2 présentent le nombre théorique des compressions possibles et le nombre moyen des compressions proposées par les annotateurs. Ce nombre moyen varie entre 1 et $min(N, 2^n)$ où N est le nombre d'annotateurs. Il est à son maximum lorsque aucun annotateur n'a choisi la même combinaison d'EDUs à éliminer ce qui est pratiquement le cas pour les phrases avec plus de 12 sous segments identifiés.

Nombre de phrases avec n segments	Nombre de segments n	Nombre théorique des compressions 2^n	Nombre moyen des compressions proposées	Nombre total des compressions proposées
74	1	2	1.00	74
76	2	4	2.60	198
37	3	8	4.59	170
15	4	16	7.00	105
5	5	32	10.80	54
1	6	64	16.00	16
1	8	256	16.00	16
1	9	512	26.00	26

TABLE 5.1 – *Nombre théorique des compressions possibles et nombre moyen des compressions proposées par les annotateurs pour DiSeg.*

Nombre de phrases avec n segments	Nombre de segments n	Nombre théorique des compressions 2^n	Nombre moyen des compressions proposées	Nombre total des compressions proposées
22	1	2	1.00	22
56	2	4	2.48	139
35	3	8	4.34	152
31	4	16	6.12	190
25	5	32	8.80	220
15	6	64	11.60	174
7	7	128	12.28	86
6	8	256	14.33	86
6	9	512	21.00	126
1	10	1024	18.00	18
1	11	2048	16.00	16
2	12	4096	28.00	56
2	13	8192	39.00	78
1	21	2097152	40.00	40

TABLE 5.2 – *Nombre théorique des compressions possibles et nombre moyen des compressions proposées par les annotateurs pour CoSeg.*

Le nombre de propositions distinctes est bien inférieur à celui d'une élimination aléatoire des segments. Clairement, si on note n le nombre d'EDUs distinctes d'un texte, alors la dimension de l'espace des solutions choisies par les annotateurs est très infé-

rieure à 2^n. Cependant il n'y a pas pour autant accord des annotateurs. La majorité des segments éliminés par au moins un annotateur ne le sont pas par une majorité.

Afin de savoir si c'est la taille de la phrase qui introduit l'ambiguïté nous séparons les phrases multi-segments du corpus en classes de la même cardinalité, nous définissons les classes comme il suit : la classe C1 contient les phrases de 20 mots ou moins, la classe C2 celles ayant entre 21 et 27 mots, C2 de 28 à 34, C4 de 35 à 45 et finalement C5 regroupe les phrases restantes.

La figure 5.2 présente la proportion des cas ambigus pour ces cinq classes et différents seuils de votes obtenus. Nous pouvons déduire que ce n'est pas la taille de la phrase qui introduit l'ambiguïté. Lorsque le seuil de votation est supérieur à deux tiers toutes les classes ont environ 40% de phrases ayant été compressées de diverses manières. On peut conclure à partir de la figure qu'il y a une ambiguïté inhérente à la tâche de compression de phrases qui ne dépend pas non plus de la granularité de la segmentation.

Même dans le cas des phrases moins complexes, il est rare qu'une seule compression fasse l'unanimité. Dans la plupart des cas la compression la plus fréquente atteint à peine 25% des votes. De fait, il apparaît qu'il existe généralement plusieurs solutions acceptables à cette tâche, aucune n'étant meilleure que l'autre. Par exemple, le tableau 5.3 liste les compressions proposées par les annotateurs d'une phrase avec six EDUs CoSeg. Le tableau présente aussi le nombre de votes recueillis sur une base de 50 annotations. Il s'agit d'un cas où il est difficile de déterminer quelle est la meilleure compression.

Il est alors délicat de décider ce qui correspond à une compression optimale tant l'espace des possibilités reste large. La tâche de compression des phrases, même réduite à l'élimination d'EDUs clairement délimitées, n'est pas déterministe. Nous avons vérifié qu'il est nécessaire de prendre en compte la subjectivité dans la tâche de compression de phrases. Nous allons donc considérer pertinentes toutes les EDUs éliminées par au moins un annotateur. Nous proposons de modéliser l'élimination des EDUS en utilisant un modèle de régression linéaire qui émule les décisions des annotateurs basé sur des caractéristiques de grammaticalité, informativité et segmentation.

5.4 Le résumé automatique et la régression linéaire

H. P. Edmundson a établie un paradigme du résumé automatique par extraction en utilisant la régression linéaire (Edmundson, 1969). Il propose de pondérer les phrases dans un document en se basant sur quatre caractéristiques qu'il trouve déterminantes : les mots-indices (*Cue words*, MI), les mots-clés (*Key words*, MC), l'emplacement (*Location*, EM) et les mots du titre (*Title words*, MT). Le score d'une phrase (équation 5.1) dépend d'une fonction linéaire qui combine ces caractéristiques et dont les poids (β_1, β_2, β_3 et β_4) ont été ajustés manuellement afin d'obtenir les meilleurs résultats vis-à-vis des résumés crées manuellement.

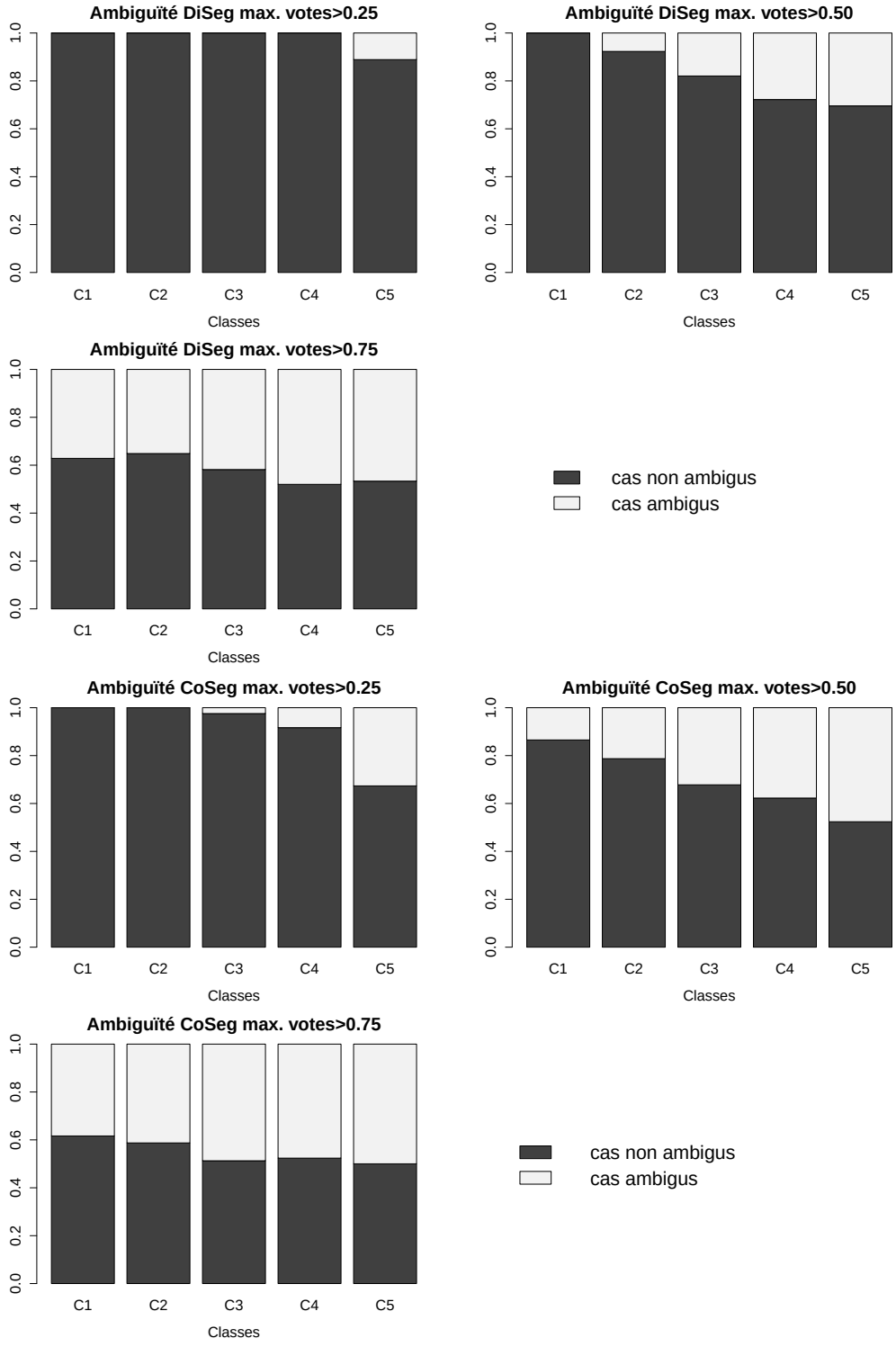


FIGURE 5.2 – *Proportion de l'ambiguïté de l'élimination d'EDUs avec différents seuils de votation.*

$$\mathbf{score}(\varphi) = \beta_1\mathbf{MI}(\varphi) + \beta_2\mathbf{MC}(\varphi) + \beta_3\mathbf{EM}(\varphi) + \beta_4\mathbf{MT}(\varphi) \tag{5.1}$$

Edmundson a d'abord constitué un corpus de 200 articles scientifiques en chimie. Ensuite, il a divisé le corpus en deux parties, une partie pour l'apprentissage et une autre pour le test. Les trois premières caractéristiques correspondent à des attributs des mots trouvés dans la partie d'apprentissage après avoir exclu « *les mots d'usage* ». En fait, MI est la seule caractéristique préalablement extraite à partir du corpus d'apprentissage tandis que le reste doivent être extraites à partir du document à résumer. En effet, MI correspondaient à des motifs lexicaux particuliers tels que « *argue* », « *propose* » ou « *in this paper* », qui ont été repérés et triés par fréquence d'apparition dans le

Votes %	τ%	Compression proposée
27.1	29.5	Los efectos de la LSD sobre el sistema nervioso central son extremadamente variables.
2.1	47.7	Los efectos de la LSD sobre el sistema nervioso central son extremadamente variables el entorno en que se use la droga.
8.3	47.7	Los efectos de la LSD sobre el sistema nervioso central son extremadamente variables y dependen de la cantidad que se consuma.
2.1	52.3	Los efectos de la LSD sobre el sistema nervioso central son extremadamente variables y dependen de la cantidad que se consuma, la personalidad.
2.1	59.1	Los efectos de la LSD sobre el sistema nervioso central son extremadamente variables la pureza de Èsta, el estado de ánimo y las expectativas del usuario.
8.3	68.2	Los efectos de la LSD sobre el sistema nervioso central son extremadamente variables y dependen de la cantidad que se consuma, el estado de ánimo y las expectativas del usuario.
10.4	77.3	Los efectos de la LSD sobre el sistema nervioso central son extremadamente variables y dependen de la cantidad que se consuma, la pureza de ésta, el estado de ánimo y las expectativas del usuario.
2.1	81.8	Los efectos de la LSD sobre el sistema nervioso central son extremadamente variables y dependen de la cantidad que se consuma, la pureza de ésta, la personalidad, el estado de ánimo y las expectativas del usuario.
2.1	86.4	Los efectos de la LSD sobre el sistema nervioso central son extremadamente variables y dependen de la cantidad que se consuma, el entorno en que se use la droga, el estado de ánimo y las expectativas del usuario.
35.4	100	Los efectos de la LSD sobre el sistema nervioso central son extremadamente variables y dependen de la cantidad que se consuma, el entorno en que se use la droga, la pureza de ésta, la personalidad, el estado de ánimo y las expectativas del usuario.
Traduction	100	Les effets du LSD sur le système nerveux central sont extrêmement variables et dépendent de la quantité consommée, de l'entourage où l'acide est consommé, de la pureté de celui-ci, de la personnalité du consommateur, de son état d'esprit et de ses attentes.

TABLE 5.3 – *Exemple des compressions proposées par les annotateurs.*

corpus. Par rapport aux MC, les mots du document à résumer sont triés en ordre descendant de fréquence. Le nombre de mots les plus fréquentes d'une phrase est calculé. L'emplacement d'une phrase (EM) indique simplement sa position dans le texte. Le MT est le nombre de mots dans la phrase qui apparaissent aussi dans un titre ou sous-titre.

En général, un modèle de régression linéaire simple est un outil statistique pour identifier la relation entre une variable expliquée y et un vecteur de variables explicatives $(x_1, x_2, \ldots, x_p)$ (Cook et Weisberg, 2009). L'équation (5.2) correspond à la forme générale d'une régression linéaire simple où l'on suppose implicitement une notion préalable de *causalité* dans le sens où y dépend de $(x_1, x_2, \ldots, x_p)$.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p \tag{5.2}$$

La régression est dite linéaire parce qu'elle impose une forme fonctionnelle linéaire des paramètres $\beta = (\beta_0, \beta_1, \ldots, \beta_p)$. Ainsi, on cherche à estimer le vecteur des paramètres à partir des valeurs réelles mesurées au préalable dans un corpus. La façon classique d'estimer les paramètres du modèle consiste à utiliser la méthode de moindres carrés. On minimise la somme des carrés des écarts entre les valeurs prédites et celles observées par rapport aux paramètres (Wasserman, 2004). Afin de simplifier les explications d'ajustement et qualité de la régression, nous étudierons le cas le plus simple, où nous avons une seule variable explicative (équation 5.3) :

$$y = f(x) = \beta_0 + \beta_1 x \tag{5.3}$$

Ainsi, pour n observations $\{(x_i, y_i), i = 1, 2, \ldots, n\}$, $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ est la moyenne de y. La variabilité de ces observations est mesurée par la somme des carrés totale ou SCT :

$$SCT = \sum_{i=1}^{n} (y_i - \bar{y})^2 \tag{5.4}$$

et la somme des carrés de résidus ou SCR :

$$SCR = \sum_{i=1}^{n} (y_i - f_i)^2 \tag{5.5}$$

Afin, de calculer la qualité de l'ajustement, il est d'usage d'utiliser le coefficient de détermination R^2 ainsi que la statistique F (Cook et Weisberg, 2009). R^2 exprime le rapport entre la variance expliquée par le modèle et la variance totale. D'une valeur comprise entre 0 (rien n'est expliqué) et 1 (tout est expliqué), il mesure l'adéquation entre le modèle et les données observées. L'interprétation géométrique du SCT et du SCR est présentée dans la figure 5.3. Les aires des carrés à droite représentent les carrés des résidus par rapport à la régression ; celles à gauche représentent les carrés de résidus par

rapport à la valeur moyenne. Ainsi, nous avons :

$$R^2 = 1 - \frac{SCR}{SCT}. \tag{5.6}$$

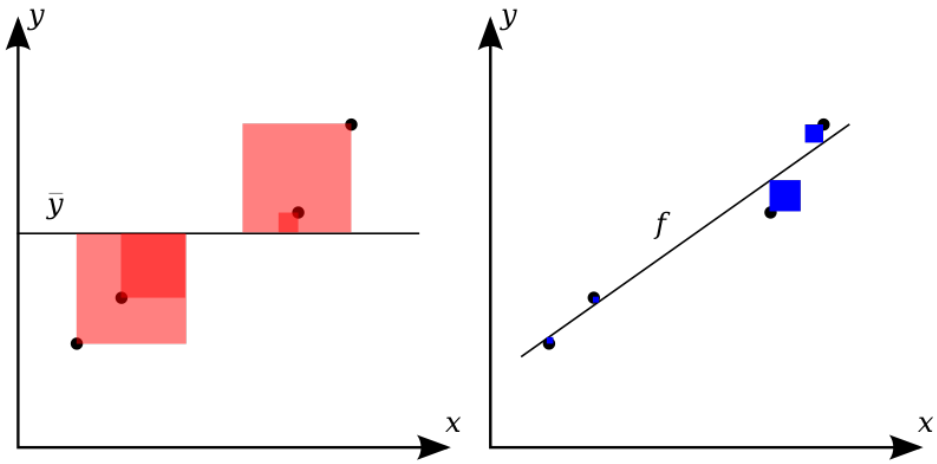


FIGURE 5.3 – *Illustration du coefficient de détermination pour une régression linéaire. Image prise de Wikipédia* http://en.wikipedia.org/wiki/Coefficient_of_determination.

Un défaut du coefficient R^2 est qu'il croît en fonction du nombre de variables explicatives. Or, on sait qu'un excès de variables produit des modèles peu robustes (Cook et Weisberg, 2009). Pour cette raison nous nous intéressons à une version légèrement différente qui prend en compte le nombre de variables explicatives k dont $\beta_i \neq 0$. L'équation (5.7) correspond au coefficient de détermination ajusté ou $R^2_{\text{ajusté}}$.

$$R^2_{\text{ajusté}} = R^2 - \frac{k(1-R^2)}{n-k-1} = 1 - \frac{(1-R^2)(n-1)}{n-k-1}. \tag{5.7}$$

En ce qui concerne la statistique F de l'équation (5.8), elle s'intéresse à la significativité globale d'un modèle et correspond au test statistique de Fisher ayant par hypothèse nulle H_0 : il y a au moins un coefficient β_i qui explique la variable de réponse y contre l'hypothèse H_a, où tous les coefficients sont zéro. Dans le cas d'une régression simple (équation 5.3), seul le paramètre β_1 est concerné.

$$F = (n-2)\frac{R^2}{1-R^2} \tag{5.8}$$

Dans la section 5.5 nous montrerons comment nous avons modélisé les décisions des annotateurs (d'éliminer ou de garder un segment) afin d'ajuster la régression.

5.5 Modèle de régression linéaire pour prédire l'élimination de segments

Les données issues de la campagne d'annotation, décrite dans la section 5.2, nous servent à déterminer le vecteur des paramètres β qui estiment l'élimination d'un segment. Les variables explicatives utilisées correspondent essentiellement à une des quatre catégories selon leur nature : variables d'informativité (énergie textuelle, chapitre 4), variables de grammaticalité (modèles de langage probabilistes, chapitre 3), variables de segmentation (segmentation discursive, chapitre 2) et variables de longueur. Le tableau 5.4 montre la liste de variables choisies et leurs descriptions. L'idée générale consiste à modéliser la probabilité d'élimination de segments effectuée par les annotateurs, en utilisant les variables du tableau 5.4. Notre variable expliquée est basée sur le quotient définie par l'équation (5.9).

$$\mathbf{P_{elim}}(s) = \frac{\text{nombre d'éliminations du segment } s}{\text{total d'annotations sur le segment } s}. \tag{5.9}$$

Une valeur de $\mathbf{P_{elim}}(s) = 1$ signifie que le segment s a été éliminé systématiquement par tous les annotateurs. Par contre $\mathbf{P_{elim}}(s) = 0$ indique que tous les annotateurs ont retenu le segment. L'équation (5.10) définit un modèle linéaire du $\mathbf{P_{elim}}(s)$ selon les variables du tableau 5.4. Nous considérons quatre groupes de variables selon leur type. Le premier groupe ($\beta_1 Es + \beta_2 E\varphi + \beta_3 \widetilde{E}$) mesure l'informativité ($\mathbf{Ener}(s, \varphi)$) ; le deuxième ($\beta_4 Gs + \beta_5 G\varphi + \beta_6 \widetilde{G}$) mesure la grammaticalité ($\mathbf{Gram}(s, \varphi)$) ; le troisième ($\beta_7 S + \beta_8 P + \beta_9 \widetilde{P}$) mesure l'impact de la segmentation ($\mathbf{Seg}(s, \varphi)$) et le quatrième ($\beta_{10} Ls + \beta_{11} L\varphi + \beta_{12} \widetilde{L}$), la longueur ($\mathbf{Lon}(s, \varphi)$).

$$\mathbf{P_{elim}}(s, \varphi) = \mathbf{Ener}(s, \varphi) + \mathbf{Gram}(s, \varphi) + \mathbf{Seg}(s, \varphi) + \mathbf{Lon}(s, \varphi) \tag{5.10}$$

L'équation (5.10) représente un modèle linéaire « *complet* » dans le sens où ils utilise toutes les variables explicatives du tableau 5.4. Afin de faciliter la lecture, dorénavant nous ajoutons un préfixe "D" aux variables quand il s'agit de segments produits par DiSeg et "C" pour ceux de CoSeg.

Les résultats de la signification statistique pour les modèles complets, sur la base de 2 877 résumés (section 5.2), sont présentés dans les tableaux 5.6 et 5.7, pour DiSeg et CoSeg respectivement. La première colonne montre le nom de la variable, la deuxième la valeur du coefficient β, la troisième l'écart type, la quatrième la valeur de la statistique t (le coefficient divisé par l'écart type), la cinquième la probabilité que le coefficient soit significativement différent de 0 et la sixième est une référence visuelle pour repérer la signification statistique de l'estimation (par rapport à la valeur de probabilité de la statistique t). La façon d'interpréter cette référence est décrite dans le tableau 5.5.

