*GAËL HARRY ADÉLIO ANDRÉ DIAS*

❖

# EXTRACTION AUTOMATIQUE D'ASSOCIATIONS LEXICALES À PARTIR DE CORPORA

Dissertation présentée pour l'obtention du diplôme de Docteur en Informatique de l'Université Nouvelle de Lisbonne, Faculté des Sciences et Technologie. Cette dissertation a été préparée dans le cadre du protocole de co-tutelle entre l'Université Nouvelle de Lisbonne et l'Université d'Orléans – France.

LISBONNE
2002

# Remerciements

Une thèse de doctorat est l'aboutissement de nombreuses années d'études qui ont vu se succéder professeurs, camarades et amis. Pour être juste, toutes ces personnes qui ont marqué positivement ce long chemin peu indulgent devraient être mentionnées dans cette partie qui leur est réservée de droit. Malheureusement, cette liste est bien trop grande pour être reportée intégralement. Ainsi, bon nombre de personnes ne verront pas leur nom mentionné bien que leur présence ait toujours été appréciée à juste titre. A tous ces anonymes, un grand merci.

Néanmoins, la réalisation de ce travail n'a été possible que grâce à l'appui incessant d'un petit groupe de personnes que je me dois de remercier personnellement.

Je tiens à remercier tout particulièrement Sylvie Guilloré pour m'avoir transmis le plus important des atouts : le goût pour le traitement automatique des langues naturelles. En particulier, sa confiance en mes capacités et son soutien de tous les jours m'ont permis d'affronter tous les obstacles survenus dans les meilleures conditions possibles.

José Gabriel Pereira Lopes mérite mes sincères remerciements pour m'avoir toujours appuyé durant ces quatre années de travail quotidien. Sa disponibilité sans pareil et son soutien tant au niveau professionnel que personnel ont été les fondements de la réussite de cette recherche. Sans son aide incessante et ses connaissances hors pair, ce travail n'aurait certainement pas atteint les niveaux d'exigence démontrés.

Je tiens également à remercier Jean-Claude Bassano pour sa patience et sa compréhension de tous les moments. En particulier, son soutien inébranlable vis-à-vis de mon travail m'a permis de développer sereinement l'ensemble de mes idées. A lui, un grand merci.

Les réunions et les discussions entre collègues ont été à l'origine d'un partage scientifique et culturel des plus fructueux. Tout au long de cette recherche, celles-ci ont permis de confronter mes idées à celles de mes camarades pour pouvoir ainsi mieux les défendre. Je tiens donc à remercier tous les membres du GLINT de l'Université Nouvelle de Lisbonne ainsi que tous les membres du LIFO de l'Université d'Orléans. En particulier, je tiens à remercier Pierre Réty pour m'avoir aidé à formaliser les calculs de complexité et Helena Ferreira de l'Université de la Beira Interior pour avoir pressenti la règle de récursivité de l'Expectative Mutuelle.

Je tiens enfin à remercier les membres du jury pour leurs remarques pertinentes et pour leur disponibilité face à la lourde tâche que représente la lecture de ce rapport.

Je veux bien sûr remercier mes parents, ma soeur Séverine, mes beaux parents et toute ma famille pour avoir su me soutenir dans les moments difficiles, me motiver dans les moments d'égarements et me comprendre dans les moments de doute. Sans eux, rien de tout ce travail n'aurait été possible. Malgré mes absences répétées et mes longs silences, mon coeur ne vous a jamais oublié. Cette thèse vous est entièrement dédiée.

Mes amis ont également contribué pour une grande part au succès de ce travail. En effet, les sorties ainsi que les confidences sont des moments nécessaires pour assurer un équilibre émotionnel qui est régulièrement mis à dure épreuve. En particulier, un grand merci à Pedro, Célia, Sofia, Ana, Paulo et Simon.

Finalement, je tiens à remercier de tout mon coeur ma femme Sandra qui a partagé mes bonheurs comme mes incertitudes tout au long de ces quatre années de travail. J'espère pouvoir la récompenser de toute sa patience lors de la réalisation de son mémoire de DEA. A toi, Xana, le plus grand des merci ■

# Résumé en Français

L'acquisition automatique d'associations lexicales à partir de corpora revêt une importance cruciale dans le cadre du Traitement Automatique des Langues Naturelles. Ainsi, une association lexicale est une combinaison récurrente de mots simples qui se trouvent ensemble plus souvent que par le simple fait du hasard dans un domaine donné. Les associations lexicales définissent en fait des phénomènes linguistiques tels que les expressions idiomatiques, les idiotismes, les mots composés, les synapsies ou encore les lexies complexes. Du fait du cararctère non compositionnel du sens des associations lexicales, leur identification s'est rapidement révélée primordiale pour la réalisation d'analyses et de synthèses de textes qui prennent en compte toutes les subtilités de la langue. Dans le cadre de ce rapport, nous introduisons une nouvelle architecture purement statistique qui permet l'extraction d'associations lexicales contiguës ou non à partir de corpora non annotés. Dans ce but, nous avons défini trois concepts originaux : les modèles $N$-gram positionnels, l'Expectative Mutuelle et l'algorithme GenLocalMaxs. Ainsi, l'énoncé initial est d'abord transformé en un ensemble de $N$-grams positionnels i.e. en vecteurs ordonnés d'unités lexicales simples. Une mesure d'association, l'Expectative Mutuelle, est ensuite chargée de définir le degré de cohésion de chaque $N$-gram positionnel. Finalement, l'algorithme GenLocalMaxs identifie les $N$-grams positionnels pertinents grâce au repérage de valeurs de cohésion localement maximales. De nombreux efforts ont également été déployés afin d'évaluer cette méthodologie. Dans ce cadre, nous avons proposé la normalisation de cinq mesures d'association couramment utilisées dans le domaine afin de valider les améliorations mises en évidence tant par l'Expectative Mutuelle que l'algorithme GenLocalMaxs ■

# Résumé en Anglais

The automatic acquisition of lexical associations from corpora is a crucial issue for Natural Language Processing. A lexical association is a recurrent combination of words that co-occur together more often than expected by chance in a given domain. In fact, lexical associations define linguistic phenomena such as idioms, collocations or compound words. Due to the fact that the sense of a lexical association is not compositionnal, their identification is fundamental for the realization of analysis and synthesis that take into account all the subtilities of the language. In this report, we introduce a new statistically-based architecture that extracts from naturally occurring texts contiguous and non contiguous. For that purpose, three new concepts have been defined : the positionnal $N$-gram models, the Mutual Expectation and the GenLocalMaxs algorithm. Thus, the initial text is fisrtly transformed in a set of positionnal $N$-grams i.e ordered vectors of simple lexical units. Then, an association measure, the Mutual Expectation, evaluates the degree of cohesion of each positionnal $N$-gram. Finally, the GenLocalMaxs algorithm extracts the pertinent positionnal $N$-grams based on the identification of local maximum values of Mutual Expectation. Great efforts have also been carried out to evaluate our metodology. For that purpose, we have proposed the normalisation of five well-known association measures and shown that both the Mutual Expectation and the GenLocalMaxs algorithm evidence significant improvements comparing to existent metodologies ■

# Table des matières

# Liste des tableaux

# Table des figures

# Préface

Avant de passer à l'essence même de mon travail, il convient de replacer l'ensemble de cette recherche dans son contexte organisationnel et de déterminer ses motivations.

Cette thèse de doctorat a été avant tout le fruit d'une collaboration fructueuse entre l'Université d'Orléans et l'Université Nouvelle de Lisbonne. Cette relation n'a pas été fortuite. En effet, c'est lors de mon DEA que j'ai pris les premiers contacts avec José Gabriel Pereira Lopes de l'Université Nouvelle de Lisbonne. Sylvie Guilloré et Jean-Claude Bassano, mes professeurs à l'Université d'Orléans, me conseillaient alors de partir à Lisbonne pour réaliser mon stage de DEA en traduction automatique. Après un semestre d'une expérience enrichissante, de très forts liens se nouaient entre José Gabriel Pereira Lopes et moi-même ainsi qu'entre les deux groupes de Langage Naturel des deux Universités.

Cette collaboration se matérialiserait un an plus tard par la signature d'un contrat de thèse en co-tutelle entre l'Université d'Orléans et l'Université Nouvelle de Lisbonne après une courte escapade à l'Université de Sheffield pour travailler avec Robert Gaizauskas et Yorick Wilks.

A partir de ce moment, je commençais à travailler principalement à Lisbonne dans le cadre du projet DIXIT[1]. De courts séjours au Laboratoire d'Informatique Fondamentale d'Orléans — LIFO — étaient néanmoins organisés grâce à la participation au projet RELING[2] dont l'un des buts principaux était la réalisation de thèses co-tutelles entre les

[1]Projet financé par le Ministère des Sciences et de la Technologie Portugais - Ref. 2/2.1/TIT/1670/95 : DIXIT - Systèmes de dialogue intentionnel et d'aide à la traduction - Chercheur principal : José Gabriel Pereira Lopes.

[2]Projet financé par la Fondation pour les Sciences et la Technologie Portugaise - Réseaux Formation/Recherche en Contraintes et Langage Naturel - Chercheur principal pour le Langage Naturel : José

Universités de Lisbonne, d'Orléans et l'INRIA de Rocquencourt.

Après un an de travail, mes efforts étaient récompensés par l'obtention d'une bourse de doctorat concédée par la Fondation Portugaise pour les Sciences et la Technologie pour une durée de trois ans[3]. Les bases pour la réalisation d'un bon travail étaient alors construites.

L'axe principal de ma recherche se fixait alors sur l'identification automatique d'associations lexicales à partir de corpora. En effet, dans les systèmes de dialogue, la phase de réalisation lexicale est particulièrement importante. Celle-ci consiste à choisir les mots les plus adéquats qui permettront de transmettre l'idée originelle du message sans la déformer. Dans ce contexte, nous nous sommes rapidement vus confronter à la complexité de l'analyse lexicale et particulièrement à la réalisation par unités polylexicales. En effet, l'utilisation d'associations lexicales pour transmettre un unique concept sémantique est particulièrement fréquent. Or, nous ne disposions d'aucune ressource contenant ce type d'information pour le Portugais. Ainsi, la réalisation d'un logiciel d'extraction s'imposait.

Dans l'optique multilingue du projet DIXIT, notre logiciel devait alors répondre à certaines exigences. Premièrement, le logiciel devait pouvoir être appliqué indépendamment au Portugais et au Français. Deuxièmement, il était intéressant qu'il puisse être le plus flexible possible afin de pouvoir s'adapter sans problème à d'autres langues et ceci quelle que soit leur notion de segmentation[4]. Dans ces conditions, la lemmatisation, les définitions de règles syntaxiques ou de listes de mots vides n'étaient pas envisageables.

Mon travail s'est donc orienté vers les travaux privilégiant les heuristiques numériques. Dans ce cadre, nous avons introduit trois nouveaux concepts : les modèles $N$-gram positionnels, l'Expectative Mutuelle et l'algorithme de sélection GenLocalMaxs. J'espère donc que ce travail vienne combler un certain nombre de lacunes existantes et propose de nouvelles directions pour des recherches futures.

Mon rapport sera composé de quatre parties principales : une introduction pour sensibi-

Gabriel Pereira Lopes.

[3]Référence : GGPXXI/BD/3895/96.

[4]La segmentation des langues orientales se fait au niveau du caractère par exemple.

liser le lecteur au problème des associations lexicales, une partie architecture où tous les concepts théoriques et pratiques seront introduits, une section d'évaluation dans laquelle nous illustrerons les résultats obtenus et finalement une conclusion qui donnera un certain nombre de pistes pour améliorer le système.

Nous vous souhaitons donc une bonne lecture ■

.

# Première partie

# Préliminaires

**“Les considérations statistiques sont essentielles à la compréhension de l’activité et du développement du langage”**

---

*Lyons [2]*

# Chapitre 1

# Introduction

La globalisation de l'outil informatique et l'avénement de la toile ont radicalement changé le panorama du Traitement Automatique des Langues Naturelles — TALN. Dans les années cinquante, la linguistique structurale et le courant de pensée chomskyen faisaient école. La formalisation et l'intuition du locuteur natif prévalaient sur l'empirisme de la linguistique descriptive. Le TALN se donnait alors comme objectif principal l'analyse exhaustive des textes et leur compréhension complète. Parallèlement, le volume des textes disponibles en format électronique ne cessait d'augmenter. Par voie de conséquence, les impératifs économiques se métamorphosaient et les objectifs du TALN accompagnaient cette tendance. Aujourd'hui, accéder rapidement à l'information, la traduire et la résumer sont les enjeux d'un futur proche. Ainsi, bien plus que l'analyse de cas d'école, il est nécessaire de confronter les techniques élaborées aux réalités du matériel écrit. A partir de ce constat, une linguistique faisant systématiquement appel à des collections de textes pour développer des dictionnaires et des grammaires descriptives, mais aussi tester des hypothèses et confronter des modèles aux réalisations effectives s'est progressivement affirmée. C'est le courant de la linguistique de corpus. Dans ce cadre, l'étude de la dimension phraséologique du langage s'est révélée primordiale pour la réussite des nouveaux projets du TALN.

## 1.1 Phraséologie

La Phraséologie se donne comme objectif principal l'analyse des figements des éléments textuels dont le langage fait foison. Suivant cette nouvelle perspective d'analyse, J. Sinclair [3] défend que la langue met à disposition un grand nombre de phrases préconstruites dont les éléments se constituent en choix uniques. C'est le principe de l'idiomaticité. A.