On observe des différences importantes entre les tableaux 5.6 et 5.7. D'abord, la qualité de l'ajustement du modèle est supérieure pour CoSeg. Ensuite, il y a quelques

variables qui n'ont pas la même signification statistique dans les deux types de segmentation. Par exemple, l'énergie textuelle d'un segment DiSeg n'a pas d'impact sur l'éstimation de la probabilité d'élimination alors que l'énergie d'un segment CoSeg est plus importante. Puis, le nombre total de segments dans la phrase est à peine significatif pour DiSeg tandis que pour CoSeg il a plus d'importance. En revanche, les variables de longueur du segment et de la phrase sont plus déterminantes en utilisant DiSeg. Ces divergences semblent mettre l'accent sur l'un de nos résultats initiaux : il faut que les segments à éliminer ne soient pas très longs. La probabilité d'élimination pour des segments DiSeg est plus influencée par la longueur des segments que par leur contenu. Cela est dû au fait que les annotateurs n'ont pas eu beaucoup de choix pour éliminer des segments au moment de l'annotation.

Les résultats des modèles complets nous ont encouragé à trouver les modèles optimaux contenant uniquement des variables explicatives significatives pour la prédiction (celles marquées avec une ou plusieurs étoiles). Afin de trouver les modèles optimaux, nous avons généré toutes les régressions linéaires possibles. Notre stratégie a été d'utiliser le concept de masque, c'est-à-dire, un vecteur binaire qui vérifie pour chaque élément i d'un l'ensemble s'il doit apparaître dans un sous-ensemble ou non, selon si les valeurs du masque sont 0 (l'élément ne doit pas apparaître) ou 1 (l'élément doit apparaître). Par exemple, soit

$$\{Es, E\varphi, \widetilde{E}, Gs, G\varphi, \widetilde{G}, S, P, \widetilde{P}, Ls, L\varphi, \widetilde{L}\},$$

l'ensemble de variables à considérer et le masque

$$(0,0,0,0,1,0,0,0,1,1,0,1).$$

Variable	Type	Description
Es	Ener	L'énergie textuelle du segment
$E\varphi$	Ener	L'énergie textuelle de la phrase
$\widetilde{E}$	Ener	Le rapport entre l'énergie textuelle du segment et celle de la phrase, il est défini par $Es/E\varphi$
Gs	Gram	La valeur du segment dans un modèle de langage de $n-$grammes
$G\varphi$	Gram	La valeur de la phrase dans un modèle de langage de $n-$grammes
$\widetilde{G}$	Gram	Le rapport entre la valeur du segment dans un modèle de langage de $n-$grammes et celle de la phrase, il est défini par $Gs/G\varphi$
S	Seg	Le nombre total de segments dans la phrase
P	Seg	La position du segment dans la phrase
$\widetilde{P}$	Seg	La position relative du segment dans la phrase, elle est définie par P/S
Ls	Lon	La longueur du segment en nombre de mots
$L\varphi$	Lon	La longueur de la phrase en nombre de mots
$\widetilde{L}$	Lon	Le rapport entre la longueur du segment, en nombre de mots, et celle de la phrase, il est défini par $Ls/L\varphi$

TABLE 5.4 – *Liste de variables explicatives utilisées pour l'ajustement de la régression linéaire.*

Code	Valeur de la $\mathbf{P}(>\|t\|)$	Interprétation
****	0.000	hautement significatif
***	0.001	très significatif
**	0.010	assez significatif
*	0.050	peu significatif
	0.100	non significatif

TABLE 5.5 – *Interprétation de codes de référence visuelle pour les tableaux de résultats des modèles linéaires.*

	Estimation	Écart type	Valeur t	$\mathbf{P}(>\|t\|)$	Code
DEs	-0.027	0.016	-1.595	0.111	
$DE\varphi$	0.026	0.009	2.710	0.007	***
$D\widetilde{E}$	-0.008	0.092	-0.092	0.926	
DGs	0.005	0.002	1.897	0.058	*
$DG\varphi$	-0.003	0.001	-2.159	0.031	**
$D\widetilde{G}$	0.056	0.257	0.221	0.824	
DS	0.024	0.014	1.668	0.096	*
DP	-0.032	0.019	-1.642	0.101	
$D\widetilde{P}$	0.672	0.069	9.679	0.000	****
DLs	0.016	0.006	2.784	0.005	***
$DL\varphi$	-0.007	0.002	-2.699	0.007	***
$D\widetilde{L}$	-0.768	0.262	-2.929	0.003	***
statistique F= 107.1 avec 12 et 433 degrés de liberté, valeur p : < 2.2e-16					
R^2= 0.695, R^2 ajusté=0.690					

TABLE 5.6 – *Modèle linéaire complet pour des segments DiSeg.*

La régression associée serait donc

$$\beta_5 G\varphi + \beta_9 \widetilde{P} + \beta_{10} Ls + \beta_{12} \widetilde{L}.$$

Ce qui réduit le problème à l'implémentation d'un compteur binaire efficace.

Après la génération, nous avons doublement trié tous les modèles, d'abord par rapport au coefficient de détermination ajusté (R^2 ajusté) et ensuite selon la statistique F (section 5.4). Les résultats pour les modèles optimaux correspondent aux tableaux 5.8 et 5.9 d'où les régressions linéaires qui optimisent la prédiction de la valeur du $\mathbf{P_{elim}}$ correspondent à l'équation 5.11 pour les segments DiSeg et à l'équation 5.12 pour CoSeg.

Avec l'élimination de variables sans signification statistique les résultats ont légèrement amélioré . Les résultats de CoSeg continuent à être supérieurs à DiSeg. La différence la plus importante est que la position d'un segment CoSeg n'affecte pas la probabilité d'élimination mais ceci n'est pas vrai pour les segments DiSeg. En tout cas, c'est la position relative qui détermine le plus l'élimination des segments que la position absolue dans le deux types de segmentation.

	Estimation	Écart type	Valeur t	$\mathbf{P}(>\|t\|)$	Code
CEs	-0.057	0.017	-3.372	0.000	****
$CE\varphi$	0.057	0.006	8.488	0.000	****
$C\widetilde{E}$	0.033	0.100	0.337	0.736	
CGs	0.005	0.002	1.795	0.073	*
$CG\varphi$	-0.002	0.001	-2.425	0.015	**
$C\widetilde{G}$	0.099	0.199	0.502	0.615	
CS	-0.013	0.006	-2.192	0.028	**
CP	-0.004	0.007	-0.669	0.503	
$C\widetilde{P}$	0.395	0.051	7.720	0.000	****
CLs	0.015	0.006	2.297	0.021	**
$CL\varphi$	-0.004	0.002	-1.951	0.051	*
$C\widetilde{L}$	-0.584	0.220	-2.653	0.008	***
statistique F= 153.5 avec 12 et 809 degrés de liberté, valeur p : < 2.2e-16 R^2= 0.748, R^2 ajusté=0.741					

TABLE 5.7 – *Modèle linéaire complet pour des segments CoSeg.*

	Estimation	Écart type	Valeur t	$\mathbf{P}(>\|t\|)$	Code
DEs	-0.028	0.006	-4.097	0.000	****
$DE\varphi$	0.027	0.008	3.163	0.001	***
DGs	0.004	0.001	2.584	0.010	**
$DG\varphi$	-0.002	0.001	-2.559	0.010	**
DS	0.024	0.013	1.838	0.066	*
DP	-0.031	0.018	-1.687	0.092	*
$D\widetilde{P}$	0.669	0.065	10.303	0.000	****
DLs	0.015	0.004	3.756	0.000	****
$DL\varphi$	-0.007	0.002	-3.142	0.001	***
$D\widetilde{L}$	-0.713	0.070	-10.070	0.000	****
statistique F= 129.1 avec 10 et 435 degrés de liberté, valeur p : < 2.2e-16 R^2= 0.696, R^2 ajusté=0.691					

TABLE 5.8 – *Modèle linéaire optimal pour des segments DiSeg.*

Deux conclusions importantes ressortent à partir des tableaux des regresions optimales. La première est que les modèles choisis pour caractériser les trois aspects de la compression de phrases (segmentation discursive pour la longueur, énergie textuelle pour l'informativité et modèles de langage probabilistes pour la grammaticalité) se prêtent bien pour cette tâche. D'où, le fait qu'il soit possible de proposer des modèles beaucoup plus complexes en utilisant de telles techniques. La deuxième est qu'une partie importante de la variance n'est pas capturée à cause de la subjectivité inhérente. La divergence de réponses a donc un effet negatif dans la régression et on peut supposer qu'un effet similaire resortira en utilisant le même corpus annoté indépendamment de la méthode d'apprentissage utilisée.

	Estimation	Écart type	Valeur t	$\mathbf{P}(>\|t\|)$	Code
CEs	-0.053	0.005	-9.613	0.000	****
$CE\varphi$	0.059	0.005	10.508	0.000	****
CGs	0.003	0.001	2.024	0.043	**
$CG\varphi$	-0.002	0.001	-2.585	0.009	***
CS	-0.015	0.005	-2.961	0.003	***
$C\widetilde{P}$	0.369	0.029	12.451	0.000	****
CLs	0.013	0.005	2.589	0.009	***
$CL\varphi$	-0.004	0.001	-2.048	0.040	**
$C\widetilde{L}$	-0.446	0.068	-6.480	0.000	****
statistique F= 205.1 avec 9 et 812 degrés de liberté, valeur p : < 2.2e-16					
R^2= 0.748, R^2 ajusté=0.753					

TABLE 5.9 – *Modèle linéaire optimal pour des segments CoSeg.*

$$\hat{\mathbf{P}}_{\mathbf{elim\_DiSeg}}(s,\varphi) = -0.028DEs + 0.027DE\varphi + 0.004DGs - 0.002DG\varphi + 0.024DS - 0.031DP \\ +0.669D\widetilde{P} + 0.015DLs - 0.007DL\varphi - 0.713D\widetilde{L} \tag{5.11}$$

$$\hat{\mathbf{P}}_{\mathbf{elim\_CoSeg}}(s,\varphi) = -0.053CEs + 0.059CE\varphi + 0.003CGs + -0.002CG\varphi - 0.015CS + 0.369C\widetilde{P} \\ +0.013CLs - 0.004CL\varphi - 0.446C\widetilde{L} \tag{5.12}$$

Les équations 5.11 et 5.12 sont des estimateurs de la probabilité d'élimination des segments DiSeg et CoSeg respectivement. Par la suite, nous les utiliserons pour générer des phrases compressées dans un système de résumé automatique. Nous proposons deux algorithmes de résumé par compressiom dont le coeur est une expression conditionnelle pour décider si un segment doit être éliminé ou non en se basant sur des valeurs éstimées par ces équations.

5.6 Deux algorithmes de génération de résumés par élimination de segments discursifs

Nous proposons deux algorithmes de génération de résumés par élimination des segments. Tous les deux partent du même principe. D'abord, chaque phrase est découpée en segments, puis une fonction de probabilité détermine, pour chaque segment, s'il doit être éliminé. Les fonctions de probabilité correspondent à l'équation 5.11 pour les segments DiSeg et à l'équation 5.12 pour ceux de CoSeg.

Ainsi, nous avons crée deux algorithmes dont le coeur est justement une expression conditionnelle de type **if** $(\mathbf{P}_{\mathbf{elim}}(s,\varphi) > \alpha)$ **then eliminer**(s) où α est la valeur du seuil de probabilité d'élimination du segment.

L'algorithme 1 prend comme arguments la valeur α et le document à résumer *Doc*. D'abord *Doc* est segmenté en phrases et chaque phrase est à son tour découpée en segments discursifs. Ensuite, on décide, pour chaque segment s'il doit être éliminé ou non, selon la fonction 5.10. Il est important de remarquer que pour calculer cette probabilité il faut calculer préalablement les valeurs d'énergie textuelle et de grammaticalité, en considérant les deux segmentations. Finalement, un résumé avec de phrases compressées est généré.

Algorithm 1 Génération de résumés par élimination de segments.

Arguments : (seuil de probabilité $\alpha \in [0,1]$, Document *Doc*)
Segmenter(*Doc*)
for all sentences φ in *Doc* **do**
  for all segments s in φ **do**
    if $(\mathbf{P_{elim}}(s,\varphi) > \alpha)$ **then**
      Eliminer(s) de φ
    end if
  end for
end for
  return résumé

Nous avons produit des résumés en utilisant chaque algorithme. Les résumés ont été créés à partir de textes qui ne font pas partie du corpus d'apprentissage. À cet effet, nous avons divisé le corpus en trois parties : 2/3 pour l'apprentissage des paramètres à partir des données annotées et 1/3 pour générer les résumés. Pour l'algorithme 1, nous avons produit neuf résumés par document, en faisant varier la valeur α de 0.05 à 0.95, avec un pas de 0.1.

La figure 5.4 présente le taux de compression en fonction de la valeur α pour les résumés produits avec l'algorithme 1. Chaque résumé est représenté par un symbole. Les figures en haut présentent le taux de compression en nombre de mots ; celles en bas en nombre de segments. La colonne de gauche présente l'information pour DiSeg ; celle de droite pour CoSeg.

Nous observons principalement que dans tous les cas, à partir de la valeur de seuil 0,6 les textes ne sont plus compressés. Cela est en accord avec les résultats mentionnés dans la section 5.3 par rapport à l'ambiguïté. En effet, le modèle linéaire a réussi à émuler le processus d'annotation manuelle, c'est-à-dire la tendance à conserver l'information.

Nous remarquons aussi un dispersement plus évidente pour DiSeg, étant donné qu'il segmente d'une manière plus grossière que CoSeg. Comme la segmentation de ce dernier est fine, la courbe obtenue est plus régulière entre $\alpha = 0.05$ de 0.05 à $\alpha = 0.55$.

Une limite de l'algorithme 1 est son incapacité à contrôler le taux de compression τ. Pour palier à cette limite, nous avons développé l'algorithme 2 qui peut contrôler τ en variant la valeur α. La valeur du seuil de probabilité est incrémentée dans la boucle principale jusqu'à ce que le résumé ait la taille voulue.

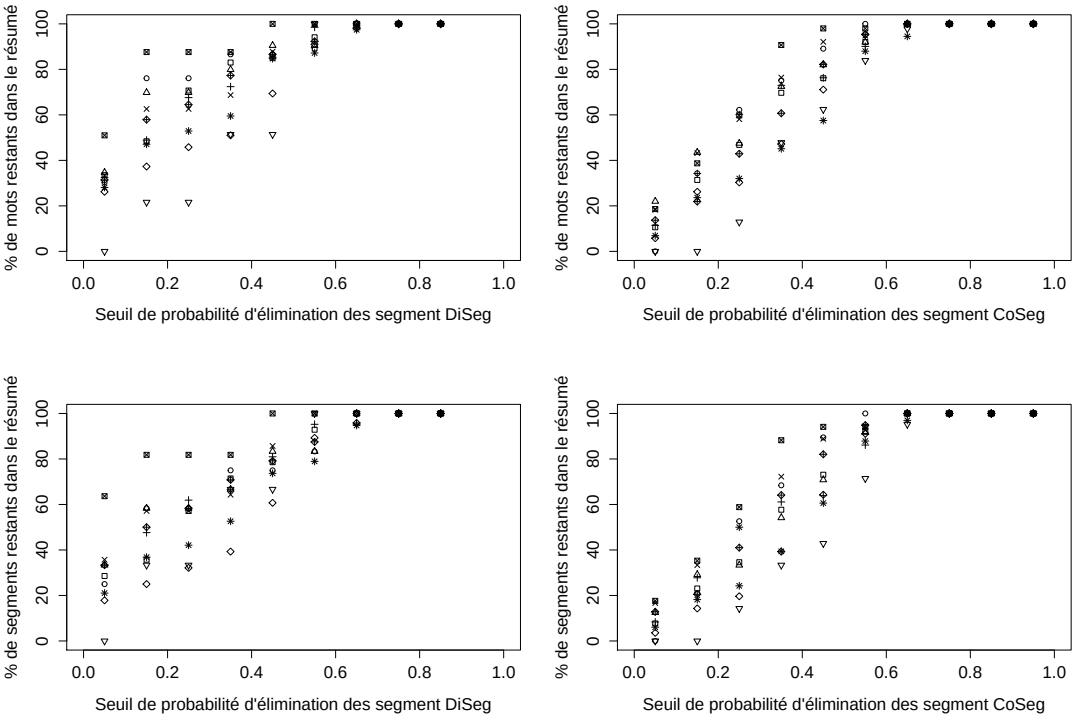


FIGURE 5.4 – *Taux de compression en fonction de la valeur de probabilité d'élimination d'EDUs.*

Pour l'algorithme 2, nous avons produit neuf résumés par document en variant la valeur τ de 10% à 90% avec un pas de 10%. En conséquence, nous avons obtenu des résumés avec des taux de compression différents. D'abord les plus courts et puis la taille augmente jusqu'à ce que le résumé corresponde au document d'origine.

Le tableau 5.10 montre les résumés les plus courts d'un texte extrait de notre corpus produits par l'algorithme 1 avec $\alpha = 0.05$. Le document d'origine est également montré afin de comparer les résultats. La qualité des résumés de l'ensemble de test est formellement évaluée dans la section 6.5. Dans l'annexe B, nous affichons tous les résumés obtenus pour ce même document en utilisant l'algorithme 2.

5.7 Conclusions du chapitre

Dans ce chapitre nous avons montré que la compression de phrases est un tâche subjective dans le sens qu'il s'agit d'un problème avec plusieurs solutions.

Nous avons décrit notre protocole d'annotation du corpus, réalisé grâce aux sciences citoyennes qui engagent des volontaires non-experts dans des tâches scientifiques et ont l'avantage de permettre de constituer un grand corpus annoté.

Pour faciliter la tâche d'annotation, nous avons créé une plate-forme Web. Le corpus

Algorithm 2 Génération de résumés par élimination de segments discursifs avec le taux de compression comme paramètre.

Arguments : (τ souhaité, Document Doc)
$\alpha \leftarrow 0$
$Segmenter(Doc)$
repeat
  for all sentences φ in Doc **do**
    for all segments s in φ **do**
      if $(\mathbf{P_{elim}}(s, \varphi) > \alpha)$ **then**
        $Eliminer(s)$ de φ
      end if
    end for
  end for
  $\alpha \leftarrow alpha + 0.01$
until τ du Résumé $\geq \tau$ souhaité
  return Résumé

que nous avons obtenu est, à notre connaissance, le premier corpus de phrases annoté en espagnol et il sera disponible au public sur Internet.

Nous avons analysé l'accord entre les annotateurs où nous avons trouvé qu'ils proposent des compressions différentes mais le nombre de solutions proposées est inférieur par rapport à la taille de l'espace de recherche. Nous proposons de simuler ces décisions avec une régression linéaire.

Nous avons présenté une régression linéaire basée sur des variables d'énergie textuelle, des modèles de langage probabilistes, des segmentations discursives et de la longueur. Ainsi, notre approche est inspirée du paradigme d'apprentissage supervisé d'Edmundson dont la probabilité d'éliminer un segment discursif a été apprise sur la base d'environ soixante mille annotations.

Nous avons réalisé une analyse des variables qui sont significatives pour déterminer l'élimination d'un segment. Pour ce faire, nous avons trouvé les modèles linéaires optimaux pour chaque segmenteur discursif.

Finalement, nous avons introduit deux algorithmes de génération de résumé par élimination de segments. Le premier nous permet de conditionner la probabilité qui détermine si un segment doit être éliminé ; le deuxième nous permet de contrôler le taux de compression.