Pawley [4] corrobore cette opinion et démontre que la "naturalité" et la perception de la "fluidité" du processus de production de la langue se doit en grande partie à l'utilisation d'un grand nombre d'expressions préfabriquées. Ainsi, l'identification de concepts décrits à l'aide de plusieurs occurrences de mots s'est peu à peu imposée comme l'une des tâches fondamentales de la compréhension et la production du langage. Cependant, son importance a longuement été méconnue. Ce n'est pas que le sujet ait été totalement ignoré : presque toutes les grammaires le traitaient dans un chapitre consacré à la formation des mots. Mais l'ampleur du phénomène échappait aux auteurs, à l'exception, peut-être, de O. Jespersen [5] qui oppose la liberté combinatoire au figement comme mécanisme de formation des mots. Ce phénomène est donc resté marginal et la perception collective simpliste : les mots composés sont ceux qui ont un trait d'union. Outre cette représentation caricaturale, le figement linguistique a également été obscurci par des dénominations floues et très hétérogènes parmi lesquelles on trouve les termes suivants : expression idiomatique, idiotisme, mot composé, synapsie, synthème et lexie composée. Cette tendance s'est cependant rapidement inversée grâce à l'accès libéralisé à de grandes quantités de textes. En effet, la nécessité de tester les applications dans des conditions réelles d'utilisation a permis de mettre en évidence les lacunes inhérentes aux études basées uniquement sur l'analyse des mots simples. En effet, les unités polylexicales s'avèrent souvent opaques dans la phase de compréhension et cause d'hésitations dans la phase de production de la langue. Par exemple, les expressions toutes faites, comme les noms composés (*coup de coeur*), les verbes composés (*mettre en évidence*), les locutions adverbiales (*en plein dans le mille*), prépositives (*en direction de*) ou conjonctives (*au fur et à mesure que*) ainsi que les déterminants composés (*un tas de*) démontrent souvent un figement sémantique et transformationnel qui rend difficile leur analyse et leur interprétation. Il est donc clair que la maîtrise de ces "mots en plusieurs mots" est essentielle pour l'apprentissage des langues.

## 1.2 Applications

Ainsi, reconnaître les unités lexicales complexes des collections de textes toujours croissantes constitue un enjeu fondamental dans le cadre de la recherche documentaire. En effet, l'identification des termes complexes permet soit d'indexer les textes avec une plus grande précision soit de guider l'utilisateur dans sa quête d'information. Dans le premier cas, la sélection de termes discriminants (ou decripteurs) pour représenter le

contenu des textes est un problème critique. Idéalement, les termes d'indexation devraient décrire directement les concepts présents dans les documents. Cependant, la plupart des systèmes de recherche d'information indexent les textes de la base documentaire à partir d'unités lexicales simples. Or, ces unités atomiques ne sont pas suffisamment spécifiques pour évoquer le contenu des textes. Afin d'améliorer la qualité de l'indexation, certains systèmes bénéficient de l'existence de thesauri prédéfinis. Dans ce cas, les descripteurs sont choisis parmi les éléments du thesaurus. Ainsi, les mécanismes pour retrouver les documents utilisent les liens directs entre le thesaurus et les textes, et quelquefois les liens de synonymie, d'hyperonomie et/ou d'hyponomie entre les éléments du thesaurus [6]. Malheureusement, de nombreux domaines ne disposent pas de thesauri spécialisés et parallèlement peu de projets incluent la construction automatique de thesauri [7] [8]. L'identification des unités lexicales complexes d'une base documentaire propose une alternative aux problèmes précédemment exprimés. En effet, les mots composés sont souvent moins ambigus et plus motivés que les unités lexicales simples et permettent ainsi une meilleure approximation des thèmes abordés. Ainsi, les termes complexes sont utilisés pour représenter le contenu des textes sous la forme d'indexes [9]. Dans le deuxième cas, le système de recherche documentaire doit pouvoir guider l'utilisateur dans sa quête d'information et ainsi lui permettre de raffiner sa requête en lui proposant une liste de descripteurs qui s'apparient à sa requête initiale. L'utilisateur doit ainsi pouvoir "voyager" dans l'espace de la base documentaire à partir des différents concepts listés par le système. Le pouvoir de représentation des unités lexicales complexes joue un rôle primordial dans cette opération en proposant une meilleure discrimination des sujets abordés par les textes [10].

La maîtrise des unités polylexicales est également indispensable pour le succès des systèmes de traduction automatique. Dans ce contexte, le figement sémantique des unités lexicales complexes constitue un obstacle bien défini de la phase de production du langage. Ceci se traduit par la difficulté des non natifs à produire des traductions "naturelles". En effet, le sens des mots qui constituent les termes complexes n'intervient pas forcément dans leur interprétation. En d'autres termes, le sens des mots composés n'est pas calculable à partir du sens des unités lexicales simples qui les composent. Le concept de compositionalité n'est donc pas applicable causant ainsi un certain nombre d'hésitations durant le processus de traduction. Ce rejet de la compositionalité est

pourtant discutable. En effet, un certain nombre de chercheurs tels que P. Dowing [11] et D. Corbin [12] défendent qu'il existe des mécanismes compositionnels de construction du sens des unités polylexicales, même si ceux-ci sont complexes. Cependant, ces mécanismes se montrent insuffisamment informatifs pour permettre la réalisation de traductions fluides et leur codage difficile n'avantage pas le traitement informatique. Par exemple, il est clair que l'expression *Assemblée Nationale* ne révèle pas un sens opaque. Bien au contraire, l'Assemblée Nationale est une assemblée qui représente la Nation. Cependant, plus que cette interprétation, le terme *Assemblée Nationale* traduit un concept mental, celui d'institution administrative. Ainsi, si l'on considère uniquement son interprétation compositionnelle, cette expression devrait être traduite en Portugais par la juxtaposition des unités atomiques *Assembleia* — *Assemblée* — et *Nacional* — *Nationale*. Or, sa traduction "naturelle" est formulée par l'expression *Assembleia da República* — *Assemblée de la République*. Bien que la première traduction n'introduise pas une représentation sémantique indéchiffrable pour un locuteur de langue portugaise, un certain effort de compréhension serait nécessaire[1]. Dans ce contexte, il est clair que l'interprétation des expressions toutes faites dépasse le simple cadre de la sémantique compositionnelle. Ainsi, les études basées sur l'analyse des unités lexicales simples sont insuffisantes pour la réalisation de traductions fluides et "naturelles"[2]. Dans ce sens, l'identification des expressions présentes dans les textes est une étape de normalisation incontournable pour la réalisation d'applications de traduction de qualité.

Nous ne prétendons pas, dans cette partie introductive, donner une liste exhaustive des applications pour lesquelles la maîtrise des unités polylexicales est un enjeu fondamental. Cependant, il est évident que les recherches en étiquetage morpho-syntaxique, desambiguïsation, construction de terminologies, résumé automatique de textes, classification de textes et bien d'autres sont particulièrement sensibles à l'utilisation de bases de données lexicales introduisant des connaissances spécifiques sur les unités lexicales complexes.

[1] Les phénomènes de figement dépendent également du contexte dans lequel ils sont utilisés. Ainsi, avant la révolution des oeillets, le terme correcte aurait été *Assembleia Nacional* et non pas *Assembleia da República* !

[2] Le système SYSTRAN en est le parfait exemple.

## 1.3 Ressources

Dans ce cadre, de nombreux travaux se sont donnés comme objectif principal la création de ressources linguistiques intégrant la notion de figement. I. Mel'cuk est l'un des premiers, pour la langue française, à donner une place centrale aux mots composés dans son *Dictionnaire Explicatif et Combinatoire du Français*. Ainsi, il définit des fonctions lexicales [13] qui visent à mettre au jour les réalisations lexicales les plus probables pour exprimer les modifications sémantiques des mots. Par exemple, le degré fort se dit *à chaudes larmes* lorsqu'il s'agit de *pleurer* et *à tout rompre* pour le verbe *applaudir*. Suivant la même lignée, les études de M. Gross et M. Silberztein menées au LADL — Laboratoire d'Analyse Documentaire et Linguistique de l'Université de Paris 7 — sur les possibilités combinatoires des mots simples, ont abouti à la réalisation d'un dictionnaire électronique des mots composés du Français [14] : le DELAC — Dictionnaire électronique des mots composés. Ces travaux s'inscrivent depuis le début des années quatre-vingt dans une perspective de listage qui a mis en évidence l'importance numérique du phénomène de figement dans l'étude du Français. Ainsi, on estime que les unités complexes occupent un cinquième de la surface des textes ! Devant le succès de ces travaux, de nombreux projets de recherche se sont organisés en Europe pour la compilation de versions localisées du DELAC. En particulier, au Portugal, I. Ranchod dirige une équipe de linguistes pour la réalisation de l'équivalent portugais du DELAC.

Cependant, la construction "manuelle" de bases de données lexicales exhaustives représente un travail dantesque impossible à réaliser. En effet, l'ensemble des unités polylexicales est ouvert et à compléter. De fait, le dynamisme du langage interdit la définition exhaustive de tous les mots composés de la langue. En particulier, une partie essentielle de la néologie lexicale s'opère par le biais de séquences complexes de mots simples. Par exemple, *vie artificielle*, *algorithme génétique*, *programmation évolutive* sont des termes complexes particulièrement récents dans le domaine de l'informatique qu'il est peu probable de trouver répertoriés dans les bases de connaissance lexicale même spécialisées. En effet, la création et la maintenance "manuelle" de banques terminologiques est un travail long, difficile et aventureux qui ne peut accompagner l'accroissement incessant des masses de textes à traiter et par conséquent l'évolution du langage. Ainsi, afin de réduire cet écart, l'un des enjeux de la recherche en TALN est de fournir des outils permettant l'extraction automatique d'unités lexicales complexes à partir de collections de textes spécialisés.

## 1.4 Méthodologies

Le coût et la difficulté de la construction de ressources lexicales ont mis à l'honneur les méthodes automatiques et semi-automatiques d'acquisition de connaissances lexicales qui considèrent les collections de textes — corpora — comme des sources précieuses d'information. Dans ce cadre, le problème du dépouillement lexicographique a traditionnellement été abordé suivant trois axes bien distincts. La première approche utilise des techniques structurelles fondées sur l'analyse syntaxique de l'énoncé [15] [16]. Dans ce cadre, l'analyse de patrons de surface spécifiques permet l'identification d'un nombre significatif d'unités lexicales complexes qui démontrent certaines régularités syntaxiques. Cependant, cette démarche nécessite des connaissances approfondies de la langue freinant ainsi son application à de nouveaux domaines [17] ou à de nouvelles langues[3]. Afin de faire face à ces problèmes, une nouvelle méthodologie propose des techniques hybrides qui associent modèles statistiques et filtres syntaxiques [21] [22] [23] [24] [25] [26]. Dans ce contexte, des patrons syntaxiques superficiels sont définis *a priori* et repèrent des séquences d'unités lexicales simples dont le degré de figement est évalué par le biais de mesures mathématiques. Malheureusement, ces systèmes n'autorisent généralement que l'extraction de termes complexes de type nominal. De plus, leur application à de nouvelles langues nécessite la redéfinition des filtres syntaxiques. Finalement, la troisième approche propose des techniques statistiques et numériques qui décèlent les associations préférentielles présentes dans les collections textuelles [27] [28] [29] [30] [31] [32] [33] [34]. Dans ce contexte, une mesure ou une combinaison de mesures mathématiques évalue le degré d'attraction qui lie entre elles toutes les unités textuelles de séquences construites à partir des énoncés. Cette méthode se différencie donc des approches précédentes par sa flexibilité d'adaptation à de nouveaux domaines et à de nouvelles langues ainsi que par l'ensemble non-restreint des termes complexes extraits.

Cependant, les systèmes proposés exhibent trois inconvénients majeurs. Premièrement, ils recourent à la définition de valeurs seuil globales dans le cadre du processus d'acquisition. Ces seuils sont des valeurs limites de fréquence ou de mesure d'association qui permettent de définir si les séquences d'unités lexicales sont d'intérêt ou non. Ils font donc l'objet d'un ajustement qui est crucial pour la réussite des expériences statistiques. Il s'agit d'un

[3]Cette limitation n'est cependant pas si criante aujourd'hui grâce aux méthodes automatiques de repérage de patrons syntaxiques à partir de textes étiquetés [18] [19] [20].

compromis entre des valeurs assez permissives pour que la collecte soit importante — taux de rappel — et des valeurs trop généreuses pour que le résultat soit précis — taux de précision. Dans le cadre de totale flexibilité dressé par cette dernière méthodologie, le concept de valeur seuil vient entâcher son développement. Il est clair qu'une solution plus fiable et plus robuste s'impose. Deuxièmement, la plupart des modèles mathématiques n'évaluent que les associations binaires — i.e. entre deux mots simples — et doivent recourir à l'utilisation de techniques d'amorçage pour l'extraction d'associations $N$-aires — i.e. entre $N$ mots simples. Celles-ci définissent un processus itératif où l'acquisition d'associations de plus de deux mots requiert un travail complémentaire où les "composés binaires" acquis initialement jouent le rôle d'amorce. Or, comme nous le verrons plus en détail dans ce rapport, le résultat des méthodes d'amorçage dépend foncièrement des associations binaires retenues lors de la première étape du processus. Là encore, une solution générique s'impose. Troisièmement, les modèles mathématiques proposés dans la littérature se montrent particulièrement sensibles aux unités lexicales fréquentes et par conséquent sous-évaluent généralement les associations entre constituants. Ainsi, B. Daille [21] et C. Enguehard [23] — entre autres — ne considèrent que les occurrences des mots pleins pour évaluer les forces de cohésion et évitent l'intégration des fragments fonctionnels, souvent fréquents, dans l'application des mesures numériques. Une fois de plus, les restrictions imposées par le choix des unités de décompte vérifient la nécessité d'interventions extérieures dans le processus d'acquisition réduisant ainsi la flexibilité des méthodes numériques. Dans ce cadre, il est clair que la définition de nouveaux modèles mathématiques est nécessaire pour résoudre les problèmes mis en évidence par les systèmes existants.