Descubrimiento de mamut emociona a científicos (résumé) seg=CoSeg, $\alpha = 0.05$, $\tau = 24.8$
Los científicos de la Universidad de Iowa planean escanear el lugar el viernes con un radar que penetra en la tierra. La excavación se extenderá por varios meses. Los huesos los encontró hace dos años el propietario del terreno.
Descubrimiento de mamut emociona a científicos (résumé) seg=DiSeg, $\alpha = 0.05$, $\tau = 33.3$
Los científicos de la Universidad de Iowa planean escanear el lugar el viernes con un radar que penetra en la tierra. La excavación se extenderá por varios meses. Los huesos los encontró hace dos años el propietario del terreno, quien desea que tanto su nombre como la ubicación de la granja sean confidenciales.
Descubrimiento de mamut emociona a científicos (document source)
El inusual descubrimiento de los huesos de un mamut en una finca de Oskaloosa ha emocionado a los expertos que estudian la vida prehistórica por los descubrimientos científicos que podrían hacer se con la enorme bestia. El hallazgo es raro porque al parecer buena parte del esqueleto del animal se encuentra en excelente estado. Eso permite a los científicos recolectar polen y evidencias de otras plantas en el sitio de la excavación que podrían revelar detalles respecto del medio ambiente de Iowa hace más de 12000 años. Los científicos de la Universidad de Iowa planean escanear el lugar el viernes con un radar que penetra en la tierra para ver si pueden determinar qué tanto de los restos del mamut siguen enterrados. La excavación se extenderá por varios meses. Los huesos los encontró hace dos años el propietario del terreno, quien desea que tanto su nombre como la ubicación de la granja sean confidenciales para proteger el sitio.
Découverte d'un mammouth enflamme les scientifiques (traduction)
L'inusuelle trouvaille des ossements d'un mammouth dans une ferme de Oskaloosa a enflammé les experts qui étudient la vie préhistorique à cause des découvertes scientifiques qui pourraient se faire avec l'énorme bête. La trouvaille est rare parce qu'il paraît qu'une bonne partie du squelette de l'animal se trouve en excellent état. Ceci permet aux scientifiques de collecter du pollen et des évidences d'autres plantes dans le site de l'excavation qui pourraient révéler des détails sur l'environnement de Iowa il y a plus de 12000 ans. Les scientifiques de l'Université d'Iowa prévoient de scanner le site vendredi avec un radar qui pénétra la terre pour voir s'ils peuvent déterminer quelle quantité des restes du mammouth sont encore ensevelis. L'excavation s'étendra durant plusieurs mois. Les ossements ont été trouvés il y a deux ans par le propriétaire du terrain, qui souhaite que son nom ainsi que l'adresse de la ferme soient confidentielles pour protéger le site.

TABLE *5.10* – *Exemples de résumés générés à partir d'un texte extrait de notre corpus après avoir utilisé l'algorithme de compression de phrases en variant la probabilité d'éliminer les segments (α).*

Chapitre 6

Test de Turing pour l'évaluation de résumés automatiques

Sommaire

Dans ce chapitre, nous abordons l'évaluation des résumés automatiques. Nous discutons la problématique que pose l'utilisation des métriques ROUGE, BLEU et FRESA pour l'évaluation des résumés avec des phrases compressées. Nous présentons alors une méthode inspirée du test de Turing dans laquelle des juges humains doivent identifier l'origine, humaine ou automatique, d'une série des résumés. Nous expliquons comment valider les réponses des juges avec le test statistique exact de Fisher.

6.1 Problématique de l'évaluation pour la compression de phrases

L'évaluation du résumé de documents n'est pas une question triviale. Cela a conduit à l'exploration de plusieurs approches à ce sujet qui apportent des réponses partielles. (Amigó et al., 2005) proposent une classification dont les deux axes principaux sont l'évaluation intrinsèque et l'évaluation extrinsèque. Une division plus générale nous semble plus utile vis-à-vis des propos de cette thèse. Deux grandes catégories d'évaluation sont distinguées par rapport aux ressources qu'elles utilisent : l'évaluation manuelle et l'évaluation automatique.

Dans l'évaluation manuelle le principe est de comparer les résumés automatiques avec ceux produits par des humains (Edmundson, 1969). La principale difficulté est qu'un seul texte peut engendrer une infinité de résumés valables. Il est impossible de laisser de coté la subjectivité inhérente à la tâche et on doit considérer que, parfois, même les évaluateurs experts ne sont pas d'accord sur la qualité d'un résumé. Un travail représentatif de ce type d'évaluation est celui de (Mani et al., 1999) où les auteurs proposent de donner aux évaluateurs humains des résumés produits par des méthodes automatiques ainsi que les documents originaux. Dans ces derniers, il y a quelques phrases-clé qui contiennent des informations pertinentes vis-à-vis de la thématique du texte. Ces phrases sont des références de contenu qui doivent impérativement être incluses dans les résumés automatiques. Les évaluateurs comparent les deux documents et vérifient que les résumés automatiques incluent bien ces phrases. Une autre approche dans la même ligne mais applicable à des résumés monodocument est proposée par (Saggion et Lapalme, 2000). On fournit aux évaluateurs une liste de « *concepts-clés* » qui doivent être mentionnés dans les résumés automatiques. Plus récemment, (Orasan et al., 2007) proposent d'évaluer la qualité du résumé grâce à un test de comparaison. Les évaluateurs sont censés de juger le meilleur résumé d'une paire dont un a été élaboré à l'aide d'un outil de résumé assisté par ordinateur (« *Computer-aided summarisation* ») et l'autre sans cet outil. Leur hypothèse est qu'il n'y a pas de différence statistiquement significative entre les deux types de résumé. Ainsi, les juges sont incapables de les distinguer. Les résultats de telles expériences sont vérifiés grâce au test du χ^2.

D'un côté, l'évaluation automatique est capable de traiter une quantité massive de documents et d'un autre côté, elle diminue la subjectivité inhérente des juges humains. La méthode d'évaluation automatique la plus répandue est ROUGE (*Recall-Oriented Understudy for Gisting Evaluation*) (Lin, 2004). Elle est utilisée lors des campagnes internationales des conférences DUC - *Document Understanding Conferences*, devenues *Text Analysis Conference* (TAC) [1] en 2008. Dans cette méthode, un résumé candidat est comparé avec plusieurs résumés élaborés par des experts (appelés résumés modèles ou références). La mesure sous-jacente est basée sur les co-occurrences des n-grammes entre le résumé candidat et les références. L'équation (6.1) correspond à la métrique ROUGE-N, où n est la taille du n-gramme, $gram_n$ et $Count_{match}(gram_n)$ est le nombre maximum de n-grammes qui apparaissent dans le résumé candidat et les références. Formellement, ROUGE est basé sur le rappel entre le résumé candidat et l'ensemble

1. http ://www.nist.gov/tac/

des résumés de référence. Il faut observer que le dénominateur est la somme totale du nombre de n-grammes dans les références.

$$\text{ROUGE–N} = \frac{\sum_{S\epsilon\{\text{résumés de référence}\}} \sum_{gram_n \epsilon S} Count_{match}(gram_n)}{\sum_{S\epsilon\{\text{résumés de référence}\}} \sum_{gram_n \epsilon S} Count(gram_n)} \tag{6.1}$$

Les résumés avec phrases compressées possèdent moins de mots que ceux qui n'ont pas été compressés. Par conséquence, les candidats avec moins de contenu sont pénalisés même si leur contenu est pertinent. Pour cette raison, la mesure ROUGE n'est pas adaptée à la compression de phrases.

Dans (Molina et al., 2010a), nous essayons d'évaluer des phrases compressées avec une métrique semi-automatique développée par IBM pour la tâche de traduction automatique (*Machine Translation* ou MT), le score BLEU (Papineni et al., 2002). Ce score est basé sur la précision entre les n-grammes d'une phrase candidate est un ensemble de phrases de référence. Il faut diviser le nombre de n-grammes de la phrase candidate présents dans les références par le nombre total de n-grammes de la phrase candidate. Dans l'équation (6.2), C correspond à la phrase à évaluer et $Count_{clip}$(n-gramme) est le maximum de fois qu'un n-gramme de la phrase candidate est dans une des références. Une pénalisation par rapport à la taille (*Brevity Penalty*, BP) est imposée aux phrases trop longues ou trop courtes. En conséquence, les phrases compressées obtiennent des scores bas. Plus la phrase candidate est courte, plus elle est pénalisée par BLEU (équation 6.3).

$$Prec_n = \frac{\sum_{C\epsilon\{\text{Candidates}\}} \sum_{\text{n-gramme } \epsilon C} Count_{clip}(\text{n-gramme})}{\sum_{C\epsilon\{\text{Candidates}\}} \sum_{\text{n-gramme } \epsilon C} Count(\text{n-gramme})} \tag{6.2}$$

$$\text{BLEU} = BP \times e^{\left(\sum_{n=1}^{N} \frac{1}{N} \mathbf{log}(Prec_n)\right)} \tag{6.3}$$

Dans (Molina et al., 2012) nous essayons d'évaluer des résumés avec des phrases compressées en utilisant la mesure FRESA (Torres-Moreno et al., 2010a; Saggion et al., 2010) qui n'a pas besoin de résumés de référence mais uniquement du document d'origine. FRESA calcule les divergences entre les distributions des fréquences de termes entre le résumé à évaluer et le document d'origine. Ces divergences correspondent à celle de Kullback-Leibler (KL) et celle de Jensen-Shannon (JS) comme il est décrit dans (Louis et Nenkova, 2008). Soit T l'ensemble de termes contenus dans le document d'origine. Pour chaque $t \in T$, C_t^T est le nombre d'apparitions de t dans le document d'origine et C_t^S est le nombre d'apparitions de t dans le résumé à évaluer.

$$\mathcal{D}(T||S) = \sum_{t \in T} \left| \log\left(\frac{C_t^T}{|T|} + 1\right) - \log\left(\frac{C_t^S}{|S|} + 1\right) \right| \tag{6.4}$$

L'équation (6.4) calcule la différence absolue entre les divergences des distributions (dans l'espace *log*). Une fois normalisées entre 0 et 1, les valeurs élevées de FRESA

(basse divergence) sont associées à la similitude entre le résumé et le texte d'origine tandis que les valeurs basses (haute divergence) impliquent une dissimilitude. L'interprétation de cette mesure n'est pas évidente. La seule conclusion qu'on peut tirer est que la valeur de divergence entre un texte et son résumé est élevée, ceci est toujours vrai pour les phrases compressées comme nous montrons dans les expériences de la section 3.3. FRESA a associé des valeurs d'haute divergence indépendamment de la stratégie suivie, y compris la compression aléatoire (Molina et al., 2012).

Visiblement, aucune des mesures présentées ne prennent en compte la structure grammaticale de la phrase compressée candidate. Il est donc possible qu'un résumé ayant des phrases compressées complètement agrammaticales obtienne un bon score. En effet, un des principaux désavantages des évaluations automatiques est qu'elles ne considèrent pas les informations syntaxiques ni sémantiques puisqu'elles sont basées sur des apparitions d'éléments lexicaux tels que les n-grammes. À notre connaissance, le problème d'évaluer des phrases compressées, de même que celui de l'évaluation du résumé automatique, restent ouverts. Nous proposons donc d'affronter les problèmes de l'évaluation autrement, en utilisant le test de Turing.

6.2 Le jeu d'imitation

Nous nous sommes inspirés des idées qu'Alan M. Turing a eu dans les années cinquante et qui continuent à être polémiques, mais qui peuvent être avantageuses pour l'évaluation de quelques tâches du *Traitement Automatique de la Langue Naturelle* et plus concrètement du résumé automatique. Nous parlons du très connu test de Turing décrit dans l'article (Turing, 1950). Dans cet article, Turing propose de discuter par rapport à la question : et si les machines pouvaient penser ? Pour éviter des explications au niveau de la définition de penser, il a établi « *le jeu de l'imitation* », aujourd'hui connu comme le test de Turing.

Dans le test, il y a deux joueurs et un juge. Le premier joueur est un être humain (A) et le deuxième une machine (B). Une autre personne (C), le juge, doit deviner l'identité de chacun des joueurs, mais il n'a pas le droit de les voir. Il est permis que les joueurs interagissent avec le juge par un terminal. Par exemple, le juge écrit des questions à l'aide d'un clavier et lit les réponses des joueurs sur un écran. Les réponses sont identifiées par deux variables x et y. À un moment du jeu, le juge doit indiquer qui a dit quoi à partir des réponses obtenues durant l'échange. Il s'agit d'associer x et y à A et B. D'après (Turing, 1950), un dialogue typique pourrait être comme celui du tableau 6.2 [2].

Évidement, l'objectif de cette expérience établie par Turing n'était pas de tromper quelqu'un, mais de se poser des questions philosophiques autour de la pensée. Concrètement, il s'agit de réfléchir à la possibilité de créer de la pensée artificiellement et de simuler artificiellement les fonctionnalités cognitives du cerveau humain. Nous profitons de quelques aspects du protocole du test qui nous semblent intéressants pour évaluer une tâche très complexe pour laquelle aucune méthode d'évaluation efficace a

2. Nous avons pris la liberté d'adapter à nos besoins la dernière ligne de ce dialogue.

été proposée. Nous sommes d'accord avec (Harnad, 2000) sur le fait que Turing a privilégié l'interaction par la langue naturelle. La langue n'est-elle pas justement un des moyens principaux pour véhiculer la pensée ?

Toutefois, les aspects philosophiques du test ne concernent pas cette étude-ci. Nous souhaitons uniquement vérifier un type bien spécifique de fonctionnalité linguistique : la génération de résumés. En effet, dans le test de Turing le juge n'a pas le droit de voir les joueurs. Avec cette restriction, Turing a mis en évidence que ce sont les aspects fonctionnels et non les aspects physiques qui doivent être jugés. Il nous semble naturel de transposer ce test pour évaluer des tâches du traitement automatique de la langue dont l'objectif est de simuler la performance des humains. À cet égard, nous envisageons une variante du test de Turing adressée à l'évaluation de résumés automatiques.

6.3 Le test de Turing revisité pour évaluer le résumé automatique

Supposons qu'un être humain (A) et une machine (B) produisent respectivement deux résumés (a) et (b), à partir du même document. (A) et (B) doivent respecter les mêmes règles afin que les productions soient homogènes et, en conséquence, comparables (par exemple, le même taux de compression). Un juge humain (C), doit déterminer lequel des résumés a été élaboré par (A) et lequel a été élaboré par (B). Il doit dévoiler l'identité de chaque joueur en s'appuyant uniquement sur la lecture de leurs résumés.

Pour l'évaluation, nous voulons vérifier si la qualité de nos résumés passe le test de Turing. Pour avoir des résultats statistiquement significatifs nous avons maximisé, dans nos possibilités, le nombre de juges. Il est important de remarquer que nos juges d'évaluation sont différents de nos premiers annotateurs.

Six résumés humains (A) ont été tirés aléatoirement de notre corpus de 2 877 ré-

(juge) : Quel est le résultat de 34 957 + 70 764 ?

(joueur) : (Un pause d'environ 30 secondes, puis la réponse) 105 621.

(juge) : Savez jouer aux échecs ?

(joueur) : Oui.

(juge) : J'ai le roi en roi 1 et aucune autre pièce. Vous avez le roi en roi 6 et un pion en pion 1. C'est à vous, Comment allez-vous jouer ?

(joueur) : (Un pause d'environ 15 secondes) roi à roi 8, mate et échec.

(juge) : S'il vous plaît, écrivez un sonnet qui aie pour thème le pont d'Avignon.

(joueur) : Ne comptez pas sur moi pour ça, je n'ai jamais pu écrire de poésie.

(juge) : Eh bien, alors lisez l'article sur « *Le pont d'Avignon* » dans Wikipédia et préparez un résumé de celui-ci.

TABLE 6.2 – *Exemple d'un dialogue dans le jeu d'imitation d'après Turing.*

sumés après l'annotation (section 5.2). Nous avons choisi six résumés produits automatiquement (B) avec l'algorithme 1 (section 5.6). Parmi les six résumés automatiques trois ont été segmentés avec DiSeg et trois avec CoSeg. Pour chaque segmenteur nous en avons sélectionné trois selon le taux de compression τ. Nous avons gardé ceux qui avaient les meilleurs scores de grammaticalité pour chacune des catégories listées dans le tableau 6.3. Notre seule intervention a été de nous assurer que les phrases commençaient par une majuscule et terminaient par un point. Les 54 juges hispanophones [3] (C) ignoraient toute cette information. Nous leur avons donné une seule consigne : déterminer pour chaque résumé s'il avait été produit par un humain ou par une machine. L'annexe C montre une copie du document distribué aux juges.

Catégorie	Nombre de mots doc. source	Nombre de mots résumé	Taux de compression τ%	Segmenteur
$\tau < 50\%$	303	49	16.1%	DiSeg
$\tau \approx 50\%$	209	104	49.6%	DiSeg
$\tau > 50\%$	156	119	76.3%	DiSeg
$\tau < 50\%$	217	57	26.2%	CoSeg
$\tau \approx 50\%$	165	76	43.4%	CoSeg
$\tau > 50\%$	234	186	79.4%	CoSeg

TABLE 6.3 – *Critères de sélection pour l'évaluation de résumés avec un test de Turing.*

6.4 La goûteuse de thé : le test exact de Fisher

Pour valider statistiquement nos résultats, nous nous inspirons de l'expérience de «*la dame du thé*» décrite dans (Agresti, 2002) grâce à laquelle Ronald A. Fisher a inventé un test statistique exact [4]. Dans l'expérience originale, une dame (Dr. Muriel Bristol) se vante d'être capable de distinguer si une tasse de thé a été versée sur un fond de lait ou pas. Pour tester sa prétention, Fisher lui demande de goûter 8 tasses de thé, 4 versées sur du lait avant d'être servies et 4 ne contenant que du thé. La goûteuse doit indiquer se réponse pour chaque tasse. Elle sait qu'il y a exactement 4 tasses ne contenant que du thé. Le test statistique, proposé par Fisher, est basé sur le comptage du nombre de bonnes et mauvaises réponses à l'aide d'un tableau de contingence comme celui de l'exemple 6.4. L'hypothèse nulle est calculée en comparant le nombre de permutations sélectionnées avec le nombre de permutations non sélectionnées.

Avec les résultats du tableau 6.4 est-il possible d'affirmer que la dame possède l'habilité dont elle se vante ? Il y a 70 façons de choisir les réponses d'un tableau de contingence avec les restrictions du test, en formant un groupe de 4 tasses parmi un total de 8. La première sélection se fait entre 8 tasses, la deuxième entre 7, la troisième entre 6 et la quatrième entre 5. Soit $8 \times 7 \times 6 \times 5 = 1\,680$ façons différentes. Comme on a déjà compté

3. Tous les juges ont un niveau d'études de Bac +4 ou plus.
4. Ainsi nommé parce que la signification de la déviation de l'hypothèse nulle peut être calculée de manière exacte.

	Réponse correcte	
Réponse de la goûteuse	Lait	Thé
Lait	a =3	b =1
Thé	c =1	d =3

TABLE 6.4 – *Tableau de contingence pour l'évaluation des réponses de la goûteuse de thé avec le test statistique exact de Fisher.*

tous les ordres de choix possibles on doit diviser par 24. En effet, 4 objets peuvent être ordonnés de 4×3×2×1= 24 façons possibles. En conséquence, la dame peut choisir de manière correcte tout au plus 1 tableau parmi 70. Soit $1/70 \approx 1.4$, qui en termes statistiques, se trouve dans la région critique à 95% de confiance car p =0.014 < 0.05.

Fisher a montré (Fisher et al., 1935) que la probabilité d'obtenir un tel tableau est donnée par la loi hypergéométrique (équation 6.5).

$$p = \frac{\binom{a+b}{a}\binom{c+d}{c}}{\binom{n}{a+c}} = \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{a!b!c!d!n!} \tag{6.5}$$

Où $\binom{l}{k}$ est le coefficient binomial est n est la somme de toutes les cellules. Pour les réponses du tableau 6.4 :

$$p = \frac{\binom{4}{3}\binom{4}{1}}{\binom{8}{4}} = 0.229$$

qui correspond à la probabilité d'obtenir les résultats du tableau 6.4 par hasard. Ces résultats n'ont pas établi d'association entre les réponses de la goûteuse et la réalité.

6.5 Validation des résultats de notre évaluation avec le test exact de Fisher

Dans notre évaluation, nous avons demandé aux 54 juges de distinguer pour chacun des 12 résumés s'il avait été crée par un être humain ou par une machine. Un tableau de contingence (tableau 6.5) a été créé pour chaque juge. Un test exact de Fisher a été appliqué aux réponses obtenues.

L'hypothèse nulle dans nos tests (H_0, hypothèse d'indépendance) est qu'il n'y a pas d'association entre les réponses et l'origine du résumé. L'hypothèse alternative (H_1) est donc qu'il existe une association positive.

Réponse du juge	Origine du résumé Humain	Machine
Humain	a	b
Machine	c	d

TABLE 6.5 – *Tableau de contingence pour l'évaluation de résumés avec le test statistique exact de Fisher.*

Nous avons utilisé la fonction *fisher.test* du logiciel R pour calculer la valeur p. On refuse H_0 si la valeur de p est basse. Dans la version du test en R, cette valeur est calculée en additionnant les probabilités de toutes les tables ayant une probabilité inférieure ou égale à celle de la table de contingence observée. Dans nos expériences nous utilisons la configuration standard du test : deux queues ayant une intervalle de confiance de 95%.

Le tableau 6.6 montre les tableaux de contingence et les valeurs de p associées à l'évaluation du jeu d'imitation, adapté au résumé automatique pour les 54 juges. Parmi les 54 juges, seulement une personne (Juge 1 dans le tableau) a présenté évidence statistique d'avoir bien distingué entre les deux types de résumé.

En ce qui concerne les 53 juges restants dans le tableau 6.6, il n'y a pas d'arguments statistiques pour rejeter l'hypothèse que leurs choix aient pu être obtenus par hasard. Pourtant il n'est pas vraisemblable que 53 juges aient décidé de répondre de manière aléatoire. Nous nous penchons donc par le fait qu'ils ont trouvé autant de qualité dans le résumé manuels que dans les automatiques.

Le fait qu'un juge ait été capable de distinguer les deux types de résumés implique qu'il y a des caractéristiques du résumé automatique qui pourraient être perçues. Après une lecture minutieuse, nous avons analysé les caractéristiques langagières des résumés. Nous y trouvons que la présence ou absence de marqueurs discursifs peut donner une impression d'artificialité. Par exemple, le texte *La Tundra* (automatique) se compose de phrases séparées uniquement liées par le thème. Il y a un manque de cohésion entre les phrases. En revanche, le texte *Introducción a las matemáticas* (manuel), présente une grande quantité d'exemples sur l'utilité des mathématiques. L'emphase mise sur l'énumération d'exemples est propre au raisonnement logique de l'homme. Ainsi, l'annotateur a décidé de conserver les exemples car il les a trouvé importants.

6.6 Évaluation de résumés selon le type de segmentation et la taille

Nous utilisons le test exact de Fisher afin de vérifier si on observe des résultats différents selon le segmenteur automatique employé dans l'évaluation de l'origine des résumés par les juges. Le tableau de contingence 6.7 montre le nombre de fois que les juges ont identifié correctement et incorrectement les résumés produits par la machine selon le segmenteur. L'idée est de savoir s'il y a suffisamment d'évidence pour affirmer

Juge id	Contingence		p	H_0	Juge id	Contingence		p	H_0	Juge id	Contingence		p	H_0
Juge 1	4	0	0.030	faux	Juge 2	3	2	0.500	vrai	Juge 3	5	5	0.772	vrai
	2	6				3	4				1	1		
Juge 4	1	5	0.998	vrai	Juge 5	3	3	0.716	vrai	Juge 6	4	4	0.727	vrai
	5	1				3	3				2	2		
Juge 7	5	5	0.772	vrai	Juge 8	4	2	0.283	vrai	Juge 9	5	3	0.272	vrai
	1	1				2	4				1	3		
Juge 10	1	3	0.969	vrai	Juge 11	3	4	0.878	vrai	Juge 12	4	2	0.283	vrai
	5	3				3	2				2	4		
Juge 13	3	3	0.716	vrai	Juge 14	3	3	0.716	vrai	Juge 15	2	3	0.878	vrai
	3	3				3	3				4	3		
Juge 16	3	5	0.969	vrai	Juge 17	1	3	0.969	vrai	Juge 18	3	2	0.500	vrai
	3	1				5	3				3	4		
Juge 19	3	2	0.500	vrai	Juge 20	2	4	0.960	vrai	Juge 21	3	3	0.716	vrai
	3	4				4	2				3	3		
Juge 22	3	4	0.878	vrai	Juge 23	2	6	1.000	vrai	Juge 24	1	4	0.992	vrai
	3	2				4	0				5	2		
Juge 25	2	2	0.727	vrai	Juge 26	4	2	0.283	vrai	Juge 27	2	4	0.960	vrai
	4	4				2	4				4	2		
Juge 28	4	4	0.727	vrai	Juge 29	1	3	0.969	vrai	Juge 30	3	4	0.878	vrai
	2	2				5	3				3	2		
Juge 31	5	5	0.772	vrai	Juge 32	2	4	0.960	vrai	Juge 33	3	4	0.878	vrai
	1	1				4	2				3	2		
Juge 34	2	4	0.960	vrai	Juge 35	3	4	0.878	vrai	Juge 36	2	4	0.960	vrai
	4	2				3	2				4	2		
Juge 37	4	5	0.909	vrai	Juge 38	3	4	0.878	vrai	Juge 39	4	4	0.727	vrai
	2	1				3	2				2	2		
Juge 40	4	2	0.283	vrai	Juge 41	4	4	0.727	vrai	Juge 42	1	5	0.998	vrai
	2	4				2	2				5	1		
Juge 43	2	3	0.878	vrai	Juge 44	2	6	1.000	vrai	Juge 45	5	6	1.000	vrai
	4	3				4	0				1	0		
Juge 46	4	3	0.500	vrai	Juge 47	2	3	0.878	vrai	Juge 48	4	2	0.283	vrai
	2	3				4	3				2	4		
Juge 49	3	3	0.716	vrai	Juge 50	2	3	0.878	vrai	Juge 51	4	3	0.500	vrai
	3	3				4	3				2	3		
Juge 52	2	5	0.992	vrai	Juge 53	2	5	0.992	vrai	Juge 54	3	3	0.716	vrai
	4	1				4	1				3	3		

TABLE 6.6 – *Résultats du test de Turing orienté vers l'évaluation de résumé automatique avec 54 juges.*

qu'une segmentation particulière permet plus facilement d'identifier l'origine des résumés. En effet, l'hypothèse nulle est que le degré d'identification est indépendant du type de segmentation. Les résultats donnent un $p - value = 0.4965$ au 95% avec l'intervalle de confiance de $[0.63; 2.76]$. Nous acceptons donc H_0 car $p - value > 0.05$: le fait qu'un résumé ait été segmenté avec DiSeg ou Coseg ne détermine pas l'identification par les juges.

Segmenteur	Origine correctement identifiée	Origine erronément identifiée
DiSeg	45	63
CoSeg	19	35

TABLE 6.7 – *Évaluation de l'influence de type de segmentation pour l'identification des résumés.*

Pour vérifier l'influence du taux de compression du résumé dans les choix des juges nous utilisons le test du χ^2. Cette fois, nous ne pouvons pas utiliser le test exact de Fisher car le tableau de contingence associé est 3×2 (voir tableau 6.8) et ce test ne peut être réalisé que pour des tableaux 2×2. Ainsi, les résultats du test donnent un $p - value = 0.05476$, à peine supérieure à la valeur critique, ce qui nous a conduit à réaliser d'autres analyses. En effet, si l'on compare les valeurs espérées sous l'hypothèse nulle dans le tableau 6.8, on vérifie que, pour les résumés avec $\tau > 50\%$, il a été plus difficile d'identifier l'origine artificielle. Ce fait peut être aussi confirmé dans le tableau 6.9 des variances résiduelles normalisées où l'écart pour les résumés avec $\tau > 50\%$ est plus de deux fois supérieur à la moyenne. On peut inférer que pour le juges, il a été beaucoup plus difficile d'identifier correctement un résumé produit automatiquement lorsqu'il avait était moins compressé.

Catégorie	Correctement identifiée (observée)	Erronément identifiée (observée)	Correctement identifiée (espérée)	Erronément identifiée (espérée)
$\tau < 50\%$	27	27	25	29
$\tau \approx 50\%$	30	24	25	29
$\tau > 50\%$	18	36	25	29

TABLE 6.8 – *Évaluation de l'influence du τ pour l'identification des résumés.*

Catégorie	Correctement identifiée	Erronément identifiée
$\tau < 50\%$	0.668	-0.668
$\tau \approx 50\%$	1.671	-1.671
$\tau > 50\%$	-2.339	2.339

TABLE 6.9 – *Écart-type des variances résiduelles par rapport au τ pour l'identification des résumés.*

6.7 Conclusions du chapitre

Nous avons traité l'évaluation des résumés automatiques. Comme nous l'avons montré auparavant, les méthodes d'évaluations ROUGE, BLEU et FRESA, efficacement utilisées pour évaluer des résumés automatiques par extraction, ne permettent pas d'évaluer des résumés par compression de phrases étant donné qu'elles ne tiennent pas compte de la grammaticalité.

Nous avons donc proposé une méthode inspirée du test de Turing dans laquelle des juges humains devaient dévoiler l'origine de plusieurs résumés. Moyennant le test exact de Fisher, nous avons calculé la fiabilité des réponses des juges.

Après 54 tests, nous avons obtenu des résultats intéressants. En effet, selon le test exact de Fisher, les réponses d'un seul juge sont très peu probablement liées au hasard. Pour le reste des réponses, nous pouvons faire deux interprétations différentes : soit les 53 juges restants ont donné des réponses aléatoires ; soit ils n'ont pas trouvé assez d'évidences pour distinguer l'origine manuelle ou automatique des résumés. Or, il est peu vraisemblable que 53 juges, tous volontaires motivés, aient décidé de faire le test au hasard, nous penchons donc sur la deuxième interprétation : notre système a réussi à les tromper. Dans ce cas là, il faudrait toujours considérer que même si le système à trompé la majorité des juges, un a été capable de repérer l'artificialité des résumés. Ainsi, comme nous l'avions prévu, la contrainte principale de notre système est le manque de cohésion textuelle qui est d'autant plus évidente que le texte est court.

Conclusions et perspectives de recherche

Notre conclusion principale est que la tâche de compression de phrases doit être reconsidérée à partir des nouvelles hypothèses concernant l'informativité. De manière générale, les gens hésitent beaucoup par rapport à l'identification des fragments importants. Nous avons observé que ce phénomène est indépendant de la granularité des segments et de la longueur de la phrase. Face au doute, les annotateurs ont une forte tendance à conserver l'information. Il y a donc un certain niveau de subjectivité inhérent à la tâche. À partir d'un grand corpus, nous avons obligé un algorithme à « *apprendre* » ce que l'humain considère important. Le résultat a été un résumeur qui décide d'éliminer des segments discursifs à des valeurs de probabilité très basses. D'ailleurs il a appris à tout laisser à partir d'une valeur de probabilité d'élimination (selon les réponses de l'annotation). Néanmoins, certaines caractéristiques de la phrase et de ses segments discursifs sont corrélées avec la décision d'éliminer ou préserver un segment. Nous considérons donc que les résultats de cette étude ouvrent la discussion par rapport à la subjectivité de l'informativité dans le résumé automatique : Devrait-on la considérer dans des modèles pour le résumé automatique ?

Notre étude a révélé que la contrainte principale de la méthode proposée est le manque de cohésion. En effet, nous avons prouvé que plus les résumés sont courts, plus les juges distinguent leur origine artificielle. Nous avons trouvé les deux raisons principales du manque de cohésion :

1. La suppression excessive de marqueurs discursifs contenus dans les segments éliminés.
2. L'incapacité à identifier les synonymes et les paraphrases.

La première raison s'explique par l'utilisation de la segmentation discursive : les marqueurs discursifs disparaissent en effaçant un segment. Pour résoudre ce problème il est parfois possible de laisser le marqueur discursif mais, il faut savoir, en tout moment, quelle est sa fonction dans la phrase ; plus précisément, à quelle type de relation il correspond. Bien que la désambiguïsation de segments n'est pas une tâche triviale, il existe actuellement des études linguistiques visant à surmonter ce défi (da Cunha, 2013). Quant à la deuxième raison, elle s'explique par le recours à l'énergie textuelle : la méthode lie des mots lemmatisés. Elle est donc incapable de reconnaître des mots ou

des segments avec la même signification écrits de manières différentes. Pour considérer la sémantique on utilise un thésaurus. Une étude récente propose une amélioration du modèle afin de mesurer la proximité sémantique plutôt que la proximité lexicale (Boudouma et al., 2012). Nous comptons approfondir dans ces deux voies.

Bien que nous nous sommes restreints au domaine du résumé automatique, la méthode proposée peut aussi être vue comme une «*simplification textuelle*». La différence essentielle est que dans le résumé, il y aura des phrases qui seront extraites pour conformer le résumé final mais quelques-unes seront annulées entièrement. En ce qui concerne la simplification, on garderait, au moins, un segment de chaque phrase. La simplification textuelle pourrait être utilisée comme un pré-traitement pour l'indexation d'un moteur de recherche. Elle permettrait la navigation des documents via des phrases simplifiées. L'idée principale est de relier des fragments complexes des documents lorsqu'ils partagent des simplifications proches.

Nous avons montré que la segmentation discursive intra-phrase est intéressante pour évaluer des segments contenant des informations importantes et également pour générer des versions grammaticales des phrases compressées. De ce point de vue, contrairement aux approches classiques de résumé par élimination des segments discursifs, nous avons montré qu'il n'est pas nécessaire d'élaborer une analyse discursive approfondie du texte en entier afin d'identifier les segments éliminables. Dans notre approche, aucun de ces traitements n'est nécessaire :

1. Identification des noyaux/satellites.
2. Identification de type de relation rhétorique.
3. Création de l'arbre rhétorique.

Nous pensons même que les concepts et les méthodes abordés dans cette étude peuvent être appliqués au domaine de l'analyse rhétorique. Par exemple, nous envisageons d'utiliser l'énergie textuelle pour l'identification des noyaux/satellites. Nous avons aussi développé CoSeg, un nouveau segmenteur discursif en espagnol optimisé pour la compression de phrases. Finalement, notre recherche a abouti à la création d'un segmenteur discursif multilingue qui utilise peu de ressources linguistiques et nous avons introduit un nouveau protocole d'évaluation de performance des segmenteurs discursifs (Saksik et al., 2013).

Nous avons adapté la méthode d'énergie textuelle pour l'appliquer à la tâche de compression de phrases, jusqu'à présent utilisée uniquement soit pour l'élaboration de résumés par extraction, soit pour le regroupement sémantique de définitions (Molina, 2009; Molina et al., 2010b). Nous avons montré que la transformation logarithme améliore deux aspects de l'énergie textuelle : elle corrige sa distribution et la borne entre 0 et 1.

En ce qui concerne l'évaluation, nous avons analysé les raisons pour lesquelles les évaluations ROUGE, BLEU et FRESA ne sont pas bien adaptées à la compression des phrases. Nous avons revisité le test de Turing pour évaluer la qualité des résumés automatiques. Nous avons établi un protocole expérimental et nous l'avons validé grâce à

une méthode combinatoire exacte qui n'avait pas été utilisée à ce propos, mais qui ouvre des perspectives prometteuses. Nous considérons que cette méthode peut être utilisée pour l'évaluation d'autres tâches du *Traitement Automatique de la Langue Naturelle*. Par exemple, pour évaluer la traduction automatique, la génération automatique, la simplification automatique et la génération de reformulation et paraphrase.

Nous avons approfondi l'étude de compression de phrases en espagnol. Ainsi, nous avons constitué le premier corpus de phrases compressées en cette langue. Ces données sont disponibles sur le Web afin d'encourager des futures recherches : http://molina.talne.eu/sentence_compression/data/. Comme nous avons décidé de faire appel aux « *sciences citoyennes* », nous avons développé une plate-forme d'annotation permettant d'obtenir de nouvelles données de manière efficace (Molina, 2013). La plate-forme s'est avérée tellement pratique et flexible qu'elle a été utilisée dans le domaine d'analyse d'opinion, dans le cadre du projet ANR Imagiweb [5] où elle a servi à l'annotation de *tweets* [6].

Bien qu'il reste encore un grand chemin à parcourir pour réussir à élaborer un résumeur automatique d'une qualité comparable à celle d'un être humain, nous considérons que la présente étude a contribué à éclaircir un peu plus ce domaine.

5. http://www.agence-nationale-recherche.fr/en/anr-funded-project/?tx_lwmsuivibilan_pi2[CODE]=ANR-12-CORD-0002
6. http://molina.talne.eu/sentaatool/info/systeme_description.html

Bibliographie

(Afantenos et al., 2010) S. Afantenos, P. Denis, et L. Danlos, 2010. Learning recursive segments for discourse parsing. *Cornell University ArXiv :1003.5372,Computation and Language (cs.CL)*. 36

(Agresti, 2002) A. Agresti, 2002. *Categorical data analysis*, Volume 359. Wiley interscience. 100

(Aluísio et al., 2008) S. M. Aluísio, L. Specia, T. A. Pardo, E. G. Maziero, et R. P. Fortes, 2008. Towards brazilian portuguese automatic text simplification systems. Dans les actes de *8th ACM symposium on Document engineering*, 240–248. ACM. 26

(Amigó et al., 2005) E. Amigó, J. Gonzalo, A. Peñas, et F. Verdejo, 2005. Qarla : a framework for the evaluation of text summarization systems. Dans les actes de *43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, Ann Arbor, MI, Etats-Unis, 280–289. ACL. 96

(Amit et al., 1987) D. J. Amit, H. Gutfreund, et H. Sompolinsky, 1987. Statistical mechanics of neural networks near saturation. *Annals of Physics 173*(1), 30–67. 60

(Barzilay et al., 1999) R. Barzilay, K. R. McKeown, et M. Elhadad, 1999. Information fusion in the context of multi-document summarization. Dans les actes de *37th annual meeting of the Association for Computational Linguistics*, 550–557. ACL. 23

(Baxendale, 1958) P. B. Baxendale, 1958. Machine-made index for technical literature —- an experiment. *IBM Journal of Research and Development 2*(4), 354–361. 15

(Bellot et al., 2013) P. Bellot, V. Moriceau, J. Mothe, E. SanJuan, et X. Tannier, 2013. Contextualisation de textes courts : le cas des tweets. 63

(Boudin, 2008) F. Boudin, 2008. *Exploration d'approches statistiques pour le résumé automatique de texte*. doctorat en informatique, Laboratoire Informatique d'Avignon, Université d'Avignon et des Pays de Vaucluse, Avignon, France. 17

(Boudin et Torres-Moreno, 2007) F. Boudin et J.-M. Torres-Moreno, 2007. NEO-CORTEX : A Performant User-Oriented Multi-Document Summarization System. Dans les actes de *Computational Linguistics and Intelligent Text Processing (CICLing'07)*, Volume 4394 de *Lecture Notes in Computer Science*, 551–562. Springer. 27

(Boudouma et al., 2012) R. Boudouma, R. Touahni, et R. Messoussi, 2012. L'energie conceptuelle de mémoires auto associatives à base d'ontologie du domaine. Dans les actes de *The 7th IEEE international conference Sciences of Electronics Technologies Information and Telecommunication*. IEEE. 108

(Box et Cox, 1964) G. Box et D. Cox, 1964. An analysis of transformations. *Journal of the Royal Statistical Society. Series B (Methodological)*, 211–252. 68

(Castellón et al., 1998) I. Castellón, M. Civit, et J. Atserias, 1998. Syntactic parsing of unrestricted spanish text. Dans les actes de *Proceedings First International Conference on Language Resources and Evaluation (LREC'98), Granada, Spain*. 36

(Chamberlain et al., 2009) J. Chamberlain, M. Poesio, et U. Kruschwitz, 2009. A new life for a dead parrot : Incentive structures in the phrase detectives game. Dans les actes de *Webcentives Workshop at WWW*, Volume 9. 77

(Chen et Goodman, 1999) S. Chen et J. Goodman, 1999. An empirical study of smoothing techniques for language modeling. *Computer Speech & Language 13*(4), 359–393. 52, 53

(Clarke et Lapata, 2006) J. Clarke et M. Lapata, 2006. Constraint-based sentence compression : An integer programming approach. Dans les actes de *COLING/ACL'06 on Main Conference Poster Sessions*, Sydney, Australie, 144–151. 26

(Cohn et Lapata, 2009) T. Cohn et M. Lapata, 2009. Sentence Compression as Tree Transduction. *Journal of Artificial Intelligence Research 34*, 637–674. 26

(Collins, 1996) M. J. Collins, 1996. A new statistical parser based on bigram lexical dependencies. Dans les actes de *34th annual meeting on Association for Computational Linguistics*, 184–191. ACL. 23, 25

(Cook et Weisberg, 2009) R. D. Cook et S. Weisberg, 2009. *Applied regression including computing and graphics*, Volume 488. Wiley-Interscience. 84, 85

(Corston-Oliver et Dolan, 1999) S. H. Corston-Oliver et W. B. Dolan, 1999. Less is more : Eliminating index terms from subordinate clauses. Dans les actes de *37th annual meeting of the Association for Computational Linguistics*, 349–356. ACL. 24