L'étude que nous présentons dans ce rapport résulte des trois problèmes énoncés précédemment. Notre objectif fondamental se résume donc à proposer un système d'extraction d'unités lexicales complexes totalement flexible. Afin de répondre à celui-ci, nous avons dû dans un premier temps refuser quelconque annotation des textes. En effet, il est évident que la lemmatisation et l'étiquetage morpho-syntaxique sont deux processus qui ne sont pas forcément disponibles pour les langues étudiées même si le développement d'analyseurs de plus en plus performants permet de résoudre ce problème — du moins pour l'étiquetage morpho-syntaxique — à partir de faibles quantités de textes préalablement

annotés "à la main" [20][4]. Mais, plus important, l'introduction d'informations linguistiques implique l'addition de contraintes qui ne sont pas originellement comprises dans les textes. Par exemple, il n'est pas certain que la lemmatisation systématique [35] ou la morphologie dérivationnelle, avec notamment le regroupement des mots appartenant à la même famille dérivationnelle — *stemming* [36], améliorent les performances du processus d'acquisition. Par ailleurs, les traitements linguistiques sont lourds et parfois imprécis. Dans ce dernier cas, les risques de l'utilisation de patrons syntaxiques lors du processus d'extraction sont évidents. Parallèlement, l'épurage des textes à partir de listes de mots vides a été jusqu'ici accepté sans préoccupation par la communauté scientifique. Il faudrait pourtant évaluer la perte d'information due à l'utilisation exclusive des mots pleins pour le calcul des associations textuelles [21] [23]. En effet, cette approche définit les mots fonctionnels comme étant dénués de sens lexicographique. Or, ceci est loin d'être sûr et pose de nombreux problèmes d'interprétation. Par exemple, la dénomination internationale de la ville portugaise Porto est exprimée par l'unité complexe *O Porto* qui pourrait être grossièrement traduite en Français par *Le Porto*. Dans ce cas, l'article défini *O* qui peut être considéré comme un fragment fonctionnel fait partie intégrante de l'unité complexe et ne peut en aucun cas en être retiré. Ainsi, notre préoccupation a été de n'utiliser que l'information explicitement présente dans les textes et dans toute son intégralité. Nous sommes bien entendu convaincus de la nécessité de l'apport linguistique dans le processus d'acquisition de connaissances lexicales. Dans ce sens, G. Grefenstette [37] montre que le dosage efficace entre méthodes linguistiques et statistiques bénéficie une grande partie des applications du TALN. M. Sussna [38] postule même que la description la plus riche est nécessairement la plus appropriée. En effet, les travaux statistiques basés sur l'étude des corpora permettent d'identifier des régularités lexicales mais n'autorisent en aucun cas leur enregistrement direct sans validation préalable. Proposer le plus grand nombre d'unités lexicales complexes pertinentes afin de diminuer le traitement linguistique qui, comme nous l'avons vu, implique un certain nombre de contraintes qu'il reste à évaluer de façon décisive, doit donc être l'objectif primordial d'un analyseur syntaxique.

A partir de ce constat, nous avons développé une méthodologie qui se base dans un premier temps sur la segmentation du texte en $N$-grams positionnels — séquences conti-

[4]Dans cette étude, N. Marques *et al.* utilisent à peine 5 000 mots étiquetés morpho-syntaxiquement alors que les méthodes usuelles nécessitent de cent fois plus de ressources annotées.

nues ou non d'unités lexicales ou mots. En effet, de nombreux travaux lexicographiques [39] [3] [40] [41] montrent que la plupart des relations lexicales associent des mots qui sont séparés les uns des autres par au plus cinq autres mots [5]. Or, les unités lexicales complexes sont des relations lexicales spécifiques. Dans ce contexte, une unité lexicale complexe peut être représentée par un vecteur ordonné de mots simples c'est-à-dire un $N$-gram positionnel où $N$ correspond au nombre de mots du vecteur. L'introduction des positions de chaque mot par rapport à une unité de référence appelée mot pivot introduit une nouveauté par rapport aux modèles $N$-grams classiques qui utilisent des séquences continues d'unités lexicales ou bien ne considèrent que des ensembles de mots désordonnés. En effet, l'organisation des unités textuelles en groupes cohérents fortement liés est un phénomène complexe qui ne peut être limité aux simples modèles $N$-gram classiques présentés dans [42] :193. Dans ce cadre, tant le contexte immédiat antérieur que postérieur d'un mot influencent sa présence dans un énoncé. Ainsi, l'occurrence d'un mot n'est généralement pas déterminée par la seule séquence des mots qui le précèdent mais plutôt par son environnement immédiat — contexte antérieur et postérieur. Afin de rendre compte de forme précise de tous ces phénomènes, nous avons dû introduire cette nouvelle représentation qui nécessite un traitement computationnel diligent. Le chapitre 4 donne toutes les bases théoriques de cette nouvelle représentation.

A partir de la définition des modèles $N$-gram positionnels, les méthodes de la statistique textuelle préconisent l'application de mesures d'association pour calculer le degré de cohésion liant entre eux tous les mots des séquences préalablement construites et ainsi déterminer leur pertinence. Cependant, la plupart de ces mesures n'ont été définies que pour les modèles digrams d'unités textuelles — modèles $N$-gram pour $N = 2$ — et ne permettent donc pas de mesurer les forces d'attraction qui existent entre tous les constituants d'un $N$-gram générique — $\forall N, N \geq 2$. On les appelle mesures d'association binaires. Par conséquent, l'acquisition d'associations de plus de deux mots requiert un travail complémentaire où les paires d'associations acquises initialement jouent le rôle d'amorce [31] [43]. Ainsi, à partir des associations binaires préalablement extraites, une nouvelle étude du voisinage du couple de mots est effectuée. Les unités lexicales dont la probabilité d'apparition dépasse une certaine valeur sont alors concaténées aux deux unités de l'association binaire pour former une unité complexe de taille supérieure à deux.

[5] Cette barrière est restrictive et il est évident que certaines associations dépassent cette limite.

Le processus se répète jusqu'à ce qu'il n'existe plus d'unités lexicales à associer. C'est ce que l'on appelle l'amorçage. Cette méthodologie n'est évidemment pas envisageable dans le cadre des modèles $N$-gram positionnels. En effet, une mesure d'association doit être capable de mesurer le degré de pertinence d'un $N$-gram positionnel quelconque c'est-à-dire indépendamment de la valeur de $N$. Dans ce contexte, nous proposons dans le chapitre 5, une nouvelle mesure d'association appelée Expectative Mutuelle. L'Expectative Mutuelle est une mesure probabiliste de l'approche non paramétrique qui permet de calculer les forces d'attraction qui lient entre eux tous les mots d'un $N$-gram positionnel pour tout $N$ tel $N \geq 2$. Elle est formellement définie à partir des notions d'Expectative Normalisée et de fréquence relative qui calquent respectivement les concepts de *support* et de *confiance* formulés par R. Agrawal [44] dans le cadre de la définition de règles d'association.

En particulier, nous verrons dans le chapitre 8 que l'Expectative Mutuelle démontre la bonne propriété de ne pas sous-évaluer les associations qui contiennent des unités textuelles fréquentes. Nous verrons que ce n'est pas le cas pour un certain nombre de mesures d'association binaires pour lesquelles nous proposerons une normalisation dans le chapitre 6 définissant ainsi une méthodologie "universelle" pour le calcul des forces d'attraction entre tous les mots d'un $N$-gram positionnel générique.

Après la définition du degré de cohésion des $N$-grams positionnels, la dernière étape d'un système d'extraction correspond à identifier l'ensemble des séquences pertinentes. Cette identification peut être définie comme étant la tâche qui consiste à détecter, parmi l'ensemble des $N$-grams positionnels pondérés selon leur degré de cohésion, un sous-ensemble d'éléments qui partagent certaines caractéristiques propres au concept de mot composé. En particulier, on dira que ces $N$-grams positionnels sont pertinents. Par conséquent, l'extraction d'unités lexicales complexes dépend intrinsèquement de la définition de l'ensemble de ces caractéristiques et donc de la notion de pertinence. Paradoxalement, peu de travaux se sont attaqués à ce problème. En effet, la plupart des approches préconisent l'utilisation de valeurs seuil de fréquence ou de mesure d'association qui sont loin d'être satisfaisantes. Dans ce cadre, la plupart des approches se basent sur la définition de valeurs limites — ou seuil — qui permettent de diviser en deux groupes distincts l'ensemble des séquences d'unités textuelles [22] [33] [23] [45] [31] [32] [34] [23] [28] [27] [21] : d'un côté les pertinentes et de l'autre le reste. Cependant, il n'existe

pas de fondement théorique permettant de définir ces valeurs limites. De ce fait, elles sont généralement imposées par l'expérimentation et, le genre, la longueur, le domaine ou la langue considérés sont autant de paramètres à prendre en compte lors de leur définition. Afin de répondre à l'ensemble de nos objectifs de flexibilité, les valeurs seuil doivent être évitées et une nouvelle définition de pertinence doit être proposée. Dans ce cadre, J. Silva *et al.* [30] proposent d'étudier l'environnement immédiat des séquences d'unités textuelles afin de déterminer leur pertinence. Ainsi, pour les modèles $N$-gram positionnels, nous proposerons dans le chapitre 5, l'algorithme GenLocalMaxs qui définit un $N$-gram positionnel comme pertinent si l'ajout et l'élimination d'un mot du $N$-gram positionnel n'entachent en rien son degré de cohésion — dans le cadre de notre étude, l'Expectative Mutuelle. Le GenLocalMaxs est en fait la généralisation de l'algorithme LocalMaxs présenté par J. Silva *et al.* [30] pour les modèles $N$-grams classiques.

En guise de résumé, la combinaison des méthodes suivantes

- Modèles $N$-gram positionnels,
- Expectative Mutuelle,
- Normalisation des mesures d'association binaire,
- Algorithme GenLocalMaxs.

propose une solution à la définition de valeurs seuil, aux méthodes d'amorçage ainsi qu'aux modèles mathématiques sensibles aux fragments textuels fréquents. A partir de cet ensemble de nouvelles méthodologies, notre analyseur probabiliste sera capable d'extraire pour n'importe quelle langue et sans intervention extérieure, un ensemble d'unités lexicales complexes contiguës et non contiguës prouvant ainsi sa totale flexibilité. La seconde partie de ce rapport démontrera en particulier tout le bien fondé de cette nouvelle approche comparativement aux méthodes existentes lors d'une évaluation par comparaison.

## 1.5 Plan

Afin de rendre compte de tous les aspects importants de notre étude, nous avons divisé ce rapport de recherche en quatre parties principales : une partie dans laquelle nous abordons les notions de base inhérentes à l'extraction d'unités lexicales complexes, une seconde partie où nous définissons l'architecture du système, une troisième partie dans laquelle nous évaluons les résultats obtenus à partir de collections de textes en Français et en Portugais et finalement une partie de conclusion dans laquelle nous introduirons

nos perspectives de travail futur.

Dans la première partie de ce rapport, nous aborderons deux points principaux. Premièrement, nous analyserons les travaux qui ont été proposés suivant les trois axes de recherche que constituent les méthodes syntaxiques, hybrides et numériques. Cette analyse servira de motivation aux solutions innovatrices que nous nous proposons d'introduire. Dans un deuxième temps, nous essayerons de donner une définition consistante aux phénomènes de figement qui ont souvent été obscurcis par des dénominations floues et hétérogènes. Dans ce cadre, nous accorderons une place égale à chacun des deux courants linguistiques que sont la linguistique classique et la linguistique de corpus.

Dans la deuxième partie, nous donnerons une large place à la formalisation de notre méthodologie. Dans ce cadre, nous introduirons, dans un premier temps, la notion de modèles $N$-gram positionnels qui nous permettra de mettre en évidence la structure des unités polylexicales. Plus particulièrement, nous parlerons de pré-traitement des textes, de segmentation en $N$-grams positionnels, de complexité et de codification des données. Dans le deuxième chapitre, nous définirons formellement la mesure d'Expectative Mutuelle à partir de la notion d'Expectative Normalisée. Ainsi, nous serons en mesure d'attribuer une valeur de cohésion à chacun des $N$-gram positionnels construits préalablement — $\forall N, N \geq 2$. Dans ce même chapitre, nous présenterons une méthode originale qui nous permettra d'identifier les unités lexicales complexes de l'ensemble des $N$-grams positionnels associés à leur mesure d'association sans recourir à la définition de valeurs seuil. Ceci sera l'objectif d'un algorithme d'analyse de maxima locaux, le GenLocalMaxs. Finalement, dans un dernier chapitre, afin de préparer la dernière partie de ce rapport, nous proposerons une méthodologie définissant cinq mesures d'association binaires de forme $N$-aire. Celle-ci nous permettra en particulier de proposer une évaluation uniformisée des résultats à partir d'une seule plateforme utilisant l'algorithme GenLocalMaxs pour sélectionner les unités polylexicales.

Dans la troisième partie de ce rapport, nous aborderons la phase d'évaluation des résultats. Dans un premier temps, nous résumerons les différentes méthodologies d'évaluation existentes pour proposer notre propre idée de l'évaluation dans le cadre de nos recherches sur les phénomènes de figement : l'évaluation par comparaison. Dans un deuxième temps,

nous proposerons une analyse qualitative des résultats obtenus à partir du système tel que nous l'avons développé et nous comparerons les ensembles d'unités polylexicales extraits par chaque mesure d'association normalisées sur la base du GenLocalMaxs. Dans le chapitre suivant, nous aborderons une évaluation quantitative concernant la longueur des unités polylexicales extraites ainsi que leur fréquence à partir des résultats obtenus avec l'Expectative Mutuelle comparés avec ceux du coefficient d'association [27], du coefficient Dice [45], de la probabilité conditionnelle symétrique [30], du test $\Phi^2$ [29] et du coefficient de vraisemblance Loglike [28] préalablement normalisés. Bien entendu, nous évaluerons également les taux de précision et de couverture de notre extracteur dans ce même chapitre pour chaque mesure d'association testée.