(da Cunha, 2013) I. da Cunha, 2013. A symbolic corpus-based approach to detect and solve the ambiguity of discourse markers. Dans les actes de *Computational Linguistics and Intelligent Text Processing*. Springer. 107

(da Cunha et al., 2007) I. da Cunha, S. Fernández, P. Velázquez-Morales, J. Vivaldi, E. SanJuan, et J. M. Torres-Moreno, 2007. A new hybrid summarizer based on vector space model, statistical physics and linguistics. Dans les actes de *6th Mexican International Conference on Advances in Artificial Intelligence (MICAI'07)*, Aguascalientes, Mexique, 872–882. Springer-Verlag. 27

(da Cunha et al., 2010) I. da Cunha, E. SanJuan, J.-M. Torres-Moreno, M. Lloberes, et I. Castellón, 2010. Discourse segmentation for spanish based on shallow parsing. Dans G. Sidorov, A. Hernández Aguirre, et C. Reyes García (Eds.), *Advances in Artificial Intelligence*, Volume 6437 de *Lecture Notes in Computer Science*, 13–23. Springer Berlin / Heidelberg. 36, 42

(da Cunha et al., 2012) I. da Cunha, E. SanJuan, J.-M. Torres-Moreno, M. Lloberes, et I. Castellón, 2012. Diseg 1.0 : The first system for spanish discourse segmentation. *Expert Systems with Applications 39*(2), 1671–1678. 36

(da Cunha et Wanner, 2005) I. da Cunha et L. Wanner, 2005. Towards the Automatic Summarization of Medical Articles in Spanish : Integration of textual, lexical, discursive and syntactic criteria. Dans les actes de *Workshop Crossing Barriers in Text Summarization Research*, Borovets, Bulgarie. Recent Advances in Natural Language Processing. 27

(da Cunha et al., 2007) I. da Cunha, L. Wanner, et M. T. Cabré, 2007. Summarization of specialized discourse : The case of medical articles in Spanish. *Terminology 13*(2), 249–286. 27

(Daelemans et al., 2004) W. Daelemans, A. Höthker, et E. T. K. Sang, 2004. Automatic sentence simplification for subtitling in dutch and english. Dans les actes de *4th International Conference on Language Resources and Evaluation*, 1045–1048. 24, 26

(Dalianis, 2001) H. Dalianis, 2001. Swesum : A text summerizer for swedish. Rapport technique TRITA-NA-P0015 IPLab-174, Royal Institute of Technology, Stockholm Sweden. 27

(Das et Martins, 2007) D. Das et A. F. Martins, 2007. A survey on automatic text summarization. *Literature Survey for the Language and Statistics II course at CMU 4*, 192–195. 17

(de Loupy et al., 2010) C. de Loupy, M. Guégan, C. Ayache, S. Seng, et J.-M. Torres-Moreno, 2010. A French Human Reference Corpus for Multi-Document Summarization and Sentence Compression. Dans N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner, et D. Tapias (Eds.), *7th International Conference on Language Resources and Evaluation*, Valletta, Malte. ELRA. 28

(Deerwester et al., 1990) S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, et R. Harshman, 1990. Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Science 41*(6), 391–407. 32

(Edmundson, 1969) H. P. Edmundson, 1969. New Methods in Automatic Extraction. *Journal of the Association for Computing Machinery 16*(2), 264–285. 16, 81, 96

(Elhadad, 1992) M. Elhadad, 1992. *Using argumentation to control lexical choice : A functional unification-based approach*. Thèse de Doctorat, Computer Science Department, Columbia University, New York. 23

(Fernández, 2009) S. Fernández, 2009. *Applications exploratoires des modèles de spins au Traitement Automatique de la Langue*. doctorat en physique statistique, Département de Physique de la Matière et des Matériaux, Université Henri Poincaré, Nancy, France. 60, 61, 62

(Fernández et al., 2007a) S. Fernández, E. SanJuan, et J.-M. Torres-Moreno, 2007a. Energie textuelle des mémoires associatives. Dans les actes de *Proceedings de la conférénce Traitement Automatique de la Langue Naturelle (TALN'07)*, Volume 1, Toulouse, France, 25–34. 60, 62

(Fernández et al., 2007b) S. Fernández, E. SanJuan, et J.-M. Torres-Moreno, 2007b. Textual Energy of Associative Memories : performants applications of Enertex algorithm in text summarization and topic segmentation. Dans les actes de *Mexican International Conference on Artificial Intelligence (MICAI'07)*, Aguascalientes, Mexique, 861–871. Springer-Verlag. 60

(Fernández et al., 2008) S. Fernández, E. SanJuan, et J.-M. Torres-Moreno, 2008. Enertex : un système basé sur l'énergie textuelle. Dans les actes de *Proceedings de la conférence Traitement Automatique des Langues Naturelles (TALN'08)*, Avignon, France, 99–108. 27, 61, 62

(Fernández et al., 2009) S. Fernández, E. SanJuan, et J.-M. Torres-Moreno, 2009. Résumés de texte par extraction de phrases, algorithmes de graphe et énergie textuelle. Dans les actes de *XVI^{es} rencontres de la Société francophone de classification*, Grenoble, France, 101–104. 62, 63

(Fisher et al., 1935) R. A. Fisher et al., 1935. The design of experiments. *The design of experiments.*. 101

(Galley et McKeown, 2007) M. Galley et K. McKeown, 2007. Lexicalized markov grammars for sentence compression. *the Proceedings of NAACL/HLT*, 180–187. 26

(Grefenstette, 1998) G. Grefenstette, 1998. Producing intelligent telegraphic text reduction to provide an audio scanning service for the blind. Dans les actes de *AAAI Spring Symposium on Intelligent Text summarization (Working notes)*, Stanford University, CA, Etats-Unis, 111–118. 22, 24

(Harnad, 2000) S. Harnad, 2000. Minds, machines and turing. *Journal of Logic, Language and Information* *9*(4), 425–445. 99

(Hertz et al., 1991) J. A. Hertz, A. S. Krogh, et R. G. Palmer, 1991. *Introduction to the theory of neural computation*, Volume 1. Westview press. 61, 62

(Hopfield et Tank, 1985) J. J. Hopfield et D. W. Tank, 1985. "neural" computation of decisions in optimization problems. *Biological cybernetics* *52*(3), 141–152. 60

(Iordanskaja et al., 1991) L. Iordanskaja, R. Kittredge, et A. Polguere, 1991. Lexical selection and paraphrase in a meaning-text generation model. *Natural language generation in artificial intelligence and computational linguistics*, 293–312. 23

(Irwin, 1995) A. Irwin, 1995. *Citizen science : a study of people, expertise, and sustainable development*. Burns & Oates. 77

(Ježek et Steinberger, 2008) K. Ježek et J. Steinberger, 2008. Automatic text summarization (the state of the art 2007 and new challenges). *Proceedings of Znalosti 2008*, 1–12. 17

(Jing et McKeown, 1999) H. Jing et K. R. McKeown, 1999. The decomposition of human-written summary sentences. Dans les actes de *22nd annual international ACM SIGIR conference on Research and development in information retrieval*, 129–136. ACM. 18, 22, 32

(Jing et Croft, 1994) Y. Jing et W. B. Croft, 1994. An association thesaurus for information retrieval. Dans les actes de *Proceedings of RIAO*, Volume 94, 146–160. 22

(Jonnalagadda et Gonzalez, 2010) S. Jonnalagadda et G. Gonzalez, 2010. Sentence simplification aids protein-protein interaction extraction. *Cornell University ArXiv :1001.4273*. 51

(Ketui et al., 2012) N. Ketui, T. Theeramunkong, et C. Onsuwan, 2012. A rule-based method for thai elementary discourse unit segmentation (ted-seg). Dans les actes de *Knowledge, Information and Creativity Support Systems (KICSS), 2012 Seventh International Conference on*, 195–202. IEEE. 36

(Kittredge et Mel'cuk, 1983) R. Kittredge et I. Mel'cuk, 1983. Towards a computable model of meaning-text relations within a natural sublanguage. Dans les actes de *Proc. of 8th International Joint Conference on Artificial Intelligence (IJCAI'83)*, 657–659. 23

(Klein et Manning, 2003) D. Klein et C. D. Manning, 2003. Accurate unlexicalized parsing. Dans les actes de *41st Annual Meeting on Association for Computational Linguistics*, Volume 1, 423–430. ACL. 51

(Knight et Marcu, 2000) K. Knight et D. Marcu, 2000. Statistics-based summarization – step one : Sentence compression. Dans les actes de *17th National Conference on Artificial Intelligence and 12th Conference on Innovative Applications of Artificial Intelligence*, Austin, TX, Etats-Unis, 703–710. 24, 25

(Knight et Marcu, 2002) K. Knight et D. Marcu, 2002. Summarization beyond sentence extraction : a probabilistic approach to sentence compression. *Artificial Intelligence 139*(1), 91–107. 24

(La Rocca, 2012) M. La Rocca, 2012. *Los marcadores del discurso del español. Un inventario comparado*. Aracne Editrice. 35

(Labadié et al., 2008) A. M. Labadié et al., 2008. *Segmentation thématique de texte linéaire et non-supervisée : Détection active et passive des frontières thématiques en Français*. Thèse de Doctorat, Université Montpellier II-Sciences et Techniques du Languedoc. 63

(Landauer et al., 1998) T. K. Landauer, P. W. Foltz, et D. Laham, 1998. An introduction to latent semantic analysis. *Discourse Processes 25*(2), 259–284. 32

(Lara et al., 1979) L. Lara, R. Chande, et M. Hidalgo, 1979. *Investigaciones lingüísticas en lexicografía*, Volume 89. Colegio de México, Centro de Estudios Lingüísticos y Literarios. 53

(Lin, 2004) C.-Y. Lin, 2004. ROUGE : A Package for Automatic Evaluation of Summaries. Dans M.-F. Moens et S. Szpakowicz (Eds.), *Workshop Text Summarization Branches Out (ACL'04)*, Barcelone, Espagne, 74–81. ACL. 29, 96

(Linke-Ellis, 1999) N. Linke-Ellis, 1999. Closed captioning in america : Looking beyond compliance. Dans les actes de *TAO Workshop on TV Closed Captions for the Hearing Impaired People, Tokyo, Japan*, 43–59. 24

(Louis et Nenkova, 2008) A. Louis et A. Nenkova, 2008. Automatic Summary Evaluation without Human Models. Dans les actes de *First Text Analysis Conference (TAC'08)*, Gaithersburg, MD, Etats-Unis. 56, 97

(Luhn, 1957) H. P. Luhn, 1957. A statistical approach to mechanized encoding and searching of literary information. *IBM Journal of research and development* *1*(4), 309–317. 15

(Luhn, 1958) H. P. Luhn, 1958. The Automatic Creation of Literature Abstracts. *IBM Journal of Research and Development* *2*(2), 159–165. 15

(Mani et al., 1999) I. Mani, D. House, G. Klein, L. Hirschman, T. Firmin, et B. Sundheim, 1999. The tipster summac text summarization evaluation. Dans les actes de *ninth conference on European chapter of the Association for Computational Linguistics*, 77–85. ACL. 96

(Mani et Maybury, 1999) I. Mani et M. T. Maybury, 1999. *Advances in automatic text summarization*. MIT Press. 15

(Mann et Thompson, 1988) W. C. Mann et S. A. Thompson, 1988. Rhetorical Structure Theory : Toward a functional theory of text organization. *Text* *8*(3), 243–281. 28, 33

(Manning et Schütze, 1999) C. D. Manning et H. Schütze, 1999. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge. 52

(Marcu, 2000) D. Marcu, 2000. The rhetorical parsing of unrestricted texts : A surface-based approach. *Computational Linguistics* *26*(3), 395–448. 41

(Maziero et al., 2007) E. Maziero, T. Pardo, et M. Nunes, 2007. Identificação automática de segmentos discursivos : o uso do parser palavras. Série de relatórios do núcleo interinstitucional de lingüística computacional, Universidade de Sao Paulo, São Carlos, Brésil. 36

(McDonald, 2006) R. McDonald, 2006. Discriminative sentence compression with soft syntactic evidence. Dans les actes de *Proceedings of EACL*, Volume 6, 297–304. 26

(Mihalcea, 2004) R. Mihalcea, 2004. Graph-based ranking algorithms for sentence extraction, applied to text summarization. Dans les actes de *Proceedings of the ACL 2004*

on Interactive poster and demonstration sessions, Morristown, NJ, USA, 20. Association for Computational Linguistics. 62

(Molina, 2009) A. Molina, 2009. Agrupamiento semántico de contextos definitorios. Mémoire de Master, Universidad Nacional Autónoma de México – Posgrado en Ciencia e Ingeniería de la Computación, México. 108

(Molina, 2013) A. Molina, 2013. Sistemas web colaborativos para la recopilación de datos bajo el paradigma de ciencia ciudadana. *Komputer Sapiens 1*, 6–8. 77, 109

(Molina et al., 2010a) A. Molina, I. da Cunha, J.-M. Torres-Moreno, et P. Velazquez-Morales, 2010a. La compresión de frases : un recurso para la optimización de resumen automático de documentos. *Linguamática 2*(3), 13–27. 26, 32, 97

(Molina et al., 2010b) A. Molina, G. Sierra, et J.-M. Torres-Moreno, 2010b. La energía textual como medida de distancia en agrupamiento de definiciones. Dans les actes de *Journées d'Analyse Statistique de Documents (JADT'10)*, Rome, Italie. 63, 108

(Molina et al., 2012) A. Molina, J.-M. Torres-Moreno, I. da Cunha, E. SanJuan, et G. Sierra, 2012. Sentence compression in spanish driven by discourse segmentation and language models. *Cornell University ArXiv :1212.3493, Computation and Language (cs.CL), Information Retrieval (cs.IR)*. 52, 97, 98

(Molina et al., 2011) A. Molina, J.-M. Torres-Moreno, E. SanJuan, I. da Cunha, G. Sierra, et P. Velázquez-Morales, 2011. Discourse segmentation for sentence compression. Dans les actes de *Advances in Artificial Intelligence*, LNCS, 316–327. Springer-Verlag, Berlin, Heidelberg. 32, 36, 37, 54, 55, 121

(Orasan et al., 2007) C. Orasan, L. Hasler, et S. St, 2007. Computer-aided summarisation : how much does it really help. *Proceedings of Recent Advances in Natural Language Processing (RANLP 2007)*, 437–444. 96

(Papineni et al., 2002) K. Papineni, S. Roukos, T. Ward, , et W. J. Zhu, 2002. BLEU : a method for automatic evaluation of machine translation. Dans les actes de *40th Annual Meeting of the Association for Computational Linguistics (ACL'02)*, Philadelphie, PA, Etats-Unis, 311–318. ACL. 29, 97

(Radev et al., 2002) D. Radev, E. Hovy, et K. McKeown, 2002. Introduction to the special issue on summarization. *Computational linguistics 28*(4), 399–408. 17

(Rayward et Bowden, 2004) W. B. Rayward et M. E. Bowden, 2004. *The History and heritage of scientific and technological information systems*. American Society of Information Science and Technology and the Chemical Heritage Foundation. 15

(Robert-Ribes et al., 1999) J. Robert-Ribes, S. Pfeiffer, R. Ellison, et D. Burnham, 1999. Semi-automatic captioning of tv programs, an australian perspective. Dans les actes de *Workshop on TV closed captions for the hearing impaired people*, 87–100. 24

(Roze et al., 2012) C. Roze, L. Danlos, et P. Muller, 2012. Lexconn : a french lexicon of discourse connectives. *Discours. Revue de linguistique, psycholinguistique et informatique* (10). 35

(Saggion et Lapalme, 2000) H. Saggion et G. Lapalme, 2000. Concept identification and presentation in the context of technical text summarization. Dans les actes de *ANLP/NAACL Workshop on Automatic Summarization*, Seattle, WA, Etats-Unis, 1–10. ACL. 96

(Saggion et al., 2010) H. Saggion, J.-M. Torres-Moreno, I. da Cunha, et E. SanJuan, 2010. Multilingual summarization evaluation without human models. Dans les actes de *23rd International Conference on Computational Linguistics : Posters (COLING'10)*, Beijing, Chine, 1059–1067. ACL. 56, 97

(Saksik et al., 2013) R. Saksik, A. Molina, L. Andréa, et Torres-Moreno, 2013. Segmentaçao discursiva automática : uma avaliaçao preliminar em francês. Dans les actes de *4th Meeting RST and Discourse Studies, STIL 2013 Symposium in Information and Human Language Technology*. 44, 108

(Schmid, 1995) H. Schmid, 1995. Treetagger–a language independent part-of-speech tagger. *Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart*, 43. 28

(Siddharthan, 2006) A. Siddharthan, 2006. Syntactic simplification and text cohesion. *Research on Language and Computation* 4(1), 77–109. 52

(Sierra et al., 2010) G. Sierra, J.-M. Torres-Moreno, et A. Molina, 2010. Regroupement sémantique de définitions en espagnol. Dans les actes de *Proceedings of Evaluation des méthodes d'extraction de connaissances dans les données (EGC/EvalECD'10)*, Hammamet, Tunisie, 41–50. 63

(Sleator et Temperley, 1995) D. D. Sleator et D. Temperley, 1995. Parsing english with a link grammar. *arXiv preprint cmp-lg/9508004*. 52

(Snow et al., 2008) R. Snow, B. O'Connor, D. Jurafsky, et A. Ng, 2008. Cheap and fast—but is it good ? : evaluating non-expert annotations for natural language tasks. Dans les actes de *Conference on Empirical Methods in Natural Language Processing*, 254–263. ACL. 77

(Soricut et Marcu, 2003) R. Soricut et D. Marcu, 2003. Sentence level discourse parsing using syntactic and lexical information. Dans les actes de *HLT-NAACL*, Edmonton, Canada, 149–156. 32

(Sporleder et Lapata, 2005) C. Sporleder et M. Lapata, 2005. Discourse chunking and its application to sentence compression. Dans les actes de *conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT '05, Stroudsburg, PA, USA, 257–264. ACL. 32

(Spärck-Jones, 2007) K. Spärck-Jones, 2007. Automatic summarising : The state of the art. *Information Processing and Management* 43(6), 1449–1481. 17

(Steinberger et Jezek, 2006) J. Steinberger et K. Jezek, 2006. Sentence compression for the lsa-based summarizer. Dans les actes de *7th International conference on information systems implementation and modelling*, 141–148. 32

(Steinberger et Tesar, 2007) J. Steinberger et R. Tesar, 2007. Knowledge-poor multilingual sentence compression. Dans les actes de *7th Conference on Language Engineering (SOLE'07)*, Le Caire, Egypte, 369–379. 32

(Stolcke, 2002) A. Stolcke, 2002. Srilm – an extensible language modeling toolkit. Dans les actes de *Intl. Conf. on Spoken Language Processing*, Volume 2, Denver, 901–904. 53

(Tofiloski et al., 2009) M. Tofiloski, J. Brooke, et M. Taboada, 2009. A syntactic and lexical-based discourse segmenter. Dans les actes de *ACL-IJCNLP*, 77–80. 34, 36

(Torres-Moreno, 2012) J.-M. Torres-Moreno, 2012. *Résumé automatique de documents : une approche statistique*. Hermès-Lavoisier, France. 17, 63

(Torres-Moreno et al., 2010a) J.-M. Torres-Moreno, H. Saggion, I. da Cunha, et E. SanJuan, 2010a. Summary Evaluation With and Without References. *Polibits : Research journal on Computer science and computer engineering with applications 42*, 13–19. 97

(Torres-Moreno et al., 2010b) J.-M. Torres-Moreno, H. Saggion, I. da Cunha, P. Velazquez-Morales, et E. SanJuan, 2010b. Evaluation automatique de résumés avec et sans réferences. Dans les actes de *Proceedings de la conference Traitement Automatique des Langagues Naturelles (TALN'10)*, Montréal, QC, Canada. ATALA. 56

(Torres-Moreno et al., 2001) J.-M. Torres-Moreno, P. Velázquez-Morales, et J.-G. Meunier, 2001. Cortex : un algorithme pour la condensation automatique des textes. Dans les actes de *Conference de l'Association pour la Recherche Cognitive*, Volume 2, Lyon, France, 365–366. 27

(Turing, 1950) A. M. Turing, 1950. Computing machinery and intelligence. *Mind 59*(236), 433–460. 98

(Vivaldi et Rodríguez, 2001) J. Vivaldi et H. Rodríguez, 2001. Improving term extraction by combining different techniques. *Terminology 7*(1), 31–47. 27

(Wasserman, 2004) L. Wasserman, 2004. *All of statistics : a concise course in statistical inference*. Springer Verlag. 84

(Waszak et Torres-Moreno, 2008) T. Waszak et J.-M. Torres-Moreno, 2008. Compression entropique de phrases contrôlée par un perceptron. Dans les actes de *Journées internationales d'Analyse statistique des Données Textuelles (JADT'08)*, Lyon, France, 1163–1173. 26

(Witbrock et Mittal, 1999) M. J. Witbrock et V. O. Mittal, 1999. Ultra-Summarization : A Statistical Approach to Generating Highly Condensed Non-Extractive Summaries. Dans les actes de *22nd Conference SIGIR'99*, Berkeley, CA, Etats-Unis, 315–316. ACM. 22

(Yousfi-Monod et Prince, 2006) M. Yousfi-Monod et V. Prince, 2006. Compression de phrases par élagage de l'arbre morpho-syntaxique. *Technique et Science Informatiques 25*(4), 437–468. 26, 28

Annexe A

Segments discursifs éliminés

Fragments éliminés par les annotateurs correspondant aux segments discursifs identifiés par DiSeg pour quatre genres : encyclopédique (Wikipédia), journalistique, résumés d'articles scientifiques, littéraire (Molina et al., 2011). Les relations marquées avec * ont été désambiguïsées par le contexte.