En guise de conclusion et de perspectives de travail futur, nous présenterons une série d'expériences réalisées à partir de collections de textes de différents types. Dans ce cadre, un texte pourra être considéré comme une séquence de caractères ou bien comme un ensemble d'étiquettes morpho-syntaxiques si l'on dispose de celui-ci en forme annotée. Cette analyse aura comme objectif de montrer l'intérêt de la définition de systèmes d'extraction aussi flexibles que possible. En particulier, on notera qu'il est envisageable d'extraire morphèmes, congnats et patrons syntaxiques à partir d'une méthodologie initialement définie pour l'étude des formes graphiques. Finalement, nous proposerons une nouvelle mesure d'association qui intègre de forme judicieuse les connaisssances linguistiques et lexicales dans une même architecture permettant d'augurer un avenir prometteur des méthodes statistiques ■

# Chapitre 2

# Etude Bibliographique

Les systèmes d'extraction d'associations lexicales ont traditionnellement été développés suivant trois axes bien définis. Chronologiquement, les méthodes numériques ont été les premières à proposer une solution. Dans ce cadre, la notion de figement est évaluée à partir de mesures d'association dont le rôle est de mettre en valeur les cooccurrences les plus probables. Ainsi, les séquences démontrant une forte valeur de cohésion sont identifiées comme pertinentes. Afin de faire face au nombre important de phénomènes linguistiques extraits par les méthodes statistiques, un courant idéologique basé sur des méthodes hybrides s'est développé. Il propose d'allier mesures d'association et filtres linguistiques simples. Le but est alors de restreindre l'ensemble des associations lexicales possibles au sous-ensemble des unités lexicales complexes de type nominal. L'introduction de méta-connaissances diminue cependant son champ d'application et les études deviennent de plus en plus spécifiques des langues considérées. Cette tendance aboutit à la définition d'une troisième méthode d'extraction. Dans ce contexte, patrons syntaxiques et heuristiques langagières sont les seules informations nécessaires au repérage des noms composés présents dans les collections textuelles. Dans ce chapitre, nous dressons l'état actuel de la recherche dans chacune de ces trois catégories.

## 2.1 Approche Syntaxique

Peu de travaux ont été proposés concernant l'approche purement syntaxique. Ceci se doit particulièrement aux constats faits sur la rigidité lexicale des mots composés et le rejet de règles syntaxiques flexibles. Suivant cette approche, on soulignera cependant la réalisation de trois prototypes : LEXTER par D. Bourigault [15], TERMINO par S. David et P. Plante [16] et *Termight* par I. Dagan et K. Church [26]. Le plus complet

d'entre eux est certainement LEXTER.

LEXTER définit une méthode originale d'extraction pour le Français où les termes complexes sont identifiés à partir de patrons syntaxiques méticuleusement définis. D. Bourigault défend que la notion d'objet, inhérente aux termes complexes, induit des contraintes de type formel, d'ordre syntagmatique (composition synaptique) et paradigmatique (dérivation syntagmatique), qui permettent d'identifier les séquences d'unités lexicales fortement liées. Dans ce cadre, une analyse syntaxique locale par patron de surface découpe le texte en repérant des frontières potentielles entre lesquelles il est possible d'isoler des syntagmes nominaux. Ainsi, les verbes, les pronoms, les conjonctions et les séquences préposition-article indéfini constituent des frontières qui permettent de détacher "en négatif" les unités lexicales complexes de type nominal. D. Bourigault défend ainsi l'idée que ces patrons de coupe ne sont jamais constituants des termes complexes[1]. L'idéal est donc d'identifier le plus grand nombre possible de schémas de ce type de façon à procéder à un découpage sévère qui conduise à serrer au plus près les termes complexes d'un texte. Cependant, cette sévérité doit être relative, et un certain nombre de schémas morpho-syntaxiques appellent un examen poussé. Considérons par exemple le schéma constitué par la préposition *sur* et l'article défini *le*. L'analyse d'un grand nombre de configurations de ce type sur différents corpora montre que le plus souvent celles-ci marquent des frontières entre groupes nominaux comme dans l'exemple suivant extrait de [15].

*on raccorde le câble d'alimentation sur le coffret de décharge de batterie*

Il faut néanmoins résister à la tentation de considérer toutes ces configurations comme des marqueurs de frontière. Il existe en effet une proportion suffisante de cas où de telles configurations sont parties intégrantes des termes comme dans les deux exemples suivants énoncés par D. Bourigault dans [15].

*action sur le bouton poussoir de réarmement*
*action sur le système d'alimentation de secours*

Pour être en mesure d'effectuer un découpage satisfaisant, le système doit donc disposer d'une liste de noms susceptibles de se construire avec un complément introduit par la

[1] Cette hypothèse n'est cependant pas évidente comme le montreront les prochains exemples.

préposition *sur*, c'est-à-dire un ensemble d'informations de sous-catégorisation. Dans ce cadre, D. Bourigault propose une procédure d'apprentissage endogène où le corpus est à la fois objet du traitement et source d'information. Une fois isolés les groupes nominaux maximaux, un module de décomposition selon lequel tout terme est composé d'une tête $T$ et d'une expansion $E$ finalise l'analyse. Par exemple, une règle de décomposition est définie dans la figure 2.1 pour la séquence *Nom+Adjectif*. Dans ses publications, D. Bourigault n'illustre que très rarement les résultats d'extraction de son logiciel. Il est donc difficile de juger sa réelle performance. Il semble cependant que ceux-ci soient disparates. En effet, ils doivent révéler un taux de bruit important dû à la complexité des phrases qui font l'objet de nombreuses combinaisons syntaxiques : propositions relatives, conjonctions de coordination etc. D'autre part, LEXTER n'autorise que l'extraction de termes complexes qui dérivent de la composition nominale réduisant drastiquement le taux de couverture du processus d'acquisition. Finalement, LEXTER recourt à un grand nombre de contraintes linguistiques par le biais d'heuristiques qui limitent son champ d'application au Français. Entre autres, LEXTER utilise les catégories grammaticales, les traits morphologiques (en particulier genre et nombre) et les formes lemmatisées des formes du texte pour définir le plus précisément les règles de formation des termes complexes.

$$Nom + Adjectif \rightarrow \begin{array}{l} T : Nom \\ E : Adjectif \end{array}$$

FIG. 2.1 – Règle de décomposition

Du fait de leurs similitudes évidentes avec LEXTER[2], nous ne donnerons que l'essence des deux autres systèmes TERMINO et *Termight*. D'une part, S. David et P. Plante ont développé le logiciel TERMINO autour de l'idée directrice qu'il est nécessaire d'analyser en profondeur les phénomèmes syntaxiques des unités polylexicales afin de les repérer le plus précisément possible. Pour se faire, TERMINO [16] dispose d'un analyseur morphologique qui lemmatise et étiquette les formes des textes. Une fois le texte enrichi linguistiquement, une grammaire locale repère les syntagmes nominaux qui sont ensuite étudiés selon un ensemble d'heuristiques afin de déterminer s'ils forment ou non des

[2]Et de notre point de vue, leur plus faible importance.

synapsies. Ainsi, trois opérations principales peuvent être mises en évidence : 1) analyse de l'ensemble de leurs catégories syntaxiques, 2) étude de la position de leur tête et 3) vérification de leur contenu à partir d'une liste de mots vides. Les groupes nominaux satisfaisant ces trois contraintes sont automatiquement identifiés comme pertinents par TERMINO. D'autre part, I. Dagan et K. Church proposent une étude similaire dans le contexte des terminologies multilingues. Dans ce cadre, ils proposent le logiciel *Termight* [26] qui repère les unités lexicales complexes à partir de leurs régularités syntaxiques. Ainsi, dans un premier temps, *Termight* repère les séquences de mots qui correspondent aux patrons syntaxiques préalablement définis à l'aide d'expressions régulières. Les groupes nominaux ainsi retenus sont ensuite classés par ordre décroissant de fréquence et de fonction de mot de tête. Finalement, les termes sont validés manuellement — i.e. par l'utilisateur — à l'aide d'un concordanceur.

Le caractère purement syntaxique de ces trois logiciels impose un certain nombre de contraintes qui sont communes aux trois systèmes présentés : extraction exclusive de termes complexes du type nominal, dépendance intrinsèque à la langue étudiée et absence de critère de mesure du figement[3]. Afin d'illustrer les limitations de ces systèmes, nous présentons un ensemble d'unités lexicales complexes que les contraintes imposées par les trois architectures ne permettraient pas d'extraire :

- *l'offre et la demande* et *consommateurs d'alcool et de drogues* à cause de l'utilisation de la conjonction de coordination *et*,
- *prendre une cuillère à soupe* et *exporter une variable* pour être des locutions verbales.

Comme nous l'avons déjà mentionné, l'étude proposée par I. Dagan et K. Church met en avant l'utilisation du critère de fréquence pour l'identification des unités polylexicales. Cette particularité nous permet d'introduire avec élégance les travaux hybrides qui font appel à la fois aux critères syntaxiques et aux mesures numériques de figement. En effet, il est clair que l'apport numérique est fondamental pour le succès des extracteurs d'unités polylexicales.

[3]On pourrait cependant argumenter que l'utilisation de la fréquence dans le logiciel *Termight* dresse les bases de l'utilisation de telles mesures de cohésion. Cependant, cette fréquence n'est utilisée qu'à titre informatif et auncun traitement particulier ne lui est dédié. Ceci nous a motivé à ne pas la considérer comme une mesure de cohésion a part entière.

## 2.2 Approche Hybride

Les méthodes hybrides d'extraction proposent d'allier mesures d'association et filtres linguistiques simples pour l'identification des unités lexicales complexes. Dans ce cadre, deux méthodologies ont été proposées. De nombreux systèmes réduisent dans un premier temps l'espace de recherche de l'ensemble des associations lexicales à partir d'un pré-traitement linguistique des textes. Ainsi, des patrons syntaxiques sont utilisés pour identifier certaines séquences lexicales de prédilection — la plupart d'entre elles de type nominal. Des combinaisons de mesures d'association sont ensuite appliquées pour évaluer le degré de cohésion de chaque séquence préalablement retenue. C'est l'approche syntaxico-numérique. Même si cette méthode d'extraction prédomine dans la littérature, d'autres travaux ont été proposés. Ceux-ci mettent en avant un processus d'extraction inverse. Ainsi, l'espace de recherche est d'abord épuré par l'application de mesures numériques qui identifient les séquences de mots les plus figées. Ensuite, seules les suites de mots dont la structure s'accorde à un certain nombre de patrons syntaxiques définis de forme *ad hoc* sont sélectionnés — encore une fois, les associations lexicales de type nominal sont préférées. On appellera cette méthode, méthode numérico-syntaxique.

### 2.2.1 Syntaxico-Numérique

C. Enguehard [23] est certainement l'une des premières à proposer un système d'extraction hybride de termes complexes. Son système ANA n'utilise pas l'enrichissement morpho-syntaxique du corpus mais définit un certain nombre de contraintes que l'on dira linguistiques même si celles-ci sortent légèrement des barrières de la linguistique classique. Dans ce cadre, C. Enguehard propose un système qui se divise en deux modules : familiarisation et découverte. Dans une première phase, le module familiarisation extrait automatiquement par le biais de mesures statistiques ou manuellement quelques éléments de connaissance sur la langue et le domaine utilisés sous la forme de quatre listes : une liste de mots fonctionnels tels que *a, alors, après*, une liste de mots fortement liés comme *de la, et la, est le*, une liste de mots schémas tels que *de, des, du* et une liste de *bootstraps* qui sont des concepts du domaine comme par exemple *automate, centrale, chaudière* dans le cadre de l'énergie nucléaire. A partir de ces quatre listes, le module découverte repère un ensemble d'unités lexicales complexes potentielles. Dans ce cadre, C. Enguehard définit trois types d'événements qui forment les structures typiques des termes complexes.

- **Expression** : cooccurrence de deux concepts,
- **Candidat** : cooccurrence d'un concept et d'un mot séparés par un mot schéma,
- **Expansion** : cooccurrence d'un concept et d'un mot.

Une fois repérées, les unités lexicales complexes candidates sont considérées pertinentes si elles sont suffisamment figées et fréquentes. Cette fréquence est une valeur seuil établie en fonction du volume de texte traité. Cette technique a donc permis d'extraire un certain nombre de termes complexes du domaine de la commercialisation du miel : *analyse transactionnelle*, *création de marque*, *tradition de production* etc. C. Enguehard a également appliqué son système sur des corpora scientifiques écrits en Anglais. De cette forme, elle démontre que les méthodes hybrides sont plus facilement "transportables" que les méthodes purement syntaxiques. Les unités lexicales complexes suivantes ont ainsi pu être détectées : *signal strength*, *temperature coefficient of velocity*, *velocity of sound* etc.

Durant la même période, chez IBM, J. Justeson et S.M. Katz [22] se donnent pour objectif de construire une terminologie de termes techniques de différents domaines. Dans ce cadre, ils se trouvent évidemment confrontés à la surenchère de l'utilisation des termes complexes. Ils proposent donc une étude linguistique précise de l'ensemble des unités complexes représentatrices de leur domaine d'étude. Leurs travaux aboutissent aux résultats suivants : entre 90% et 99% des termes complexes utilisés dans les domaines médicaux et phisiologiques sont des syntagmes nominaux[4]. D'autre part, ils soulignent que les unités polylexicales retenues dépassent rarement quatre mots. Dans ce contexte, ils proposent un algorithme d'extraction basé sur une grammaire locale qui permet de repérer les syntagmes nominaux à considérer et sur une barrière de fréquence qui vérifie leur pertinence. Ainsi, à partir d'un texte annoté morpho-syntaxiquement, J. Justeson et S.M. Katz utilisent la grammaire suivante pour repérer les syntagmes répertoriés où $A$ est un adjectif, $N$ un nom et $P$ une préposition :

$$((A|N)^+|((A|N)^*(NP)^?)(A|N)^*)N$$

Une fois sélectionnées les séquences lexicales, leur fréquence d'occurrence sert d'identifiant de pertinence. Ainsi, plus les séquences sont fréquentes et plus elles peuvent être considérées pertinentes. Suivant ce simple algorithme, les deux auteurs extraient des

[4]Nous verrons que ces chiffres sont peu fiables et ne peuvent pas être élargis à d'autres langues. En particulier, M.L. Herviou-Picard [25] propose une étude similaire spécifique du Français.

termes tels que : *linear function*, *Gaussian random variable*, *degree of freedom* etc.

Suivant la même lignée, M.L. Herviou-Picard [25] propose une étude pour la création de terminologies pour le Français. Ses travaux s'inscrivent dans une politique générale de traitement massif de l'information textuelle au sein d'EDF : recherche documentaire, capitalisation du savoir-faire, gestion de la documentation technique et diffusion sélective d'informations. Dans ce cadre, elle propose une étude de la structure syntaxique des termes complexes présents dans le thesaurus EDF [46]. Les résultats montrent que 70% des termes complexes correspondent à des groupes nominaux simples — i.e. $NA$, $NPN$, $NPDN$ et $NPNA$ où $A$ est un adjectif, $N$ un nom, $P$ une préposition et $D$ un déterminant. Il reste donc 30% des unités polylexicales qui définissent des structures syntaxiques plus complexes. Nous sommes donc loin des 90% proposés par J. Justeson et S.M. Katz dans le cas de l'Anglais. A partir de ce constat, une grammaire procède à une analyse ascendante du corpus et extrait automatiquement des groupes nominaux qui sont potentiellement des termes. Dans un deuxième temps, une série de mesures statistiques proposent d'évaluer leur degré d'attraction. Ce sont l'effectif, la variance et la densité locale. En particulier, l'effectif correspond au nombre d'occurrences d'un candidat terme dans le corpus. La variance mesure l'hétérogénéité d'une expression dans les documents d'un corpus, signe de sa nature terminologique. Et finalement, la densité locale évalue si les documents utilisant un même terme ont des contenus similaires. Ainsi, M.L. Herviou-Picard défend que ces trois indicateurs doivent être utilisés conjointement afin de rendre compte de l'ensemble des phénomènes associatifs.

Il est clair que la fréquence d'occurrence proposée par J. Justeson et S.M. Katz est insuffisante pour déterminer la pertinence des groupes nominaux. Dans ce cadre, les travaux de M.L. Herviou-Picard montrent une avancée significative dans le cadre de l'extraction terminologique. Cependant, d'autres études ont été proposées faisant appel à des mesures statistiques encore plus élaborées. Dans ce cadre, B. Daille [21] réalise une étude rigoureuse des phénomènes de figement et propose une méthodologie originale qui lui permet de comparer plusieurs mesures d'association. L'étude des caractéristiques linguistiques des termes complexes l'amène à définir deux catégories principales. Ainsi, les termes de base — i.e termes de longueur deux — sont de loin les plus nombreux et s'apparient à l'une des structures morpho-syntaxiques suivantes : $NA$, $NN$, $Nà(D)N$,

$Nde(D)N$ et $NPN$[5]. Parallèlement, les termes de longueur supérieure à deux unités sont construits récursivement à partir des termes de base suivant trois opérations : la surcomposition, la modification et la coordination. Cependant, face à ce constat, B. Daille propose d'extraire exclusivement l'ensemble des termes de base des collections textuelles laissant pour travail futur l'analyse des autres termes. Dans ce contexte, un automate d'états finis repère les séquences textuelles définies par les patrons syntaxiques préalablement décrits par un ensemble d'expressions régulières. Ces séquences sont ensuite épurées de forme à ne garder que les deux unités principales d'un terme donné. Ainsi, la séquence *ligne d'abonné* serait transformée en un couple de mots *(ligne, abonné)*. Finalement, le coefficient de vraisemblance Loglike [28] calcule le degré d'attraction de chaque couple d'unités textuelles. Le choix de cette mesure d'association est le résultat d'une évaluation comparative réalisée à partir du coefficient d'association[6] [27] et du test $\Phi^2$ [29]. En particulier, cette étude révèle le caractère démesurément spécifique du coefficient d'association qui sur-évalue les associations spécifiques de la langue au détriment des associations terminologiques. Parallèlement, B. Daille souligne la pertinence de l'utilisation de la fréquence d'occurrence comme filtre statistique.

Afin de combler la faible couverture des travaux précédents, D. Evans [47] propose une méthode d'extraction innovatrice qui permet l'acquisition de termes complexes de dimension supérieure à deux. Dans ce cadre, il fait usage de la méthode d'acquisition par amorçage. Ainsi, la première étape du processus utilise les outils linguistiques du système CLARIT [48] pour reconnaître les groupes nominaux présents dans le corpus. Un second traitement linguistique permet ensuite de repérer quatre groupes de syntagmes nominaux simples — séquences nominales qui dénotent un concept. Par exemple, à partir du groupe nominal *the quality of surface of treated stainless steel strip*[7], l'atome lexical *stainless steel*, le couple modificateur *treated strip*, le sous-composé *stainless steel strip* et la paire modificatrice extra-prépositionnelle *surface quality* devraient être identifiés. Dans cette première phase d'étude, D. Evans s'attarde sur le cas des atomes lexicaux. Dans ce cadre, les séquences de deux mots préalablement extraites sont testées statistiquement

[5] On remarquera qu'il n'existe pas un consensus clair sur la définition syntaxique des termes complexes de la part de tous les auteurs. Nous reprendrons plus en détail cette remarque dans le prochain chapitre sur les définitions des unités lexicales complexes.

[6] Traduction du terme *association ratio*.

[7] Traduction : "*la qualité de la surface des planches d'acier-inox traitées*".

afin d'évaluer leur pertinence comme terme complexe. La première heuristique revient à calculer la fréquence d'occurrence de chaque groupe nominal et de s'assurer qu'aucune paire, continue ou non, contenant l'une des deux unités textuelles initiales n'ait une fréquence supérieure. Dans un deuxième temps, les couples retenus servent d'amorce pour l'identification d'atomes lexicaux étendus c'est-à-dire contenant plus de deux unités textuelles. Ainsi, ils sont ré-intégrés dans le corpus et traités comme étant des mots simples dans la suite du processus. Le score d'association[8] est alors appliqué pour évaluer la pertinence des nouvelles associations.

Durant ces dernières années, la recherche du courant syntaxico-numérique n'a pas connu de révolution particulière. L'état de l'art s'est stabilisé autour de résultats satisfaisants pour l'acquisition de terminologies. Les travaux se tournent désormais vers l'amélioration constante des taux de précision et de couverture — rappel. Dans ce cadre, on remarquera les études de R. Feldman [24] qui propose de tester la pertinence des termes complexes à partir de l'analyse de leurs distributions dans les grandes collections textuelles. On soulignera également les travaux de U. Heid [49] qui propose une étude approfondie des phénomènes de figement pour l'Allemand, langue particulièrement problématique du fait de sa composition morphologique. Finalement, B. Daille [50] propose une étude intéressante pour le Français dans laquelle elle met en évidence l'importance de l'identification des adjectifs relationnels pour l'extraction terminologique[9].

Cependant, quelques remarques s'imposent sur cette approche. Il est clair que les résultats obtenus sont raisonnablement satisfaisants pour le domaine spécifique de la construction de terminologies. Cependant, l'ensemble des phénomènes de figement que couvrent ces systèmes sont particulièrement réduits. En effet, seuls les termes complexes de type nominal simple sont retenus lors de la phase d'extraction. Ainsi, il est peu probable que les unités lexicales complexes suivantes soient identifiées : "*Liberté, Egalité, Fraternité*", *monter une partition*, *transmission via FTP anonyme*, *agents thérapeutiques dérivés de ressources naturelles*, *World Wide Web*. Parallèlement, leur application à de nouvelles langues nécessite la redéfinition des filtres syntaxiques ce qui réduit leur champ d'action. Ils sont donc peu flexibles et difficiles à évaluer. Plus regrettable encore, c'est leur absence de solution au problème de la définition de valeurs seuil et de mesures statistiques binaires.

[8] Traduction du terme *association score*.

[9] Par exemple, l'unité polylexicale *conquête de l'espace* peut se trouver sous la forme de *conquête spatiale*.

### 2.2.2 Numérico-Syntaxique

La seconde approche hybride définit dans un premier temps un ensemble de séquences d'unités lexicales fortement liées grâce à l'application de mesures d'association. Dans une seconde phase, les séquences retenues sont épurées à l'aide de filtres linguistiques. Historiquement, cette approche n'a pas connu un grand engouement bien que l'ensemble des résultats obtenus soient particulièrement intéressants, notamment au niveau de leur diversité. F. Smadja [31] est l'un des rares chercheurs à proposer cette méthodologie. Dans ce cadre, il propose le système XTRACT. Dans un premier temps, XTRACT extrait l'ensemble des associations binaires dont le score centré réduit[10] est supérieur à une valeur seuil déterminée par l'utilisateur. Dans un deuxième temps, XTRACT examine le contexte immédiat de chaque association binaire retenue et analyse les distributions des mots cooccurrents suivant leurs positions. Cette analyse permet de ne sélectionner que les mots dont la position est raisonnablement fixe par rapport au couple considéré et de fabriquer des séquences de mots de dimension supérieure à deux unités lexicales. Une fois extraites toutes ces suites complexes, F. Smadja propose l'introduction d'informations fonctionnelles afin d'enrichir les structures acquises ou bien de simplement les rejeter. En effet, l'information basée exclusivement sur l'analyse des formes graphiques n'est pas suffisante pour mettre en évidence un certain nombre de relations syntaxiques, sémantiques ou autres importantes. Par exemple, si *make a decision* est retenue comme unité lexicale complexe, on ne sait pas que *decision* est le complément d'objet direct du verbe *make*. Dans ce cadre, F. Smadja analyse syntaxiquement les unités polylexicales candidates à partir d'une grammaire développée par S. Abney [19] pour le repérage de relations syntaxiques telles que *verbe-objet*, *sujet-verbe*, *nom-adjectif*, *nom-nom*. Ainsi, toutes les phrases contenant les unités polylexicales candidates sont tour à tour analysées et enrichies ou bien rejetées. Les résultats obtenus sont impressionnants. Alors que la première phase de traitement atteint 40% de taux de précision, l'introduction d'information linguistique catapulte ce résultat au niveau des 80%. Il est évident que ces chiffres doivent être interprétés avec précaution. En effet, on verra dans la troisième partie de notre rapport que la précision des résultats dépend forcément de l'application considérée. XTRACT se distingue donc des autres méthodes par sa flexibilité d'utilisation. En effet, son adaptation à d'autres langues ou domaines est facilitée par l'extraction d'une liste de termes candidats qui peut guider *a posteriori* la définition des filtres syntaxiques. Cependant, XTRACT montre deux

[10] Le score centré réduit est la traduction française du z-score utilisé par les anglo-saxons.

problèmes majeurs qui sont l'utilisation des méthodes d'amorçage et la définition de valeurs seuil globales. En effet, les associations binaires retenues pour la deuxième étape du système sont choisies sur la base d'une valeur seuil globale qui ne peut être définie qu'à partir d'expériences et qui, par conséquent, est sujette à un taux d'erreur non négligeable. D'autre part, le résultat des méthodes d'amorçage dépend des associations binaires retenues lors de la première étape du processus d'acquisition. Par exemple, l'identification du terme complexe *Traité de Maastricht* dépend de l'extraction préalable de l'association binaire *Traité de*. Or, le degré d'association entre *Traité* et *de* est généralement sous-évalué par les mesures statistiques du fait de la forte fréquence de la préposition *de* dans l'ensemble du corpus. Ainsi, l'unité complexe *Traité de Maastricht* ne serait probablement pas identifiée du fait de la non sélection du digram *Traité de* durant l'étape initiale. Comparativement au Français et au Portugais, il est à noter que ce problème est moins contraignant pour les langues anglo-saxonnes qui font fort usage de la composition nominale par juxtaposition. Par exemple, *Traité de Maastricht* serait traduit en anglais par l'expression *Maastricht Treaty* dans laquelle aucun fragment fonctionnel n'apparaît.

## 2.3 Approche Numérique

Comme nous l'avons vu, les approches précédentes restent particulièrement dépendantes de la langue considérée lors de l'analyse et démontrent certains inconvénients majeurs tels que l'extraction exclusive de termes complexes de type nominal — laissant ainsi sans réponse un certain nombre de phénomènes de figement importants — ainsi que la définition de valeurs seuil globales et l'utilisation de mesures d'association binaires. Afin de répondre à nos impératifs de flexibilité, il convient donc d'aborder en détail l'approche purement numérique. L'approche numérique s'est rapidement imposée comme la démarche prépondérante dans le cadre de l'acquisition automatique des mots composés. Ceci se doit en grande partie aux études linguistiques [51] [40] [52] [3] [53] [3] [41] [54] [55] qui soulignent les contraintes de figement telles que le blocage des paradigmes synonymiques ou celui des propriétés transformationnelles que subissent les unités polylexicales. Dans ce cadre, un certain nombre de chercheurs défendent que les unités lexicales complexes sont des suites de mots suffisamment figées pour qu'elles soient identifiées à partir de l'analyse de leurs régularités associatives. Ainsi, l'enrichissement linguistique n'est pas considéré nécessaire et la seule information générale présente dans les textes doit suffire au repérage des associations lexicales. La communauté scientifique s'est donc définie comme objectif

principal de définir des mesures d'association capables de mettre en évidence les attirances entre mots. Suivant ce courant, deux approches ont été privilégiées. D'une part, un certain nombre de recherches proposent la définition de mesures d'association binaires qui font appel aux méthodes d'amorçage pour l'identification d'associations de plus de deux mots et d'autre part, un certain nombre d'études proposent la définition de mesures d'association $N$-aires qui évitent les traitements par amorce.