Type	Marqueur	Relation	Fragment
N	?	?	confió en su mentora y
?	?	?	lanzando un pase para anotación de 53 yardas que les dio el triunfo
N	y	LISTE*/JOINT	y estar considerado el equipo como candidato a disputar el Campeonato Nacional
S	cuando	CIRCONSTANCE	cuando tenía diecinueve años ,
S	gerundio	ELABORATION	grabando dos canciones en cada oportunidad ,
S	gerundio	CIRCONSTANCE	(estando ambos casados con otras parejas)
S	gerundio	MÉTHODE	afirmando de ella que había tenido solamente cinco parejas , todas conocidas ,
N	mientras que	CONTRASTE	mientras que otras divas de Hollywood se acostaban con cualquiera en la primera noche (si bien manteniéndolo en secreto)
S	gerundio	CONCESSION	(si bien manteniéndolo en secreto)
S	?	ÉVIDENCE	dicen que ,
S	Sí	CONDITION	si se casó ocho veces , fue
S	porque	CAUSE	porque no era proclive a aventuras fugaces
N	y	LISTE*/JOINT	y quería formalizar cada nueva relación con una boda
S	gerundio	ELABORATION	incluso superando al llamado « *animal más bello del mundo* », Ava Gardner.
S	gerundio	MÉTHODE	hablando acerca de la prevención necesaria
S	para	BUT	para combatir el SIDA.
S	pero	ANTITHÈSE	pero no tuvo éxito.
S	gerundio	JUSTIFICATION	alegando que ya no disfrutaba jugar como antes.
S	donde	CIRCONSTANCE	donde ganó la medalla de oro con el Dream Team.
N	y	LISTE/JOINT*	y desmintió los rumores que indicaban que se había realizado otro tipo de cirugías faciales.

S	gerundio	RÉSULTAT	convirtiéndose así en uno de los artistas mejores pagados del mundo.
S	cuando	CIRCONSTANCE	cuando en una entrega de premios lo presentó como « *el auténtico rey del pop , rock y soul* ».
S	gerundio	ELABORATION	diseñando mejoras para el equipo eléctrico traído del otro lado del océano gracias a las ideas de Edison.
S	gerundio	RÉSULTAT	mejorando tanto su servicio como su economía
S	cuando	CIRCONSTANCE	cuando se le denegó un aumento de US $ 25 a la semana.
?	al	?	al unirse este al Cuerpo Militar de Telegrafistas
N	y	LISTE/JOINT*	y dependen de la cantidad que se consuma , el entorno en que se use la droga , la pureza de ésta , la personalidad , el estado de ánimo y las expectativas del usuario .
S	pero	ANTITHÈSE	pero otras quizá no tengan su primer período hasta una etapa más tardía de su adolescencia .
N	y	LISTE/JOINT*	y opera en los rangos de frecuencia de 3,1 a 10,6 GHz .
N	y	LISTE*/JOINT	y se abrieron varias envasadoras en Estados Unidos .
S	por	CAUSE	por estar lejos de las estrellas
N	y	LISTE*/JOINT	y también es muy famosa la nebulosa Cabeza de Caballo , en la constelación de Orión .
S	pero	ANTITHÈSE	pero , a la vez , mitificadora porque convierte al delincuente en héroe .

TABLE *A.1* – *Parties éliminées pour le genre encyclopédique (Wikipédia).*

Type	Marqueur	Relation	Fragment
S	a menos que	CONDITION INVERSE	a menos que no reciban los cuidados y el tratamiento adecuado.
S	aunque	CONCESSION	aunque otros no satisfacen los estándares éticos de imparcialidad respecto de los intereses económicos involucrados.
S	gerundio	CIRCONSTANCE	cantando ,
S	gerundio	CIRCONSTANCE	aullando ,
S	gerundio	CIRCONSTANCE	chiflando,
S	cómo	CIRCONSTANCE	cómo evitar el contagio
?	si	?	si habrá algún complot detrás
?	?	?	que si las autoridades reaccionaron tarde
S	cómo	CIRCONSTANCE	cómo lo viven
N	y	LISTE/JOINT*	y lo hacen sonar quienes lo juegan
S	gerundio	CIRCONSTANCE	contando con la más calificada experiencia profesional , en el campo de la salud pública .
S	gerundio	CIRCONSTANCE	corriendo la versión de que era Roche que encabeza Severin Schwan ,
S	gerundio	ELABORATION	denunciando los casos en los que esto no ocurra .
S	es decir	REFORMULATION	es decir , modificó la estructura de su RNA , y de las glicoproteínas de su superficie .
S	gerundio	ELABORATION	incluyendo sobre todo en el DF , el servicio en restaurantes
S	gerundio	ELABORATION	las voces de los jugadores avisándose
S	gerundio	ELABORATION	reclamándose,
S	gerundio	ELABORATION	instruyéndose : del entrenador ordenando , del golpeado quejándose , del balón sonando como una cachetada en un pase , como un eco temible en un tapón , como un tambor zumbante en un tiro.

S	lo cual	RÉSULTAT	lo cual , como vemos , nos hace completamente dependientes del exterior aun ante problemas en los que está en juego la vida de muchos mexicanos .
S	lo que	RÉSULTAT	lo que implica procurar no asistir a lugares de alta concentración , lavarse las manos
S	?	ELABORATION	(yo agregaría : lavarse la cara , tomar una aspirina y usar gotas antisépticas para los ojos) .
N	ni	JOINT	ni deben ser objeto de discriminación .
?	?	?	o del Africam Safari , y si a usted , querido lector
S	para	BUT	para tomar una decisión en la que esté de acuerdo la mayoría .
S	pero	ANTITHÈSE	pero en todos los casos debe ser prescrito por un médico .
S	pero	ANTITHÈSE	pero que esto fue confirmado por los laboratorios más avanzados de Canadá y Estados Unidos .
S	porque	CAUSE	porque la industria automotriz tiene un enorme peso en el sector manufacturero.
S	pues	CAUSE	pues aquí tenían gran demanda las sedas , aromas , perlas , lacas , especies y demás lujos orientales .
S	pues	CAUSE	pues conozco la capacidad de los científicos mexicanos .
S	pues	CAUSE	pues de lo contrario el problema será doble .
S	pues	CAUSE	pues el número de casos puede rebasar su capacidad de atención adecuada y oportuna .
S	pues	CAUSE	pues permitiría , en el mediano plazo , sacudirnos la agobiante dependencia del exterior en materia de salud .
S	pues	CAUSE	pues son por el bien de todos
S	sin	CIRCONSTANCE	sin dar mayor explicación ,
S	lo que	RÉSULTAT	lo que generó de inmediato la inconformidad de los vecinos , refirió .
S	gerundio	CIRCONSTANCE	transmitiéndose inicialmente entre personas en nuestro país ,
N	y	SÉQUENCE	y de ahí se diseminó al mundo .
N	y	LISTE*/JOINT	y en los pobres no se piensa en el sector privado
S	a menos que	CONDITION INVERSE	(a menos que se cometan errores garrafales , que no es el caso actual) ,
S	lo cual	ÉVALUATION	(lo cual me parece correcto , en comparación con la primera versión generadora de pánico) , no estamos todavía en el terreno de las buenas noticias .
N	?	CONCESSION	no estamos todavía en el terreno de las buenas noticias .
S	?	CONDITION	(no fuera una arma de destrucción masiva porque les tomaría como 10 minutos) ,
S	aún sin	CONCESSION	aún sin incluir los efectos económicos de la influenza .
S	como	CIRCONSTANCE	Como es natural ,
N	?	ANTITHÈSE	Como parte de estos trabajos , se incluyeron además corredores entre las plantas hechos con tepujal , un material que ayuda a conservar la humedad en la tierra ,
S	pero	ANTITHÈSE	pero también hubo oposición ,
S	porque	CAUSE	porque se privilegiaba el paso peatonal sobre las áreas verdes ,
N	y	LISTE*/JOINT	y en su lugar , se colocó pasto .
S	como	CIRCONSTANCE	como ya lo he expresado ,
S	lo que	RÉSULTAT	lo que es un logro por tratarse de un espacio independiente diseñado para ofrecer escenificaciones de calidad .
S	para	BUT	para ofrecer escenificaciones de calidad .

S	lo que	RÉSULTAT	lo que los ha convertido en símbolo de la comida popular mexicana de excelencia
S	lo que	RÉSULTAT	lo que se traduce en deficiente información proporcionada a las mujeres tanto sobre el valor de la prueba
?	?	?	como de los pasos a seguir para la detección segura y oportuna de las lesiones pre-cancerosas .
S	luego de que	CIRCONSTANCE	luego de que se intentara modificar el diseño de las jardineras y el tipo de plantas en el lugar .
S	en	CIRCONSTANCE	no sólo en el parque México , sino también en los camellones de Ámsterdam , Tamaulipas , Nuevo León y Michoacán ,
N	y	LISTE/JOINT*	y de que los vecinos también consideren que están en un lugar privilegiado por sus áreas verdes , que requieren de mantenimiento .
S	para	BUT	para qué diga que es lo que quiere ,
?	?	?	pero del lado de la ignorancia ,
S	pero	ANTITHÈSE	pero se necesitan canales
S	para	BUT	para aprovechar sus talentos .
S	por lo que	RÉSULTAT	por lo que expresó que la playa inaugurada hace semana y media por el jefe de Gobierno
S	porque	CAUSE	porque aunque se dio gusto a los inconformes ,
S	probablemente	ELABORATION	probablemente son científicamente correctas
S	pero	ANTITHÈSE	pero eso se debió de haber dicho desde el principio ,
S	ya que	CAUSE	ya que ahora se convirtieron casi en símbolo de la resistencia ciudadana a la influenza .
S	porque	CAUSE	porque en este caso se trata de radicalismo puro , que tiene eco en otras personas , que reclaman que no tuvieron conocimiento del plan para la regeneración vegetal .
?	pues	CAUSE	pues cada señora tiene su secretillo para preparar los suyos
S	pues	CAUSE	pues detecta los casos reales y descarta los falsos positivos .
S	y		y descarta los falsos positivos .
S	pues	CAUSE	pues ha sido una demanda que por años han expuesto los vecinos ,
S	pues	CAUSE	pues varía conforme la condición socio - económica y lugar de residencia (Cancer Epidemiol Biomarkers Prev , 2008 ; 17 : 2808-2817) .
N	y	LISTE*/JOINT	y los datos parecen oscurecerse , mientras los reportes sobre personas infectadas en Estados Unidos , Canadá
N	y	LISTE/JOINT*	y algunas naciones europeas muestran que la infección por el virus de la influenza porcina comienza a tomar proporciones pandémicas .
S	para	BUT	y sobre todo para incitar a la participación ciudadana .
S	ya que	CAUSE	ya que aquí se establecieron las primeras universidad , imprenta , museo , casa de moneda y academia de artes ,
S	ya que	CAUSE	ya que sin él es dudoso que la prueba tenga el impacto deseable .

TABLE *A.2* – *Parties éliminées pour le genre journalistique.*

Type	Marqueur	Relation	Fragment
S	lo que	RÉSULTAT	lo que en economías cada vez más terciarizadas supone una fuerte restricción al incremento de la productividad agregada.
S	pues	CAUSE	pues los derechos fundamentales requieren de una serie de pautas hermenéuticas distintas a las que se pueden aplicar al resto de las normas jurídicas

S	ya que	CAUSE	ya que no atenta
S	ni	JOINT	ni vulnera el sistema constitucional ni en general el orden jurídico
S	ya que	CAUSE	ya que se reducirían las interacciones entre fármacos , sus efectos adversos
S	y	LISTE/JOINT*	y favorecería el cumplimiento de unos tratamientos que cada vez incluyen más pastillas.
S	cuando	CIRCONSTANCE/ CONDITION *	cuando su uso está relacionado con contenidos inapropiados para su adecuado desarrollo.
S	gerundio	RÉSULTAT	limitándose a reducir el factor de comportamiento sísmico que controla las resistencias de diseño.
S	para que	BUT	Para que esta aplicación sea posible,
S	si	CONDITION	si no se instauran rápido pautas terapéuticas eficaces.

TABLE *A.3* – *Parties éliminées pour le genre scientifique.*

Type	Marqueur	Relation	Fragment
S	como	MÉTHODE	como quien no quiere la cosa ,
S	como	CIRCONSTANCE	como suele suceder ,
N	y	LISTE/JOINT*	y empezaban a hacer ensayos con toda clase de alas , inclusive las de cera , desprestigiadas hacía poco en una aventura infortunada .
S	como	ELABORATION	más seguro de sí mismo , como quien había viajado tanto ,
N	y	LISTE/JOINT*	y todos los días se esforzaba en ello .
N	y	LISTE/JOINT*	y comenzó a peinarse
N	y	LISTE/JOINT*	y a vestirse
N	y	LISTE/JOINT*	y a desvestirse
S	de manera que	RÉSULTAT	de manera que se dedicó a hacer sentadillas
N	y	LISTE/JOINT*	y a saltar
S	para	BUT	para tener unas ancas cada vez mejores ,
N	y	LISTE/JOINT*	y sentía que todos la aplaudían.
S	porque	CAUSE	porque les mintieron que es vieja.
S	de modo que	CIRCONSTANCE	se encamina a una casa de óptica y
S	a fin de	BUT	a fin de curarse en salud .

TABLE *A.4* – *Parties éliminées pour le genre littéraire.*

Annexe B

Exemple de résumés obtenus avec différents taux de compression

Liste de résumés obtenus pour le document « *Descubrimiento de mamut emociona a científicos* » en utilisant l'algorithme 2 décrit dans la section 5.6. Le tableau B.2 correspond au document d'origine et le tableau B.1 à sa traduction approximé.

Découverte d'un mammouth enflamme les scientifiques
L'inusuelle trouvaille des ossements d'un mammouth dans une ferme de Oskaloosa a enflammé les experts qui étudient la vie préhistorique à cause des découvertes scientifiques qui pourraient se faire avec l'énorme bête. La trouvaille est rare parce qu'il paraît qu'une bonne partie du squelette de l'animal se trouve en excellent état. Ceci permet aux scientifiques de collecter du pollen et des évidences d'autres plantes dans le site de l'excavation qui pourraient révéler des détails sur l'environnement de Iowa il y a plus de 12000 ans. Les scientifiques de l'Université d'Iowa prévoient de scanner le site vendredi avec un radar qui pénétra la terre pour voir s'ils peuvent déterminer quelle quantité des restes du mammouth sont encore ensevelis. L'excavation s'étendra durant plusieurs mois. Les ossements ont été trouvés il y a deux ans par le propriétaire du terrain, qui souhaite que son nom ainsi que l'adresse de la ferme soient confidentielles pour protéger le site.

TABLE *B.1* – *Document « Descubrimiento de mamut emociona a científicos » traduit.*

Descubrimiento de mamut emociona a científicos
El inusual descubrimiento de los huesos de un mamut en una finca de Oskaloosa ha emocionado a los expertos que estudian la vida prehistórica por los descubrimientos científicos que podrían hacer se con la enorme bestia. El hallazgo es raro porque al parecer buena parte del esqueleto del animal se encuentra en excelente estado. Eso permite a los científicos recolectar polen y evidencias de otras plantas en el sitio de la excavación que podrían revelar detalles respecto del medio ambiente de Iowa hace más de 12000 años. Los científicos de la Universidad de Iowa planean escanear el lugar el viernes con un radar que penetra en la tierra para ver si pueden determinar qué tanto de los restos del mamut siguen enterrados. La excavación se extenderá por varios meses. Los huesos los encontró hace dos años el propietario del terreno, quien desea que tanto su nombre como la ubicación de la granja sean confidenciales para proteger el sitio.

TABLE *B.2* – *Document « Descubrimiento de mamut emociona a científicos ».*

Descubrimiento de mamut emociona a científicos (résumé)
Los científicos de la Universidad de Iowa planean escanear el lugar el viernes con un radar que penetra en la tierra. La excavación se extenderá por varios meses. Los huesos los encontró hace dos años el propietario del terreno, quien desea que tanto su nombre como la ubicación de la granja sean confidenciales.

TABLE *B.3* – *Résumé obtenu en utilisant l'algorithme 2. Conditions initiales : segmenteur=DiSeg, τ souhaité* $= 10\%$*. Conditions finales : τ obtenu* $= 33.3\%$*, $\alpha = 0.01$.*

Descubrimiento de mamut emociona a científicos (résumé)
Los científicos de la Universidad de Iowa planean escanear el lugar el viernes con un radar que penetra en la tierra. La excavación se extenderá por varios meses. Los huesos los encontró hace dos años el propietario del terreno, quien desea que tanto su nombre como la ubicación de la granja sean confidenciales.

TABLE *B.4* – *Résumé obtenu en utilisant l'algorithme 2. Conditions initiales : segmenteur=DiSeg, τ souhaité* $= 20\%$*. Conditions finales : τ obtenu* $= 33.3\%$*, $\alpha = 0.02$.*

Descubrimiento de mamut emociona a científicos (résumé)
Los científicos de la Universidad de Iowa planean escanear el lugar el viernes con un radar que penetra en la tierra. La excavación se extenderá por varios meses. Los huesos los encontró hace dos años el propietario del terreno, quien desea que tanto su nombre como la ubicación de la granja sean confidenciales.

TABLE *B.5* – *Résumé obtenu en utilisant l'algorithme 2. Conditions initiales : segmenteur=DiSeg, τ souhaité $= 30\%$. Conditions finales : τ obtenu $= 33.3\%$, $\alpha = 0.09$.*

Descubrimiento de mamut emociona a científicos (résumé)
El inusual descubrimiento de los huesos de un mamut en una finca de Oskaloosa ha emocionado a los expertos que estudian la vida prehistórica por los descubrimientos científicos que podrían hacerse con la enorme bestia. Los científicos de la Universidad de Iowa planean escanear el lugar el viernes con un radar que penetra en la tierra. La excavación se extenderá por varios meses. Los huesos los encontró hace dos años el propietario del terreno, quien desea que tanto su nombre como la ubicación de la granja sean confidenciales.

TABLE *B.6* – *Résumé obtenu en utilisant l'algorithme 2. Conditions initiales : segmenteur=DiSeg, τ souhaité $= 40\%$. Conditions finales : τ obtenu $= 55.2\%$, $\alpha = 0.10$.*

Descubrimiento de mamut emociona a científicos (résumé)
El inusual descubrimiento de los huesos de un mamut en una finca de Oskaloosa ha emocionado a los expertos que estudian la vida prehistórica por los descubrimientos científicos que podrían hacerse con la enorme bestia. Los científicos de la Universidad de Iowa planean escanear el lugar el viernes con un radar que penetra en la tierra. La excavación se extenderá por varios meses. Los huesos los encontró hace dos años el propietario del terreno, quien desea que tanto su nombre como la ubicación de la granja sean confidenciales.