### 2.3.1 Mesures d'association binaires

D. Labbé [56], P. Lafon et M. Tournier [57] sont les premiers à proposer des mesures numériques qui permettent d'extraire des mots composés à partir de l'analyse exclusive des formes graphiques présentes dans les textes[11]. Cependant, le coefficient d'association de K.W Church et P. Hanks [27] est le plus souvent cité comme référence dans la littérature internationale. Afin de ne pas rompre avec cette tradition dans ce paragraphe introductif, nous commencerons par l'analyse de cette mesure de la théorie de l'information. K.W Church et P. Hanks fondent leur analyse sur des résultats de la psycholinguistique. Dans ce cadre, Meyer [58] publie une expérience dans laquelle il mesure le temps de réaction d'un individu lorsqu'il est confronté à deux tâches spécifiques : 1) classer des successions de lettres en mots et non mots, 2) prononcer une suite de caractères. Ainsi, il montre dans les deux cas que la réponse à un mot — e.g. beurre — est constamment plus rapide lorsqu'il est précédé par un mot associé — e.g. pain — que lorsqu'il ne l'est pas — e.g. infirmière. Pour les deux auteurs, cette association doit être mesurée. En effet, tous les mots contiennent un certain nombre d'informations sur ceux qui les entourent. En particulier, dans le cas des mots composés, cette information est très importante. Prenons l'exemple du domaine politique. L'adjectif *Premier* impose pratiquement l'occurrence du nom commun *Ministre*. Suivant ces constatations, K.W. Church et P. Hanks proposent d'utiliser une variante de l'information mutuelle — mesure de la théorie d'information — appelée coefficient d'association dont l'objectif est de calculer le taux d'information que deux mots partagent l'un par rapport à l'autre. Pour se faire, le texte est d'abord découpé en groupes de deux mots qui ne sont pas forcément contigus. Ainsi, pour un mot pivot donné, un ensemble de mots qui ont tendance à se trouver dans son voisinage sont sélectionnés pour former une paire d'occurrence avec le mot pivot. Dans ce cadre, il faut évidemment commencer par définir le voisinage — unité de contexte — à l'intérieur

[11] Du moins dans le cadre des mesures d'association binaires hors fréquence d'occurrence.

duquel on considèrera que deux formes sont cooccurrentes. Cette unité peut ressembler à la phrase ou encore être constituée par un contexte de longueur fixe. K.W. Church et P. Hanks utilisent une fenêtre de cinq mots à droite — i.e. en avant — de l'unité pivot. Le coefficient d'association est alors appliqué à chaque couple et met en évidence le degré d'association qui lie entre elles les deux unités qui composent la paire d'unités textuelles. Finalement, les deux auteurs définissent une valeur seuil égale à trois selon laquelle un couple est une unité polylexicale ou non. Ainsi, si un couple met en évidence une valeur de coefficient d'association supérieure à trois, il est considéré pertinent. Dans le cas contraire, le couple est rejeté. Les résultats obtenus sont encourageants. Ils rendent manifestes un certain nombre de phénomèmes de figement tels que les verbes composés — *set up*, *save from*, les noms composés — *computer scientist*, *United States* et les phrases idiomatiques — *bread and butter*, *drink and drive*. Dans le cadre de leurs études, K.W. Church et P. Hanks font également ressortir des relations sémantiques entre mots — *man woman*, *doctor nurse* — ainsi que des relations lexicales — *coming from*, *keeping from*. Cette méthodologie ne dépend donc ni de la langue ni du matériel textuel utilisé. Elle est par conséquent totalement flexible. Cependant, elle dénote certaines insuffissances. Dans une premier temps, seules les associations lexicales binaires sont extraites réduisant ainsi la couverture du processus d'acquisition. Deuxièmement, l'utilisation de valeurs seuil n'est pas satisfaisante comme nous le détaillerons dans le prochain paragraphe. En effet, ces valeurs seuil ne définissent pas ce que sont les associations lexicales mais plutôt définissent une liste d'associations ordonnée par degré de pertinence. Finalement, plusieurs études [43] [17] ont montré que le coefficient d'association tend à donner une importance prépondérante aux associations lexicales peu fréquentes dans les textes[12]. Or, il est clair que la fréquence joue un rôle important pour le repérage des unités polylexicales comme le remarque justement B. Daille.

Dans le but de résoudre les insuffisances caractéristiques du coefficient d'association, d'autres mesures d'association ont été proposées. En particulier, certains chercheurs se sont donnés comme objectif d'adapter des mesures statistiques connues pour le cas spécifique du matériel textuel. Dans ce cadre, W. Gale [29] propose une mesure statistique supportée par la théorie du test d'hypothèse pour les séries statistiques bivariées : le test $\Phi^2$. Son but est de mesurer les liens qui unissent deux mots à partir

[12]Nous le confirmerons également dans ce rapport.

de l'analyse d'un tableau de contingence. Prenons l'exemple du couple *New York*. Un tableau de contingence pourrait être construit de la forme suivante 5.1 où $k$ est la fonction de fréquence d'occurrence et $\neg$ correspond à la non apparition du terme qui lui est associé.

| | $York$ | $\neg York$ |
|---|---|---|
| $New$ | $k(NewYork)$ | $k(New\neg York)$ |
| $\neg New$ | $k(\neg NewYork)$ | $k(\neg New\neg York)$ |

TAB. 2.1 – Tableau de Contingence pour *New York*

A partir de ce tableau de contingence, il suffit d'appliquer les tests d'indépendance définis par la théorie statistique pour identifier les paires de mots qui s'attirent. W. Gale définit ainsi l'hypothèse d'indépendance qui consiste à considérer deux mots indépendants si leur probabilité de cooccurrence est égale au produit de leurs probabilités marginales. Le but est alors de montrer que si l'hypothèse n'est pas valide, les deux mots forment un couple cohérent.

Dans la même lignée, T. Dunning [28] attaque la supposition de normalité qui est sous-jacente à un grand nombre de tests statistiques dont le $\Phi^2$. En effet, les tests tels que le score centré réduit ou le $\chi^2$ qui se basent sur des observations de grande taille, trébuchent face à la réalité du matériel textuel qui se distingue par la prédominance d'événements peu fréquents. Ainsi, l'idée qui consiste à estimer que les résultats des observations effectuées sur les textes suivent une distribution normale — ou approximativement normale — est caduque pour des énoncés de taille réduite. Dans ce cadre, T. Dunning utilise le test de vraisemblance maximum qui permet d'analyser avec succès les tableaux de contingence dont les comptages ne sont pas forcément élevés. Les résultats obtenus à partir d'un corpus de 31 777 mots[13] mettent en évidence l'extraction de différents types d'associations lexicales binaires — *mineral water*, *health insurance*, *can be*, *continue to*, *great deal*, etc.

Afin de résoudre le problème de l'acquisition exclusive d'associations binaires révélé par les travaux précédents, S. Shimohata [32] propose d'utiliser la méthode d'amorçage pour

[13]La taille d'un tel corpus peut être considérée faible comparativement aux corpora de 100 millions de mots qui sont maintenant courants.

identifier les cooccurrences de plus de deux caractères[14]. Pour ce faire, il utilise la mesure d'entropie [59]. Son objectif est de mesurer la quantité d'information contenue dans une séquence de caractères sur l'occurrence du prochain caractère et du caractère précédent. Ainsi, S. Shimohata propose de mesurer l'entropie de l'ensemble des caractères qui se trouvent immédiatement *devant* et immédiatement *après* une séquence de caractères notée *str* et de ne retenir comme cooccurrents que les caractères démontrant une faible valeur d'entropie c'est-à-dire une forte cohésion. Cette décision est opérée par le biais d'une valeur seuil définie de façon *ad hoc*. Les caractères retenus sont alors concaténés à la chaîne *str* et le processus se répète jusqu'à ce que plus aucun caractère ne soit identifié. Finalement, deux mesures de probabilité classent les unités polylexicales selon leur degré de pertinence. Ainsi, son système permet d'extraire, pour l'Anglais, un certain nombre d'associations lexicales complexes qui mettent en évidence la diversité des phénomènes de figement dans les textes — "*the current functional area*", "*All Rights are reserved.*", "*such as*", "*,for example,*". On remarquera que l'utilisation de tous les caractères, virgules et points confondus, permet l'identification de patrons lexicaux intéressants.

Malheureusement, ces mesures d'association ne sont pas parfaites. Loin s'en faut. En particulier, elles ne permettent pas l'extraction d'associations de plus de deux unités textuelles sans recourir aux méthodes d'amorçage qui comme nous l'avons vu précédemment posent un certain nombre de problèmes. Dans ce cadre, plusieurs chercheurs se sont donnés comme objectif de définir des mesures d'association $N$-aires.

### 2.3.2 Mesures d'association $N$-aires

Les premières études tentant de mettre en évidence les suites de mots pertinentes sans recourir aux méthodes d'amorçage imposées par les mesures d'association binaires peuvent être attribuées à A. Salem [34]. A. Salem propose de repérer toutes les séquences de mots qui se répètent dans les énoncés. C'est pourquoi, il suggère une analyse détaillée de chaque segment répété par le biais des tableaux des segments répétés — $TSR$ — et de leurs inventaires alphabétiques, hiérarchiques et distributionnels. Seulement, son travail doit être considéré plus comme une analyse des séquences de mots que comme la définition d'une méthode d'extraction.

[14] La langue Japonaise implique l'utilisation des caractères comme unités de base des traitements.

Ce n'est que huit ans plus tard qu'une méthode alternative aux segments répétés est proposée. Dans ce contexte, K. Frantzi et S. Ananiadou [33] définissent une mesure numérique appelée *C-value* qui permet de calculer la pertinence d'une séquence illimitée de mots[15]. Cette mesure se base sur trois concepts fondamentaux : la fréquence d'occurrence de chaque segment, la fréquence d'occurrence à l'intérieur de segments plus longs et le nombre de ces segments "longs". L'utilisation de la fréquence d'occurrence comme facteur de pertinence est claire. En effet, les termes complexes tendent à apparaître fréquemment dans les textes. Cependant, seule la fréquence d'occurrence ne permet pas de garantir le caractère terminologique de ces séquences. Considérons par exemple les quatre segments du tableau 2.2 où *soft contact lenses* peut être traduit par *lentilles de contact souples*.

*soft contact lenses*

*hard contact lenses*

*contact lenses*

*soft contact*

TAB. 2.2 – Exemples

Il est clair que tous ces segments ne sont pas des termes complexes. Cependant, l'utilisation exclusive de la fréquence comme critère de pertinence ne permet pas de différencier indubitablement les unités polylexicales correctes des autres. En effet, supposons que les deux premiers segments dépassent un certain seuil de fréquence permettant de les juger pertinents. Dans ce cas, les deux autres segments devraient être repérés *a fortiori*. En effet, leur fréquence est au pire égale à celle des deux premières associations. Or, cette situation n'est pas souhaitable. Le problème surgit du fait que *contact lenses* devrait être extrait mais pas *soft contact*. La différence entre ces deux segments peut être formulée par le fait que le premier *contact lenses* peut apparaître par lui-même dans le texte alors que le second *soft contact* apparaît nécessairement à l'intérieur du segment plus long *soft contact lenses*. Plus encore, plus *contact lenses* apparaîtra dans différents segments "longs" plus il sera pertinent. Il est donc nécesssaire d'introduire les deux concepts définis précédemment pour éviter les situations mises en évidence par l'utilisation exclusive de la fréquence. C'est l'objectif de la *C-value*. Une fois évalué chacun des segments, K. Frantzi et S. Ananiadou présentent un algorithme qui extrait par défaut l'ensemble des unités

[15]Du moins théoriquement.

polylexicales. Ainsi, un segment n'est pas un terme complexe s'il dénote une fréquence égale à celle d'un segment plus long qui le contient. Les autres segments sont ensuite classés suivant leur valeur de *C-value*. L'idée de K. Frantzi et S. Ananiadou est donc comparable à celle du repérage des segments répétés proposée par A. Salem à la différence qu'elle utilise les comptages des sous-segments — quasi-segments — pour identifier les segments les plus pertinents. L'un des inconvénients majeurs de cette méthodologie réside inlassablement dans l'utilisation des valeurs seuil comme élément de décision pour le processus d'acquisition. Les résultats énoncés le prouvent. Suivant la valeur du seuil, la précision des résultats oscille entre 82% et 45% !

Un certain nombre de travaux ont suivi les pas de K. Frantzi et S. Ananiadou. Parmi ceux-ci, on citera les travaux de R. Schneider [60] qui propose une nouvelle mesure d'association $N$-aire. Celle-ci permet d'évaluer le degré de cohésion de n'importe quel segment grâce à "l'importance relative" de chaque mot du corpus, calculée à partir de la fréquence d'occurrence des mots et de leurs variantes orthographiques. Suivant la même idée, G. Chartron [61] propose de repérer les cooccurrences de plusieurs termes à partir du coefficient d'implication réciproque. Celui-ci calcule le rapport entre la fréquence d'apparition d'une séquence de $N$ unités textuelles et le produit des fréquences marginales de chacun des mots constituant la séquence. Malheureusement, dans les deux cas, la méthode des valeurs seuil n'est pas évitée. Suivant un autre point de vue, P.K. Kim [62] propose une nouvelle mesure basée sur la normalisation de l'Information Mutuelle dans le cadre des segments contenant trois mots. On regrettera cependant qu'aucune méthode de généralisation exhaustive ne soit avancée. Plus récemment, D. Maynard et S. Ananiadou [63] proposent une évolution de la *C-value* qui prend en compte l'information contextuelle des segments répétés. Cette mesure s'appelle la *SNC-value* et inclut le facteur de contexte[16] *CF*. Mais, là encore, les valeurs seuil ne sont pas évitées.

## 2.4 Conclusion

Comme nous l'avons vu, plusieurs problèmes restent à résoudre quelle que soit l'approche utilisée pour l'identification des unités lexicales complexes. Selon les impératifs initiaux que nous nous sommes fixés, il est évident que l'approche statistique est la seule qui

[16] Traduction du terme *Context Factor*.

va à l'encontre de la totale flexibilité recherchée. Dans ce cadre, il est clair que la définition de mesures d'association $N$-aires est un impératif ainsi que l'élaboration de méthodes d'extraction qui ne dépendent pas de valeurs seuil globales définies de forme *ad hoc*. Ainsi, nous introduirons respectivement l'Expectative Mutuelle ainsi qu'une méthodologie de normalisation des mesures d'association binaires et également l'algorithme de sélection GenLocalMaxs. Afin de palier à l'insuffisance des représentations de suites contiguës de mots nous définirons finalement les modèles $N$-gram positionnels. Mais avant de continuer notre présentation, il nous semble nécessaire de nous arrêter quelques instants sur la définition des associations lexicales. En effet, jusqu'à présent, nous avons utilisé plusieurs termes pour faire référence à ce même concept. Ainsi, nous avons formulé sans distinction les termes *unité polylexicale*, *unité lexicale complexe*, *terme complexe*, *synapsie*, *mot composé*, *phrase idiomatique*. Nous proposons donc dans le prochain chapitre de définir les phénomèmes de figement de la langue — tant du point de vue linguistique que du point de vue statistique — et de préciser certaines notions de vocabulaire ■

# Chapitre 3

# Spécification des Associations Lexicales

Avant d'entreprendre quelque analyse sur les associations lexicales, il est nécessaire de définir avec précision les concepts qui leur sont sous-jacents. Or, l'étude de la bibliographie qui leur est dédiée ne manque pas de confondre un bon nombre de lecteurs novices ou non. En effet, le flou terminologique et les innombrables définitions sont autant de facteurs qui rendent difficiles leur appréciation. Historiquement, le traitement des associations lexicales a été relégué aux frontières de la lexicographie. Par voie de conséquence, il faudra attendre l'après-guerre pour voir à rendre ses lettres de noblesse à ce phénomène linguistique particulièrement important pour les phases de compréhension et de production du langage. Ainsi, l'avénement de la Phraséologie sous l'impulsion de divers linguistes dont J. Firth, M. Halliday, J. Sinclair et G. Gross permettra de redonner vie à un courant souvent considéré marginal par la linguistique classique. Parallèlement, les progrès incessants du traitement automatique des langues — TALN — se voyaient confrontés à la surenchère de l'utilisation des associations lexicales dans les corpora électroniques. La phase de compréhension des textes apparaissait alors plus complexe que N. Chomsky le laissait entrevoir. Les ingénieurs du langage[1] étaient par conséquent amenés à proposer leur propres définitions. Dans ce contexte, Y. Choueka, M. Benson et F. Smadja — entre autres — se sont livrés à cet exercice, faisant appel à des notions strictement numériques. Dans ce chapitre, nous proposons donc de réunir l'état de l'art relatif aux diverses définitions qui ont été proposées tant au niveau linguistique que numérique. Face aux insuffisances démontrées, nous verrons que l'élaboration d'une nouvelle définition s'avère

[1] Nous faisons ici référence aux spécialistes du traitement automatique des langues.

nécessaire afin de fonder notre approche probabiliste.

## 3.1 Courant Linguistique

Au-delà de ses objectifs applicatifs bien délimités, l'étude linguistique des phénomènes sous-jacents aux associations lexicales apparaît extrêmement diverse et la littérature abondante voire égarante. En effet, derrière des étiquettes proches, se profilent des problématiques souvent divergentes. Entreprendre une synthèse est donc une tâche périlleuse. Afin de nous guider dans ce travail, nous suivrons l'excellente étude proposée par G. Gross [51]. En effet, la notion linguistique la plus proche de celle d'association lexicale est certainement celle qu'il dénote de figement.

**Polylexicalité :** La première condition que G. Gross met en évidence est celle de la polylexicalité. Dans ce cadre, l'existence d'une association lexicale implique nécessairement l'existence d'une séquence de plusieurs mots qui aient par ailleurs un caractère autonome. Ainsi, une association lexicale peut être réalisée par une ou plusieurs formes graphiques. Les associations lexicales réalisées en une seule forme graphique sont traditionnellement nombreuses dans les langues anglo-saxonnes. *Bathroom*[2] en est l'exemple classique. En effet, cette association lexicale n'est autre que la juxtaposition des deux mots autonomes *bath* et *room*. En Français et en Portugais[3], ce phénomène est plus rare. Il n' en est cependant pas absent. Par exemple, *malappris* en Français est une structure où les deux mots *mal* et *appris* sont juxtaposés. Pareillement en Portugais, *malcriado* est la concaténation de *mal* et *criado* qui veut dire *mal éduqué*. En ce qui concerne les combinaisons réalisées par plusieurs formes graphiques, elles doivent répondre à un certain nombre de contraintes pour être considérées associations lexicales. L'une d'entre elles est l'opacité sémantique.

**Opacité sémantique :** Un nombre non négligeable d'associations lexicales ont un sens opaque. Par opaque, on entend que leur sens n'est pas calculable à partir du sens de leurs constituants. Par exemple, il est peu probable qu'un étranger puisse interpréter littéralement la séquence *prendre le large* même s'il connaît le sens habituel de tous les mots qui la composent. Cette contrainte ne fait pourtant pas l'unanimité dans la communauté linguistique. En effet, B. Habert et C. Jacquemin [43] soulignent que pour la plupart des

[2]Traduction : "*Salle de bain*".

[3]Et plus généralement dans les langues d'origine latine.

travaux de linguistes, les associations lexicales "*... sont disposées sur un continuum qui va de l'opacité à la composition limpide*". Ainsi, les expériences de P. Downing [64], comme celles de P. Sébillot et P. Boucher [65], sont autant de présomptions en faveur d'une vision compositionnelle d'une bonne partie des associations lexicales de type nominal[4]. Il est clair que la notion d'opacité est critiquable. Premièrement, parce qu'elle est difficilement définissable par ses partisans. Deuxièmement, parce que ses détracteurs voient en elle l'incapacité de traiter de la sémantique, domaine connu comme résistant à la formalisation et ne donnant pas lieu à des jugements reproductibles. Cependant, derrière cette notion, se cache l'un des fondements du courant numérique. En effet, "l'opacité" d'une séquence est souvent la preuve de l'existence d'une certaine cohésion entre tous ses constituants. Dans ce contexte, plutôt que d'opacité, nous devrions parler de cohésion. En effet, la séquence *arc de triomphe* est un exemple typique d'"image unique" de Grévisse dans le sens où elle est généralement saisie globalement, comme un tout. Cependant, son sens n'est pas complètement opaque. Mais, l'attirance que démontrent ses constituants est révélatrice d'une certaine opacité de sens. La troisième caractéristique mise en évidence par G. Gross va dans le sens de proposer une caractérisation de cette cohésion.

**Blocage des propriétés transformationnelles :** Les constructions libres ont des propriétés transformationnelles qui dépendent de leur organisation interne. Ainsi, la relation entre un verbe et son complément peut faire l'objet de certains changements de structures appelés transformations. Parmi celles-ci, on soulignera la passivation, la pronominalisation, le détachement, l'extraction et la relativation. A l'opposé, les constructions figées telles que les associations lexicales sont généralement dépourvues de propriété de recomposition et ne font l'objet d'aucune modification. On dira qu'elles sont syntaxiquement figées. Observons la phrase suivante.

*Le voleur a pris le large.*

L'opacité démontrée par l'association *prendre le large* est corrélée à une absence de propriété transformationnelle, signe de sa forte cohésion. Ainsi, les phrases suivantes sont incorrectes.

– passivation : **Le large a été pris par le voleur.*

[4]Nous considérons cette désignation équivalente à celle de composition nominale proposée par B. Habert et C. Jacquemin [43].

- pronominalisation : **Le voleur l'a pris.*
- détachement : **Ce large, le voleur l'a pris.*
- extraction : **C'est le large que le voleur a pris.*
- relativation : **Le large que le voleur a pris.*

On voit donc que les phénomènes sous-jacents aux associations lexicales transcendent les différents niveaux classiques de l'analyse linguistique et qu'une description qui ne serait que syntaxique ou sémantique serait nécessairement réductrice.

**Non-actualisation des éléments :** Cette quatrième contrainte est un prolongement des deux précédentes. Nous avons vu le figement syntaxique et en quelque sorte le figement sémantique. Nous abordons maintenant le figement des éléments lexicaux constitutifs des associations lexicales. Ainsi, on peut effectivement parler d'association lexicale lorsqu'aucun de ses constituants ne peut être actualisé. Afin d'illustrer cette caractéristique, nous reprenons notre exemple initial : *Le voleur a pris le large.* Ainsi, les phrases suivantes où le déterminant du substantif *large* a été modifié, ne pourraient pas être retenues comme linguistiquement correctes.

**Le voleur a pris son large.*
**Le voleur a pris un large.*
**Le voleur a pris ce large.*

En effet, le caractère non compositionnel de cette phrase implique des contraintes d'actualisation sur ses éléments. *Prendre le large* est en soit une et une seule unité lexicale atomique qui ne peut être modifiée en son sein. Cependant, même si ce phénomène est particulièrement caractéristique des associations lexicales, il n'en est pas pour autant vrai pour toutes les séquences figées. Par exemple, en Portugais, on rencontrera *greve da fome* et *greve de fome* pour faire référence au terme *grève de la faim.* Ainsi, *de*[5] et *da*[6] sont interchangeables. Les prochaines caractéristiques que G. Gross propose dans son étude, portent justement sur les différents degrés de figement qui sont intrinsèques aux séquences lexicales.

**Blocage des paradigmes synonymiques :** L'axe paradigmatique défini en linguistique permet de remplacer un mot d'une phrase soit par un autre de la même classe

[5] *De* est la préposition *de* du Français.
[6] *Da* est la contraction entre la préposition *de* et l'article défini *a* (*la* en Français).

sémantique soit par un synonyme dans une hypothèse plus restreinte. Dans ce cadre, de la même façon que l'on dit *jouer au football*, il est possible de dire *jouer au foot* ou encore *jouer au ballon*. Evidemment, le sens de la phrase est légèrement modifié, mais celle-ci reste interprétable. Ainsi, ces possibilités de substitution dépendent de la nature des prédicats et relèvent de contraintes très générales. En ce qui concerne les associations lexicales, la possibilité de remplacement synonymique est exclue. En effet, la substitution d'un mot de la séquence par un autre de même sens retire toute correction à la séquence. Par exemple, *pomme de terre* ne peut donner lieu à des variations comme *pomme de sol* ou *pomme de terrain*. Cette même observation peut être faite pour toutes les catégories des associations lexicales. Ainsi, pour le cas des verbes, *rire aux éclats* ne peut être substitué par *sourire aux éclats* ou *rire aux fragments*. On verra cependant que cette affirmation n'est pas totalement valide. En effet, il est possible d'assister à des substitutions pour des associations lexicales qui ne sont pas totalement figées mais dont le sens est raisonnablement opaque. Par exemple, il n'est pas impossible de rencontrer les deux séquences suivantes utilisées comme équivalentes : *louper le coche* et *rater le coche*. Une certaine attention devra donc être portée à ce type de phénomène.

**Non-insertion :** Au contraire des suites libres, les associations lexicales sont particulièrement hostiles à l'introduction d'éléments en leur sein. Cette remarque est surtout vraie pour les combinaisons nominales. Par exemple, il n'est pas possible d'insérer quelconque élément à l'intérieur de la séquence *maître d'école*. En effet, on ne peut pas dire *maître de la bonne école* ou encore *maître fantastique d'école*. A l'opposé, ce phénomène n'est pas vrai pour les suites libres. Ainsi, il est fort possible d'introduire un adverbe quelconque dans la séquence nominale *constructions libres*. Par exemple, *constructions particulièrement libres* et *constructions très libres* sont des suites correctes. Le caractère cohésif des associations lexicales tend à les considérer comme des éléments lexicaux atomiques qu'il est difficile de modifier de l'intérieur. Ainsi, on dira *un fort coup de coeur* et non pas *un coup fort de coeur*. Cependant, cette hypothèse mise en avant par G. Gross n'est pas une règle nécessaire pour l'identification d'associations lexicales. En effet, suivant le degré de figement qui caractérise une séquence, il est possible d'introduire ou non des éléments supplémentaires. Ainsi, le lecteur acceptera sans difficulté le caractère figé de la combinaison *Ministre des Affaires Etrangères*. Or, dans les textes de la Communauté Européenne, il n'est pas rare de rencontrer les suites *Ministre Français des Affaires Etrangères* ou *Ministre Portugais des Affaires Etrangères*. Dans

ce contexte, la suite n'est pas totalement rigide bien que son sens soit raisonnablement opaque. Cette remarque est également vraie pour les associations lexicales de type verbal. Par exemple, on peut dire *prendre régulièrement le large* ou *franchir difficilement le pas.*

Comme nous l'avons vu, du point de vue linguistique, il est difficile de définir d'une forme cohérente ce que sont réellement les associations lexicales. Il est en effet plus facile de lister leurs propriétés de figement les plus évidentes. Cependant, comme le souligne G. Gross, "*... on constate que les suites totalement figées sont très minoritaires par rapport à celles qui ont des restrictions partielles*". Par conséquent, une analyse au cas par cas, comme ce qui se fait au LADL s'avère nécessaire pour leur identification. Le traitement automatique du langage ne peut donc pas hériter d'un corps de doctrine éprouvé qu'il resterait à appliquer. Dans ce contexte, on peut espérer que les analyses numériques viennent compléter les indices fragiles fournis par la linguistique. Dans ce cadre, l'utilisation de mesures quantitatives impose la définition de règles strictes d'identification qui contrastent avec la quasi absence de spécifications du point de vue linguistique. Ainsi, pour notre approche probabiliste, nous serons amenés à proposer une nouvelle définition de la notion d'association lexicale.

## 3.2 Courant Numérique

Dans le cadre du TALN, le développement toujours croissant d'outils informatiques de plus en plus performants a conduit un certain nombre de chercheurs à proposer leurs propres définitions au phénomène d'association lexicale.