TABLE *B.7* – *Résumé obtenu en utilisant l'algorithme 2. Conditions initiales : segmenteur=DiSeg, τ souhaité $= 50\%$. Conditions finales : τ obtenu $= 55.2\%$, $\alpha = 0.15$.*

Descubrimiento de mamut emociona a científicos (résumé)
El inusual descubrimiento de los huesos de un mamut en una finca de Oskaloosa ha emocionado a los expertos que estudian la vida prehistórica por los descubrimientos científicos que podrían hacerse con la enorme bestia porque al parecer buena parte del esqueleto del animal se encuentra en excelente estado. Los científicos de la Universidad de Iowa planean escanear el lugar el viernes con un radar que penetra en la tierra. La excavación se extenderá por varios meses. Los huesos los encontró hace dos años el propietario del terreno, quien desea que tanto su nombre como la ubicación de la granja sean confidenciales.

TABLE *B.8* – *Résumé obtenu en utilisant l'algorithme 2. Conditions initiales : segmenteur=DiSeg, τ souhaité $= 60\%$. Conditions finales : τ obtenu $= 64.2\%$, $\alpha = 0.19$.*

Descubrimiento de mamut emociona a científicos (résumé)
El inusual descubrimiento de los huesos de un mamut en una finca de Oskaloosa ha emocionado a los expertos que estudian la vida prehistórica por los descubrimientos científicos que podrían hacerse con la enorme bestia porque al parecer buena parte del esqueleto del animal se encuentra en excelente estado. Eso permite a los científicos recolectar polen y evidencias de otras plantas en el sitio de la excavación que podrían revelar detalles respecto de el medio ambiente de Iowa hace más de 12000 años. Los científicos de la Universidad de Iowa planean escanear el lugar el viernes con un radar que penetra en la tierra. La excavación se extenderá por varios meses. Los huesos los encontró hace dos años el propietario del terreno, quien desea que tanto su nombre como la ubicación de la granja sean confidenciales.

TABLE *B.9* – *Résumé obtenu en utilisant l'algorithme 2. Conditions initiales : segmenteur=DiSeg, τ souhaité $= 70\%$. Conditions finales : τ obtenu $= 85.5\%$, $\alpha = 0.24$.*

Descubrimiento de mamut emociona a científicos (résumé)
El inusual descubrimiento de los huesos de un mamut en una finca de Oskaloosa ha emocionado a los expertos que estudian la vida prehistórica por los descubrimientos científicos que podrían hacerse con la enorme bestia porque al parecer buena parte del esqueleto del animal se encuentra en excelente estado. Eso permite a los científicos recolectar polen y evidencias de otras plantas en el sitio de la excavación que podrían revelar detalles respecto de el medio ambiente de Iowa hace más de 12000 años. Los científicos de la Universidad de Iowa planean escanear el lugar el viernes con un radar que penetra en la tierra. La excavación se extenderá por varios meses. Los huesos los encontró hace dos años el propietario del terreno, quien desea que tanto su nombre como la ubicación de la granja sean confidenciales.

TABLE *B.10* – *Résumé obtenu en utilisant l'algorithme 2. Conditions initiales : segmenteur=DiSeg, τ souhaité $= 80\%$. Conditions finales : τ obtenu $= 85.5\%$, $\alpha = 0.26$.*

Descubrimiento de mamut emociona a científicos (résumé)
El inusual descubrimiento de los huesos de un mamut en una finca de Oskaloosa ha emocionado a los expertos que estudian la vida prehistórica por los descubrimientos científicos que podrían hacerse con la enorme bestia. El hallazgo es raro porque al parecer buena parte del esqueleto del animal se encuentra en excelente estado. Eso permite a los científicos recolectar polen y evidencias de otras plantas en el sitio de la excavación que podrían revelar detalles respecto de el medio ambiente de Iowa hace más de 12000 años. Los científicos de la Universidad de Iowa planean escanear el lugar el viernes con un radar que penetra en la tierra. Si pueden determinar qué tanto de los restos del mamut siguen enterrados la excavación se extenderá por varios meses. Los huesos los encontró hace dos años el propietario del terreno, quien desea que tanto su nombre como la ubicación de la granja sean confidenciales.

TABLE *B.11* – *Résumé obtenu en utilisant l'algorithme 2. Conditions initiales : segmenteur=DiSeg, τ souhaité $= 90\%$. Conditions finales : τ obtenu $= 95.8\%$, $\alpha = 0.45$.*

Descubrimiento de mamut emociona a científicos (résumé)
La excavación se extenderá por varios meses. Los huesos los encontró hace dos años el propietario del terreno.

TABLE *B.12* – *Résumé obtenu en utilisant l'algorithme 2. Conditions initiales : segmenteur=CoSeg, τ souhaité $= 10\%$. Conditions finales : τ obtenu $= 12.1\%$, $\alpha = 0.03$.*

Descubrimiento de mamut emociona a científicos (résumé)
Los científicos de la Universidad de Iowa planean escanear el lugar el viernes con un radar que penetra en la tierra. La excavación se extenderá por varios meses. Los huesos los encontró hace dos años el propietario del terreno.

TABLE *B.13* – *Résumé obtenu en utilisant l'algorithme 2. Conditions initiales : segmenteur=CoSeg, τ souhaité* $= 20\%$. *Conditions finales : τ obtenu* $= 24.8\%$, $\alpha = 0.05$.

Descubrimiento de mamut emociona a científicos (résumé)
Porque al parecer buena parte del esqueleto del animal se encuentra en excelente estado. Los científicos de la Universidad de Iowa planean escanear el lugar el viernes con un radar que penetra en la tierra. La excavación se extenderá por varios meses. Los huesos los encontró hace dos años el propietario del terreno.

TABLE *B.14* – *Résumé obtenu en utilisant l'algorithme 2. Conditions initiales : segmenteur=CoSeg, τ souhaité* $= 30\%$. *Conditions finales : τ obtenu* $= 33.9\%$, $\alpha = 0.10$.

Descubrimiento de mamut emociona a científicos (résumé)
El inusual descubrimiento de los huesos de un mamut en una finca de Oskaloosa ha emocionado a los expertos que estudian la vida prehistórica por los descubrimientos científicos que podrían hacerse con la enorme bestia porque al parecer buena parte del esqueleto del animal se encuentra en excelente estado. Los científicos de la Universidad de Iowa planean escanear el lugar el viernes con un radar que penetra en la tierra. La excavación se extenderá por varios meses. Los huesos los encontró hace dos años el propietario del terreno.

TABLE *B.15* – *Résumé obtenu en utilisant l'algorithme 2. Conditions initiales : segmenteur=CoSeg, τ souhaité* $= 40\%$. *Conditions finales : τ obtenu* $= 56.4\%$, $\alpha = 0.20$.

Descubrimiento de mamut emociona a científicos (résumé)
El inusual descubrimiento de los huesos de un mamut en una finca de Oskaloosa ha emocionado a los expertos que estudian la vida prehistórica por los descubrimientos científicos que podrían hacerse con la enorme bestia porque al parecer buena parte del esqueleto del animal se encuentra en excelente estado. Los científicos de la Universidad de Iowa planean escanear el lugar el viernes con un radar que penetra en la tierra. La excavación se extenderá por varios meses. Los huesos los encontró hace dos años el propietario del terreno.

TABLE *B.16* – *Résumé obtenu en utilisant l'algorithme 2. Conditions initiales : segmenteur=CoSeg, τ souhaité* $= 50\%$. *Conditions finales : τ obtenu* $= 56.4\%$, $\alpha = 0.22$.

Descubrimiento de mamut emociona a científicos (résumé)
El inusual descubrimiento de los huesos de un mamut en una finca de Oskaloosa ha emocionado a los expertos que estudian la vida prehistórica por los descubrimientos científicos que podrían hacerse con la enorme bestia porque al parecer buena parte del esqueleto del animal se encuentra en excelente estado. y evidencias de otras plantas en el sitio de la excavación que podrían revelar detalles respecto del medio ambiente de Iowa hace más de 12000 años. Los científicos de la Universidad de Iowa planean escanear el lugar el viernes con un radar que penetra en la tierra. La excavación se extenderá por varios meses. Los huesos los encontró hace dos años el propietario del terreno.

TABLE *B.17* – *Résumé obtenu en utilisant l'algorithme 2. Conditions initiales : segmenteur=CoSeg, τ souhaité $= 60\%$. Conditions finales : τ obtenu $= 72.7\%$, $\alpha = 0.23$.*

Descubrimiento de mamut emociona a científicos (résumé)
El inusual descubrimiento de los huesos de un mamut en una finca de Oskaloosa ha emocionado a los expertos que estudian la vida prehistórica por los descubrimientos científicos que podrían hacerse con la enorme bestia porque al parecer buena parte del esqueleto del animal se encuentra en excelente estado. Eso permite a los científicos recolectar polen y evidencias de otras plantas en el sitio de la excavación que podrían revelar detalles respecto del medio ambiente de Iowa hace más de 12000 años. Los científicos de la Universidad de Iowa planean escanear el lugar el viernes con un radar que penetra en la tierra. La excavación se extenderá por varios meses. Los huesos los encontró hace dos años el propietario del terreno.

TABLE *B.18* – *Résumé obtenu en utilisant l'algorithme 2. Conditions initiales : segmenteur=CoSeg, τ souhaité $= 70\%$. Conditions finales : τ obtenu $= 77.0\%$, $\alpha = 0.24$.*

Descubrimiento de mamut emociona a científicos (résumé)
El inusual descubrimiento de los huesos de un mamut en una finca de Oskaloosa ha emocionado a los expertos que estudian la vida prehistórica por los descubrimientos científicos que podrían hacerse con la enorme bestia. El hallazgo es raro porque al parecer buena parte del esqueleto del animal se encuentra en excelente estado. Eso permite a los científicos recolectar polen y evidencias de otras plantas en el sitio de la excavación que podrían revelar detalles respecto del medio ambiente de Iowa hace más de 12000 años. Los científicos de la Universidad de Iowa planean escanear el lugar el viernes con un radar que penetra en la tierra. Si pueden determinar qué tanto de los restos del mamut siguen enterrados la excavación se extenderá por varios meses. Los huesos los encontró hace dos años el propietario del terreno.

TABLE *B.19* – *Résumé obtenu en utilisant l'algorithme 2. Conditions initiales : segmenteur=CoSeg, τ souhaité $= 80\%$. Conditions finales : τ obtenu $= 87.3\%$, $\alpha = 0.30$.*

Descubrimiento de mamut emociona a científicos (résumé)
El inusual descubrimiento de los huesos de un mamut en una finca de Oskaloosa ha emocionado a los expertos que estudian la vida prehistórica por los descubrimientos científicos que podrían hacerse con la enorme bestia. El hallazgo es raro porque al parecer buena parte del esqueleto del animal se encuentra en excelente estado. Eso permite a los científicos recolectar polen y evidencias de otras plantas en el sitio de la excavación que podrían revelar detalles respecto del medio ambiente de Iowa hace más de 12000 años. Los científicos de la Universidad de Iowa planean escanear el lugar el viernes con un radar que penetra en la tierra. Si pueden determinar qué tanto de los restos del mamut siguen enterrados la excavación se extenderá por varios meses. Los huesos los encontró hace dos años el propietario del terreno, quien desea que tanto su nombre como la ubicación de la granja sean confidenciales.

TABLE *B.20* – *Résumé obtenu en utilisant l'algorithme 2. Conditions initiales : segmenteur=CoSeg,* τ *souhaité* $= 90\%$*. Conditions finales :* τ *obtenu* $= 95.8\%$*,* $\alpha = 0.45$*.*

Annexe C

Test d'évaluation

Document distribué aux 54 juges pour l'évaluation des résumés avec de phrases compressées. L'information de la légende de chaque tableau leur été inconnue.

Consigne

Parmi les 12 résumés ci-dessous quelques-uns ont été créés automatiquement par un programme tandis que d'autres ont été élaborés par des humains. Êtes-vous capable de les distinguer ?

La persona con el cociente intelectual más alto del mundo
Su nombre es Marilyn vos Savant y nació en San Louis (Missouri) el 11 de agosto de 1946. Marilyn vos Savant está considerada como la persona con el cociente intelectual más alto del mundo. Hoy en día es una más que reputada columnista, escritora, conferenciante y dramaturga. En 1986 comenzó una columna dominical llamada Pregunta a Marilyn (Ask Marilyn) en la revista Parade, donde responde preguntas de los lectores acerca de diversos temas. Su mayor aspiración era el convertirse en escritora. Durante su juventud trabajó en la tienda de ultramarinos de su padre. Cursó varios seminarios de filosofía en la universidad. En la actualidad está casada con el prestigioso cardiólogo Robert Jarvik. A Marilyn se le asocia con el famoso problema de Monty Hall, o bien le fue planteado a ella a través de una consulta en su columna Ask Marilyn.

TABLE C.1 – *Résumé produit à partir du document « La persona con el cociente intelectual más alto del mundo » (segmentation : DiSeg, résumeur : Humain, τ =51.83824% du document d'origine).*

La Tundra
El ambiente de la tundra está caracterizado por una sequía prolongada. Las especies más típicas de la flora son los arbustos enanos, líquenes y musgos. Algunas especies, particularmente de aves, sólo pasan el verano en la tundra, época en la que anidan. Existen pocas especies de anfibios y reptiles.

TABLE *C.2* – *Résumé produit à partir du document « La Tundra » (segmentation : DiSeg, résumeur : Machine, τ =16.17162% du document d'origine).*

El Pulque
El Pulque o Neutle se obtiene de la fermentación de la savia azucarada o aguamiel, concentrados en el corazón de el maguey, antes de que salga el pedúnculo de la inflorescencia del maguey por el proceso conocido como raspado, que consiste en quitar el centro de la planta donde crecen las hojas tiernas dejando una oquedad que se tapa con una penca del maguey. El interior es entonces raspado con una especie de cuchara, lo que provoca que el maguey suelte un jugo el cual se concentra en el hueco. Este es, luego, a intervalos de uno o dos días absorbido hacia un cuenco hueco (llamado acocote, fruto de una cucurbitácea) y depositado en un recipiente llamado odre. Este proceso lo lleva a cabo el Tlachiquero o raspador, y el jugo se recolecta durante dos meses como máximo. Después es depositado en barriles de pino o, en cubas de acero inoxidable, donde se fermenta con la bacteria Zymomonas mobilis durante uno o dos días obteniendose un líquido blanco de aspecto lechoso con un 5% de alcohol. Se debe beber inmediatamente ya que al seguirse fermentando adquiere un gusto muy fuerte.

TABLE *C.3* – *Résumé produit à partir du document « El Pulque » (segmentation : CoSeg, résumeur : Humain, τ =90.90909% du document d'origine).*

La música en el antiguo Egipto
La Música en el antiguo Egipto se empleaba en varias actividades, pero su desarrollo principal fue en los templos, donde era usada durante los ritos dedicados a los diferentes dioses y era utilizada como remedio terapéutico. Como en otros pueblos, también se consideraba un medio de comunicación con los difuntos y los músicos alcanzaban una categoría tal que algunos están enterrados en las necrópolis reales. No se conoce cómo era realmente ya que no desarrollaron un sistema para representarla, se transmitía de maestro a alumno. También arrojan luz sobre este tema los instrumentos conservados en los museos y la representación en bajorrelieves y pinturas de instrumentos y bailarines, además de lo conservado por tradición oral por los cantores coptos.

TABLE *C.4* – *Résumé produit à partir du document « La música en el antiguo Egipto » (segmentation : DiSeg, résumeur : Machine, τ =76.28205% du document d'origine).*

Efectos de la LSD
Los efectos de la LSD sobre el sistema nervioso central son extremadamente variables y dependen de la cantidad que se consuma, el entorno en que se use la droga, la pureza de ésta, la personalidad, el estado de ánimo y las expectativas del usuario. Algunos consumidores de LSD experimentan una sensación de euforia, mientras que otros viven la experiencia en clave terrorífica. Cuando la experiencia tienen un tono general desagradable, suele hablarse de mal viaje. Cuando la sustancia se administra por vía oral, los efectos tardan en manifestar se entre 30 minutos y una hora y, según la dosis, pueden durar entre 8 y 10 horas. Entre los efectos fisiológicos recurrentes están los siguientes : contracciones uterinas, fiebre, erizamiento del vello, aumento de la frecuencia cardíaca, transpiración, pupilas dilatadas, insomnio, hiperreflexia y temblores.

TABLE *C.5* – *Résumé produit à partir du document « Efectos de la LSD » (segmentation : CoSeg, résumeur : Humain, τ =90.30769% du document d'origine).*

Introducción a las matemáticas
Cada vez que vas a la tienda, juegas en la computadora o en la consola de video juegos ; cuando sigues las incidencias de un juego de béisbol o fútbol americano, cuando llevas el ritmo de una canción, estás utilizando relaciones numéricas y en tu mente realizas una serie de operaciones que tienen que ver con el lenguaje matemático. En este sentido, podemos afirmar que el pensamiento matemático está presente en la mayoría de nuestras actividades, desde las más sencillas hasta las más especializadas. Sin embargo, no siempre estamos conscientes de los conceptos, reglas, modelos, procedimientos y operaciones matemáticas que realizamos mentalmente a diario. A lo largo de esta unidad, mediante la adquisición de distintos conocimientos y la resolución de una serie de problemas y ejercicios, descubriremos cómo representar y formalizar algunas de las operaciones que mencionamos. Los cursos de matemáticas que llevaste con anterioridad, te han familiarizado con la utilización de ciertas operaciones básicas. Con ello podríamos decir que posees los conocimientos básicos para manejar algoritmos elementales. Así que reconocerás diferentes tipos de números como los naturales, los enteros, los fraccionarios (racionales) y los irracionales que son temas de esta unidad. Gracias al conocimiento de los distintos tipos de números construirás y aplicarás modelos matemáticos, los cuales trabajarás con razones y proporciones, así como con series y sucesiones, que te ayudarán a resolver diferentes situaciones de la vida cotidiana. Todos estos aprendizajes te servirán en las siguientes unidades para identificar, resolver, plantear, interpretar y aplicar diferentes procedimientos (algoritmos) con un distinto nivel de complejidad, en variedad de situaciones.

TABLE *C.6* – *Résumé produit à partir du document « Introducción a las matemáticas » (segmentation : DiSeg, résumeur : Humain, τ =84.31373% du document d'origine).*

Por qué el embarazo de las elefantas es tan largo
El período de gestación, que se prolonga por casi dos años, es una de esas rarezas de la biología que le permite al feto desarrollar suficientemente su cerebro. Los resultados de este estudio servirán para mejorar los programas de reproducción de elefantes en los zoológicos y podrían también contribuir al desarrollo de un anticonceptivo. Los elefantes son mamíferos muy sociales con un alto grado de inteligencia, similar a la de los homínidos y los delfines. Son, además, los que tienen el período de gestación más largo, que puede extenderse hasta por 680 días. Los elefantes nacen con un nivel avanzado de desarrollo cerebral, que utilizan para alimentarse mediante sus habilidosas trompas. Hasta ahora, los científicos no habían logrado entender en profundidad los procesos biológicos del maratónico embarazo de las elefantas. Pero gracias a los avances de las técnicas de ultrasonido, los veterinarios pudieron utilizar nuevas herramientas.

TABLE C.7 – *Résumé produit à partir du document « Por qué el embarazo de las elefantas es tan largo » (segmentation : DiSeg, résumeur : Humain, τ =69.85646% du document d'origine).*

Ética de robots
Existe la preocupación de que los robots puedan desplazar o competir con los humanos. Las leyes o reglas que pudieran o debieran ser aplicadas a los robots u otros entes autónomos en cooperación o competencia con humanos han estimulado las investigaciones macroeconómicas de este tipo de competencia, notablemente por Alessandro Acquisti basandose en un trabajo anterior de John von Neumann. Actualmente, no es posible aplicar las Tres leyes de la robótica, dado que los robots no tienen capacidad para comprender su significado. Entender y aplicar las Tres leyes de la robótica, requeriría verdadera inteligencia y consciencia del medio circundante, así como de sí mismo, por parte del robot.

TABLE C.8 – *Résumé produit à partir du document « Ética de robots » (segmentation : CoSeg, résumeur : Humain, τ =63.52941% du document d'origine).*

Confirman en Veracruz caso de influenza en niño de 5 años
El gobierno de Veracruz confirmó este domingo un caso de influenza porcina de la cepa H1N1 en un niño de cinco años originario de el poblado La Gloria. El subdirector de prevención y control de enfermedades de la Secretaría de Salud estatal dijo que el menor de nombre Edgar Hernández Hernández superó el cuadro de infección pulmonar.

TABLE C.9 – *Résumé produit à partir du document « Confirman en Veracruz caso de influenza en niño de 5 años » (segmentation : CoSeg, résumeur : Machine, τ =26.26728% du document d'origine).*

Hallan genes asociados a migraña
Investigadores europeos y australianos indicaron el domingo que habían localizado cuatro nuevos genes asociados con la forma más común de la migraña. Las variantes genéticas fueron detectadas en el genoma de 4800 pacientes de migraña sin aura, la forma que asumen tres de cada cuatro ataques de migraña. Estas estas variantes genéticas no fueron halladas, sin embargo, en el grupo testigo de 7000 personas libres de la enfermedad, dijeron los investigadores. El estudio también confirmó la existencia de otros dos genes de predisposición, en un trío de genes ya identificados en un trabajo anterior. La migraña afecta a aproximadamente una de cada seis mujeres y a uno de cada ocho hombres. Los nuevos genes identificados en este estudio refuerzan el argumento según el cual la disfunción de las moléculas responsables de la transmisión de señales entre las células nerviosas, contribuye a la aparición de la migraña. Además, dos de estos genes refuerzan la hipótesis de un posible papel de las venas. La investigación, publicada en la revista especializada Nature Genetics, fue realizada por un consorcio internacional dedicado a la investigación sobre la genética de la migraña.

TABLE *C.10* – *Résumé produit à partir du document « Hallan genes asociados a migraña » (segmentation : CoSeg, résumeur : Machine, τ =79.48718% du document d'origine).*

Problemas globales
Hoy se reconoce que existen problemas que denominamos globales. Estos problemas se presentan fundamentalmente por la carga de contaminantes liberados hacia la atmósfera terrestre. Por su magnitud y complejidad constituyen un grave problema que requiere medidas muy drásticas para su solución. La composición química de la atmósfera es muy inestable : cambia a través del tiempo y en función de diversas reacciones e interacciones de sus componentes. Hoy sabemos que además de los numerosos gases que la componen, existe una compleja interrelación de los gases. Esta interacción se manifiesta en el hecho de que la radiación solar aporta la energía necesaria para que se realicen las reacciones químicas que modifican la composición de la atmósfera. El diálogo entre la atmósfera y la radiación solar ha sido alterado por el hombre.

TABLE *C.11* – *Résumé produit à partir du document « Problemas globales » (segmentation : DiSeg, résumeur : Machine, τ =59.447% du document d'origine).*

Descubrimiento de mamut emociona a científicos
El inusual descubrimiento de los huesos de un mamut en una finca de Oskaloosa ha emocionado a los expertos que estudian la vida prehistórica por los descubrimientos científicos que podrían hacerse con la enorme bestia. Al parecer buena parte del esqueleto del animal se encuentra en excelente estado. Los científicos de la Universidad de Iowa planean escanear el lugar el viernes con un radar que penetra en la tierra. La excavación se extenderá por varios meses.

TABLE *C.12* – *Résumé produit à partir du document « Descubrimiento de mamut emociona a científicos » (segmentation : CoSeg, résumeur : Machine, τ =43.42857% du document d'origine).*

Annexe D

Description du corpus et des données issues de l'annotation

Le corpus de compression de phrases en espagnol est un ensemble de textes enrichis manuellement avec des annotations concernant l'élimination de segments discursifs intra-phrase. Au total, le corpus contient 60 844 décisions d'éliminer ou de préserver un segment.

Un total de 30 textes ont été segmentés en unités discursives élémentaires par deux segmenteurs discursifs : CoSeg et DiSeg. Chaque texte segmenté correspond à une source différente dans le corpus. En conséquence il y a 60 sources au total, numérotées de 0 à 59. Les sources de 0 à 29 correspondent aux 30 textes segmentés par CoSeg et ceux segmentés par DiSeg vont du 30 à 59. Il y a 60 répertoires, un pour chaque source, nommés src0, src2,...,src59. Chaque répertoire contient environ 50 fichiers XML.

- Un fichier src*i*/orig_src*i*.xml contient le texte *i* d'origine.
- Un fichier src*i*/annot_src*i*.xml contient les résultats de l'annotation pour la source *i*.
- Plusieurs fichiers src*i*/res_src*i*_usr*j*.xml contiennent les résultats de l'annotation pour la source *i* obtenus grâce aux annotateurs (*j* est l'identifiant de l'annotateur).

Les trois balises utilisées dans ces fichiers sont : <source>, <sentence> et et les attributs sont :

- segmenter = Type de segmentation : 'P' pour CoSeg et 'D' pour DiSeg
- src_id = Identifiant de la source
- src_name = Identifiant du texte
- title = Titre du document
- sent_id =Identifiant de la phrase
- sent_ener = Énergie textuelle de la phrase
- sent_lgram = Valeur de la phrase dans un modèle de langage probabiliste
- sent_words = Nombre de mots de la phrase

- seg_id = Identifiant du segment
- deleted = Nombre de fois que le segment a été éliminé
- preserved = Nombre de fois que le segment a été préservé
- seg_ener = Énergie textuelle du segment
- seg_lgram = Valeur du segment dans un modèle de langage probabiliste
- seg_words = Nombre de mots du segment

Le corpus de compression de phrases en espagnol est disponible à l'adresse http://molina.talne.eu/sentence_compression/data/

```
<?xml version="1.0" encoding="utf-8"?>
<XML>
    <source segmenter="D" src_id="30" src_name="ecol01D" title="El termómetro">
      <sentence sent_words="20" sent_lgram="-59.8736" sent_ener="212" sent_id="210">
        <segment deleted="Y" seg_id="821">Para saber qué tan caliente o frío está algo, es decir,</segment>
        <segment deleted="N" seg_id="822">si se desea conocer la temperatura, debe utilizarse un instrumento que
      ofrezca un dato confiable, el termómetro.</segment>
      </sentence>
      <sentence sent_words="19" sent_lgram="-41.0067" sent_ener="91" sent_id="211">
        <segment deleted="N" seg_id="823">Este instrumento tiene muchos usos en los hogares, en las industrias y en
      las unidades de salud.</segment>
      </sentence>
      <sentence sent_words="7" sent_lgram="-23.0591" sent_ener="99" sent_id="212">
        <segment deleted="N" seg_id="824">En casa es útil tener un termómetro</segment>
        <segment deleted="Y" seg_id="825">para saber con precisión</segment>
        <segment deleted="Y" seg_id="826">si alguien de la familia tiene fiebre.</segment>
      </sentence>
      <sentence sent_words="18" sent_lgram="-47.0171" sent_ener="320" sent_id="213">
        <segment deleted="N" seg_id="827">En la industria los termómetros miden la temperatura de hornos y calderas,
       así como de diversos materiales</segment>
        <segment deleted="Y" seg_id="828">y sustancias que cambian a través de un proceso productivo.</segment>
      </sentence>
      <sentence sent_words="37" sent_lgram="-82.3006" sent_ener="449" sent_id="214">
        <segment deleted="Y" seg_id="829">Como ves,</segment>
        <segment deleted="N" seg_id="830">con frecuencia es necesario medir la temperatura de distintas cosas, del
      aire, del cuerpo humano, de un horno o del agua de una alberca,</segment>
        <segment deleted="N" seg_id="831">por lo que existen distintos tipos de termómetros.</segment>
      </sentence>
      <sentence sent_words="19" sent_lgram="-48.4060" sent_ener="383" sent_id="215">
        <segment deleted="N" seg_id="832">No importa el tipo de termómetro, en todos ellos la temperatura se mide en
       unidades llamadas grados.</segment>
      </sentence>
      <sentence sent_words="29" sent_lgram="-69.7622" sent_ener="318" sent_id="216">
        <segment deleted="N" seg_id="833">Cada marca del instrumento es un grado</segment>
        <segment deleted="N" seg_id="834">y cada tipo de termómetro incluye una escala de medición que, por lo
      general, se da en grados centígrados.</segment>
      </sentence>
    </source>
</XML>
```

***Listing D.1** – Exemple de fichier XML du corpus contenant les informations d'un seul annotateur.*

```
<?xml version="1.0" encoding="utf-8"?>
<XML>
  <source segmenter="D" src_id="30" src_name="ecol01D" title="El termómetro">
    <sentence sent_ener="243" sent_id="210" sent_lgram="-89.9214" sent_words="33">
      <segment deleted="39" preserved="10" seg_ener="4" seg_id="821" seg_lgram="-35.6487" seg_words="13">Para saber
      qué tan caliente o frío está algo, es decir,</segment>
      <segment deleted="0" preserved="49" seg_ener="171" seg_id="822" seg_lgram="-59.8736" seg_words="20">si se
      desea conocer la temperatura, debe utilizarse un instrumento que ofrezca un dato confiable, el termómetro.</
      segment>
    </sentence>
    <sentence sent_ener="93" sent_id="211" sent_lgram="-41.0067" sent_words="19">
      <segment deleted="0" preserved="49" seg_ener="76" seg_id="823" seg_lgram="-41.0067" seg_words="19">Este
      instrumento tiene muchos usos en los hogares, en las industrias y en las unidades de salud.</segment>
    </sentence>
    <sentence sent_ener="126" sent_id="212" sent_lgram="-51.8966" sent_words="19">
      <segment deleted="3" preserved="46" seg_ener="84" seg_id="824" seg_lgram="-23.0591" seg_words="7">En casa es
      útil tener un termómetro</segment>
      <segment deleted="31" preserved="18" seg_ener="4" seg_id="825" seg_lgram="-15.8522" seg_words="4">para saber
      con precisión</segment>
      <segment deleted="27" preserved="22" seg_ener="0" seg_id="826" seg_lgram="-22.6693" seg_words="8">si alguien
      de la familia tiene fiebre.</segment>
    </sentence>
    <sentence sent_ener="326" sent_id="213" sent_lgram="-65.1852" sent_words="29">
      <segment deleted="0" preserved="49" seg_ener="269" seg_id="827" seg_lgram="-47.0171" seg_words="18">En la
      industria los termómetros miden la temperatura de hornos y calderas, así como de diversos materiales</segment>
      <segment deleted="38" preserved="11" seg_ener="0" seg_id="828" seg_lgram="-23.1637" seg_words="11">y
      sustancias que cambian a través de un proceso productivo.</segment>
    </sentence>
```

```
    <sentence sent_ener="455" sent_id="214" sent_lgram="-89.2905" sent_words="40">
      <segment deleted="45" preserved="4" seg_ener="0" seg_id="829" seg_lgram="-12.1723" seg_words="3">Como ves,</segment>
      <segment deleted="7" preserved="42" seg_ener="183" seg_id="830" seg_lgram="-66.5261" seg_words="28">con frecuencia es necesario medir la temperatura de distintas cosas, del aire, del cuerpo humano, de un horno o del agua de una alberca,</segment>
      <segment deleted="14" preserved="35" seg_ener="154" seg_id="831" seg_lgram="-22.1886" seg_words="9">por lo que existen distintos tipos de termómetros.</segment>
    </sentence>
    <sentence sent_ener="389" sent_id="215" sent_lgram="-48.4060" sent_words="19">
      <segment deleted="0" preserved="49" seg_ener="325" seg_id="832" seg_lgram="-48.4060" seg_words="19">No importa el tipo de termómetro, en todos ellos la temperatura se mide en unidades llamadas grados.</segment>
    </sentence>
    <sentence sent_ener="324" sent_id="216" sent_lgram="-69.7622" sent_words="29">
      <segment deleted="14" preserved="35" seg_ener="23" seg_id="833" seg_lgram="-23.9065" seg_words="7">Cada marca del instrumento es un grado</segment>
      <segment deleted="13" preserved="36" seg_ener="233" seg_id="834" seg_lgram="-50.2147" seg_words="22">y cada tipo de termómetro incluye una escala de medición que, por lo general, se da en grados centígrados.</segment>
    </sentence>
  </source>
</XML>
```

Listing D.2 – *Exemple de fichier XML du corpus contenant toutes les informations de l'annotation pour un texte.*

Les informations des annotations par segment sont disponibles

1. Au format html à l'adresse http://molina.talne.eu/sentence_compression/data/segdata.html
2. Au format comma-separated values (.csv) à l'adresse http://molina.talne.eu/sentence_compression/data/segdata.csv
3. Au format excel binary file Format (.xls) à l'adresse http://molina.talne.eu/sentence_compression/data/segdata.csv

Dans le fichier, chaque ligne représente un segment. Les colonnes sont séparées par le caractère de contrôle 'tab'. Un exemple du format de fichiers pour une seule phrase est :

```
srcid sentid  segid segtype deleted preserved sentener  segener sentw segw  sentlp  seglp segpos  nsegs segtext
...
30  210 821 "D" 39  10  243 4 33  13  -89.9214  -35.6487  1 2 "Para saber qué tan caliente o frío está algo, es decir,"
30  210 822 "D" 0 49  243 171 33  20  -89.9214  -59.8736  2 2 "si se desea conocer la temperatura, debe utilizarse un instrumento que ofrezca un dato confiable, el termómetro."
30  211 823 "D" 0 49  93  76  19  19  -41.0067  -41.0067  1 1 "Este instrumento tiene muchos usos en los hogares, en las industrias y en las unidades de salud."
30  212 824 "D" 3 46  126 84  19  7 -51.8966  -23.0591  1 3 "En casa es útil tener un termómetro"
30  212 825 "D" 31  18  126 4 19  4 -51.8966  -15.8522  2 3 "para saber con precisión"
30  212 826 "D" 27  22  126 0 19  8 -51.8966  -22.6693  3 3 "si alguien de la familia tiene fiebre."
30  213 827 "D" 0 49  326 269 29  18  -65.1852  -47.0171  1 2 "En la industria los termómetros miden la temperatura de hornos y calderas, así como de diversos materiales"
30  213 828 "D" 38  11  326 0 29  11  -65.1852  -23.1637  2 2 "y sustancias que cambian a través de un proceso productivo."
30  214 829 "D" 45  4 455 0 40  3 -89.2905  -12.1723  1 3 "Como ves,"
30  214 830 "D" 7 42  455 183 40  28  -89.2905  -66.5261  2 3 "con frecuencia es necesario medir la temperatura de distintas cosas, del aire, del cuerpo humano, de un horno o del agua de una alberca,"
30  214 831 "D" 14  35  455 154 40  9 -89.2905  -22.1886  3 3 "por lo que existen distintos tipos de termómetros."
30  215 832 "D" 0 49  389 325 19  19  -48.4060  -48.4060  1 1 "No importa el tipo de termómetro, en todos ellos la temperatura se mide en unidades llamadas grados."
30  216 833 "D" 14  35  324 23  29  7 -69.7622  -23.9065  1 2 "Cada marca del instrumento es un grado"
30  216 834 "D" 13  36  324 233 29  22  -69.7622  -50.2147  2 2 "y cada tipo de termómetro incluye una escala de medición que, por lo general, se da en grados centígrados."
...
```

Listing D.3 – *Exemple du format de fichier au format comma-separated values (.csv) contenant les résultats des annotations d'un texte.*

Un autre fichier au format comma-separated values (.csv) contenant les résultats des annotations mais présentés différemment est aussi disponible.

1. Au format html à l'adresse http://molina.talne.eu/sentence_compression/data/all_50BySent.html
2. Au format comma-separated values (.csv) à l'adresse http://molina.talne.eu/sentence_compression/data/all_50BySent.csv
3. Au format excel binary file Format (.xls) à l'adresse http://molina.talne.eu/sentence_compression/data/all_50BySent.xls

Dans ce fichier, chaque ligne représente une compression proposée par les annotateurs. La colonne 'Votes' correspond au nombre de votes obtenus. La colonne 'Bin' représente les segments actifs de la phrase en notation binaire. Dans le fichier, les colonnes sont séparées par le caractère de contrôle 'tab'. Un exemple du format de fichiers pour une seule phrase est :

```
G E Votes Id  Words Asegs Nsegs Bin Text
. . .
  -55.4783  1949  39  210 17  1 2 "0 1" <s> Si se desea conocer la temperatura, debe utilizarse un instrumento que
     ofrezca un dato confiable, el termómetro. </s>
  -89.9214  2080  10  210 28  2 2 "1 1" <s> Para saber qué tan caliente o frío está algo, es decir, si se desea
     conocer la temperatura, debe utilizarse un instrumento que ofrezca un dato confiable, el termómetro. </s>
  -41.0067  311 49  211 17  1 1 "1" <s> Este instrumento tiene muchos usos en los hogares, en las industrias y en
     las unidades de salud. </s>
  -31.2169  384 3 212 11  2 3 "0 1 1" <s> Para saber con precisión si alguien de la familia tiene fiebre. </s>
  -31.9646  1260  1 212 11  2 3 "1 1 0" <s> En casa es útil tener un termómetro para saber con precisión. </s>
  -41.5212  1212  5 212 14  2 3 "1 0 1" <s> En casa es útil tener un termómetro si alguien de la familia tiene
     fiebre. </s>
  -51.8966  1428  14  212 18  3 3 "1 1 1" <s> En casa es útil tener un termómetro para saber con precisión si
     alguien de la familia tiene fiebre. </s>
  -20.9991  1044  26  212 7 1 3 "1 0 0" <s> En casa es útil tener un termómetro. </s>
  -44.6171  2371  38  213 17  1 2 "1 0" <s> En la industria los termómetros miden la temperatura de hornos y
     calderas, así como de diversos materiales. </s>
  -65.1852  2371  11  213 27  2 2 "1 1" <s> En la industria los termómetros miden la temperatura de hornos y
     calderas, así como de diversos materiales y sustancias que cambian a través de un proceso productivo. </s>
  -61.3214  4047  13  214 24  1 3 "0 1 0" <s> Con frecuencia es necesario medir la temperatura de distintas cosas,
     del aire, del cuerpo humano, de un horno o del agua de una alberca. </s>
  -69.7813  4199  1 214 26  2 3 "1 1 0" <s> Como ves, con frecuencia es necesario medir la temperatura de distintas
     cosas, del aire, del cuerpo humano, de un horno o del agua de una alberca. </s>
  -80.8307  5767  25  214 32  2 3 "0 1 1" <s> Con frecuencia es necesario medir la temperatura de distintas cosas,
     del aire, del cuerpo humano, de un horno o del agua de una alberca, por lo que existen distintos tipos de
     termómetros. </s>
  -89.2905  5919  3 214 34  3 3 "1 1 1" <s> Como ves, con frecuencia es necesario medir la temperatura de distintas
     cosas, del aire, del cuerpo humano, de un horno o del agua de una alberca, por lo que existen distintos tipos
     de termómetros. </s>
  -20.0811  1720  7 214 8 1 3 "0 0 1" <s> Por lo que existen distintos tipos de termómetros. </s>
  -48.4060  2236  49  215 17  1 1 "1" <s> No importa el tipo de termómetro, en todos ellos la temperatura se mide en
     unidades llamadas grados. </s>
  -50.0678  2127  14  216 19  1 2 "0 1" <s> Y cada tipo de termómetro incluye una escala de medición que, por lo
     general, se da en grados centígrados. </s>
  -69.7622  2405  22  216 26  2 2 "1 1" <s> Cada marca del instrumento es un grado y cada tipo de termómetro incluye
     una escala de medición que, por lo general, se da en grados centígrados. </s>
  -21.7284  278 13  216 7 1 2 "1 0" <s> Cada marca del instrumento es un grado. </s>
. . .
```

***Listing D.4** – Exemple du format de fichier au format comma-separated values (.csv) contenant les résultats des annotations d'une phrase.*

Annexe E

Principales publications liées à la thèse

Une partie des expériences décrites dans cette étude apparait aussi dans les publications listées ci-dessous.

1. *La compresión de frases : un recurso para la optimización de resumen automático de documentos*. Molina A., da Cunha I., Torres Moreno J., Velázquez Morales P. : Linguamática, 2011, vol. 2, no 3, p. 13–27. ISSN : 1647–0818
2. *Discourse Segmentation for Sentence Compression*. Molina A., Torres Moreno J.M., SanJuan E., da Cunha I., Sierra G.,Velázquez-Morales P. : In Advances in Artificial Intelligence LNCS. Springer-Verlag. Vol.7094, pp. 316–327. ISSN : 0302–9743.
3. *Sentence Compression in Spanish driven by Discourse Segmentation and Language Models*. Molina A., Torres Moreno J.M., da Cunha I., SanJuan E., Sierra G. Cornell University ArXiv :1212.3493, Computation and Language (cs.CL), Information Retrieval (cs.IR), Vol.1212. 2012
4. *Segmentação discursiva automática : uma avaliação preliminar em francês*. Saksik R., Molina A., Linhares A., Torres Moreno J.M. : In Proceedings of the 4th Meeting RST and Discourse Studies.
5. *A turing test for evaluate a complex summarization task*. Molina A., Torres Moreno J.M., SanJuan E. : In Proceedings of the Conference and Labs of the Evaluation Forum (CLEF 2013).

Index