Y. Choueka [66] est le premier à définir clairement la notion d'association lexicale d'un point de vue strictement numérique. Dans ce cadre, loin des impératifs linguistiques, une association lexicale est une "*... séquence de mots adjacents qui apparaissent fréquemment ensemble*". Cependant, cette définition n'est pas suffisante pour des analyses statistiques précises. En effet, celle-ci ne suggère que la fréquence comme facteur décisif d'identification des associations lexicales. Ainsi, toutes les séquences fréquentes des textes seraient nécessairement des associations lexicales. Cette définition est somme toute trop vague. Par exemple, tous les fragments fonctionnels tels que *de la*, *et le*, *sur le* seraient définis comme des association lexicale. Ceci n'est évidemment pas souhaitable. En effet, l'écart entre les phénomènes mis en évidence par les phraséologistes et ceux effectivement

identifiés par Y. Choueka est demesurément grand.

Dans le but de répondre à cette insuffisance, M. Benson [67] propose une définition plus fidèle au phénomène d'association lexicale : "*une association lexicale ... est une combinaison arbitraire et récurrente de mots*". L'introduction du terme *arbitraire* joue un rôle fondamental dans cette définition. M. Benson prétend rendre compte des phénomènes d'associations dont le sens est opaque — i.e. non compositionnel. Ainsi, en plus du facteur fréquence, M. Benson impose une nouvelle contrainte, celle du refus de la libre combinaison des occurrences présentes dans les séquences de mots. En effet, suivant la théorie cognitive, nous devons considérer toute suite de mots qui se réfère à un concept ou à un objet du domaine comme une association lexicale. Ainsi, l'association lexicale *coup de coeur* correspond à une utilisation arbitraire des mots qui la composent et son sens est complètement opaque. Chaque constituant ne peut être interchangé avec l'un de ses synonymes. Cependant, pour une grande partie des associations lexicales, leur sens n'est pas entièrement opaque. A l'extrême, celui-ci peut même être calculé par composition. Par exemple, l'association lexicale *accord salarial* ne correspond pas à une utilisation arbitraire des mots qui la composent et son sens peut raisonnablement être approché par la composition du sens de ses constituants *accord* et *salarial*. Le caractère arbitraire supposé par M. Benson est donc trop restrictif et ne permet pas de rendre compte de l'ensemble des associations lexicales. De plus, les méthodes quantitatives ne permettent pas d'assurer le caractère arbitraire d'une combinaison de mots. En effet, pour garantir ce concept d'opacité, la seule fréquence n'est absolument pas suffisante.

L'une des définitions les plus correctes est certainement celle de F. Smadja [31] qui introduit la notion essentielle de plausibilité pour l'identification des associations lexicales. Ainsi, une association lexicale est "*... une combinaison récurrente de mots qui se trouvent ensemble plus souvent que par le simple fait du hasard et qui correspondent à une utilisation arbitraire*". Grâce à cette définition, F. Smadja met en évidence les forces d'attraction qui lient entre eux tous les mots d'une combinaison. En effet, les associations lexicales sont des séquences de mots qui démontrent un certain degré de cohésion intrinsèque. Ainsi, la présence d'un ou plusieurs mots d'une combinaison suggère souvent l'occurrence des autres constituants de l'association. Par exemple, la probabilité que les deux mots de la séquence *Communauté Européenne* apparaissent ensemble est bien plus

élevée que le produit des deux probabilités marginales d'occurrence de *Communauté* et *Européenne*, signe de leur attraction. F. Smadja édifie ainsi les bases du traitement probabiliste pour l'identification des associations lexicales. Cependant, de la même façon que nous avons renié le caractère arbitraire de la définition de M. Benson, nous le rejetons de celle de F. Smadja. Mais, plus important que cette remarque, nous devons souligner l'absence d'allusion au domaine qui est le support de l'analyse probabiliste.

L'une des propriétés fondamentales du traitement numérique consiste à prendre en compte comme paramètre de l'analyse le domaine dans lequel les décomptes des mots sont réalisés. En effet, les associations lexicales dépendent du domaine dans lequel elles sont identifiées. En particulier, une grande proportion des associations lexicales sont des termes techniques. Dans ce contexte, des combinaisons qui ne contiennent apparemment pas de mots spécifiques du domaine considéré, peuvent révéler un sens pratiquement opaque, exclusif au domaine. Par exemple, dans le domaine de la Statistique, l'association lexicale *analyse de données* correspond à un champ d'application bien défini qui fait l'objet de nombreuses études. A l'extérieur de ce domaine, la même séquence pourrait certainement être remplacée sans difficulté de compréhension par la combinaison *étude de résultats*. En effet, hors domaine, il est peu probable que cette séquence démontre un degré de cohésion suffisamment fort pour qu'elle soit identifiée comme une association lexicale. L'ensemble de ces constatations nous a donc amenés à proposer une nouvelle définition des associations lexicales dans le cadre des études numériques : "*une association lexicale est une combinaison récurrente de mots qui se trouvent ensemble plus souvent que par le simple fait du hasard dans un domaine donné*". Ainsi, nous reprenons les bases formulées par F. Smadja auxquelles nous éliminons le caractère arbitraire et ajoutons le critère essentiel du domaine de l'analyse.

## 3.3 Conclusion

Après avoir clairement défini ce que représentent les associations lexicales dans le cadre de l'approche numérique, nous pensons que le lecteur sera en mesure de comprendre sans difficulté les concepts qui seront introduits dans la seconde partie de notre rapport sur notre méthodologie d'extraction. Mais avant de poursuivre nos propos, nous nous attardons quelque peu sur la définition d'association lexicale que nous avons présentée. L'une

des particularités des méthodes numériques est de permettre une totale flexibilité d'utilisation. En particulier, elles peuvent être testées sur tout type de matériel textuel. Ainsi, il n'est pas exclus d'appliquer les méthodes quantitatives sur des textes de caractères[7], d'étiquettes morpho-syntaxiques ou de mots — entre autres. Dans le cas précis de notre étude, nous nous sommes plus particulièrement intéressés au cas des associations lexicales c'est-à-dire associations entre mots. Nous avons cependant réalisé quelques expériences intéressantes à partir de textes de caractères et d'étiquettes morpho-syntaxiques. Dans ce contexte, des associations morphologiques et syntaxiques ont été identifiées. Nous leur réservons une place importante dans la dernière partie de notre rapport. Le point capital sur lequel nous souhaitons mettre l'accent est l'extension de la définition d'association lexicale au cas des différentes unités textuelles possibles — caractère, mot, étiquette morpho-syntaxique. Ainsi, on parlera d'association textuelle pour définir quelconque association entre différentes unités textuelles de même type. Dans la suite de ce rapport, nous ferons donc référence au terme association textuelle dans le cas général et association lexicale pour le cas spécifique des associations entre mots. La définition générale d'association textuelle pourrait ainsi être formulée : "*une association textuelle est une combinaison récurrente d'unités textuelles qui se trouvent ensemble plus souvent que par le simple fait du hasard dans un domaine donné*" ■

[7] Le Japonais et le Mandarin imposent même cette étude.

# Deuxième partie

# Méthodologie

**"L'étude des textes à l'aide de la méthode statistique constitue le centre d'une sphère d'intérêts que l'on désigne par statistique textuelle"**

---

*Lebart et Salem [1]*

# Chapitre 4

# Préparation des Données Textuelles

La méthode statistique s'appuie sur des mesures et des comptages réalisés à partir d'objets que l'on veut comparer. Dans ce cadre, une normalisation de ces unités s'impose. Dans ce chapitre, nous définissons dans un premier temps les unités minimales de traitement — unités textuelles — puis les unités minimales de mesure et de comptage — modèles $N$-gram positionnels. Parallèlement, nous introduisons une méthode de codification de textes pour la mise en oeuvre des traitements informatisés.

## 4.1 Unités Textuelles

La mesure statistique s'appuie sur des mesures et des comptages réalisés à partir d'objets que l'on veut comparer. Décompter des unités, les additionner entre elles signifient les considérer comme des unités de même type ou d'une forme plus générale — lemmes ou *stems* par exemple. Pour soumettre une série d'objets à des comparaisons statistiques, il faut définir dans un premier temps une série de liens systématiques entre des cas particuliers et des catégories plus générales. Dans la pratique, l'application de ces principes implique que soit définie une norme permettant d'isoler de l'énoncé les différentes unités sur lesquelles portent les dénombrements. L'opération qui permet de découper le texte en unités minimales — unités indivisibles le temps d'une expérience — s'appelle la segmentation. Une fois définie une norme de segmentation, les méthodes de la statistique textuelle s'appliquent sans adaptation particulière aux comptages réalisés à partir des différentes unités minimales — unités textuelles. Dans le cadre de notre recherche, nous définissons

trois normes de segmentation pour un corpus étiqueté morpho-syntaxiquement : la segmentation en caractères, en formes graphiques et en étiquettes morpho-syntaxiques. Mais avant tout, pour que la segmentation puisse être envisagée, il est nécessaire de ne retenir de la forme informatisée de l'énoncé que son essence c'est-à-dire son contenu indépendamment des normes typographiques de formatation dont il a été l'objet. Cette étape s'appelle communément le pré-traitement et consiste à faire abstraction de l'encodage des textes.

### 4.1.1 Pré-traitement

De nombreux projets de constitution de corpora en format électronique ont adopté l'initiative d'encodage des textes *TEI — Text Encoding Initiative* — qui définit un inventaire des divers éléments qui peuvent constituer un document littéraire. Cette initiative représente en ce sens une avancée dans la description et la formalisation des types de documents en circulation dans les différentes communautés langagières. En particulier, grâce à la définition d'un document type appelé *DTD — Document Type Definition*, cette initiative met à disposition un ensemble de représentations logiques qui permettent de rendre explicites certaines particularités qui sont exprimées implicitement dans les textes — structure, typographie, etc. Cette opération s'appelle le balisage et revient à introduire dans l'énoncé des balises *SGML — Standard General Markup Language* — relevant du *DTD* considéré . Nous montrons un énoncé balisé dans l'exemple suivant ainsi que son équivalent en version "papier" c'est-à-dire après sa transcription inverse.

> *<text> Le survol des problématiques linguistiques sur la composition nominale met en évidence l'absence de critères purement linguistiques permettant aux théoriciens de délimiter l'ensemble des "noms composés". </text>*

Enoncé balisé

> *Le survol des problématiques linguistiques sur la composition nominale met en évidence l'absence de critères purement linguistiques permettant aux théoriciens de délimiter l'ensemble des "noms composés".*

Enoncé après transcription inverse

Bien que l'ensemble des balises fournisse un certain nombre d'informations fondamentales pour l'analyse et la compréhension de l'énoncé [68] [69], leur exploitation pose un certain

nombre de problèmes qui dépassent le cadre de notre travail. D'une part, il n'est pas facile de comprendre dans quelle mesure la structure des textes joue un rôle dans l'identification des associations textuelles. D'autre part, de nombreux corpora en format électronique ne suivent pas les recommandations du *TEI* et sont compilés suivant différents formats — *rtf*, *doc*, *html*, *txt*, *ps*, *pdf* etc — qui utilisent diverses normes de balisage. Une étude spécifique des balises *SGML* serait par conséquent limitatrice et contraire à nos objectifs et motivations de départ. Par ailleurs, une étude exhaustive de toutes les formes d'encodage n'est pas envisageable dans le cadre de notre étude. Par conséquent, nous appliquons un pré-traitement au corpus afin de le segmenter en unités textuelles de base. Le pré-traitement du corpus consiste à éliminer toutes les balises logiques qui ont été introduites lors de son encodage, de forme à ne retenir comme données de base que l'essence du texte et non pas l'ensemble des données textuelles associées aux balises logiques de la formatation considérée. Nous avons donc développé un certain nombre d'heuristiques pour le cas spécifique de l'encodage recommandé par l'initiative *TEI* qui correspond à la norme la plus courante des corpora compilés. Cette étape ne présentant pas d'intérêt particulier au niveau théorique, nous assumerons — lucidement — l'idée que le lecteur aura compris l'objectif du pré-traitement et par conséquent nous ne nous attarderons pas plus sur cette notion.

A partir de l'énoncé pré-traité, nous définissons dans la partie suivante trois types de segmentation qui nous permettent d'envisager l'extraction de trois formes d'associations textuelles différentes : la segmentation en caractères, en formes graphiques et en étiquettes morpho-syntaxiques.

### 4.1.2 Segmentation en caractères

La segmentation est l'opération qui consiste à découper un texte en unités indivisibles — unités que l'on ne décomposera pas plus par la suite.

> *Le survol des problématiques linguistiques sur la composition nominale met en évidence l'absence de critères purement linguistiques permettant aux théoriciens de délimiter l'ensemble des "noms composés".*

Enoncé pré-traité

*L e * s u r v o l * d e s * p r o b l é m a t i q u e s * l i n g u i s t i q u e s * s u r * l a * c o m p o s i t i o n * n o m i n a l e * m e t * e n * é v i d e n c e * l ' a b s e n c e * d e * c r i t è r e s * p u r e m e n t * l i n g u i s t i q u e s * p e r m e t t a n t * a u x * t h é o r i c i e n s * d e * d é l i m i t e r * l ' e n s e m b l e * d e s * " n o m s * c o m p o s é s " .*

Enoncé segmenté en caractères

Or, la seule véritable unité indivisible qui compose les textes est le caractère — un texte n'étant qu'une suite ordonnée de caractères. La segmentation en caractères revient donc à découper l'énoncé caractère à caractère de façon à effacer toute structure apparente du texte. En particulier, E. Planas [69] lui donne le nom de *couche 1.* Nous illustrons cette décomposition dans l'exemple précédent dans lequel le caractère "espace" a été traduit pour des raisons évidentes de présentation par le caractère "*". Les méthodes statistiques s'appliqueront par conséquent aux comptages réalisés à partir des caractères permettant ainsi d'identifier un certain nombre d'associations entre caractères — unités morphologiques.

### 4.1.3 Segmentation en formes graphiques

Traditionnellement, un énoncé est défini comme étant une suite organisée de mots d'un certain vocabulaire. Cependant, le repérage des mots d'un énoncé est une tâche difficile [14]. En effet, un certain nombre de caractères fonctionnent tantôt comme composants tantôt comme séparateurs de mots. C'est le cas du trait d'union et de l'apostrophe pour le Français et le Portugais.

| Caractère non séparateur | Caractère séparateur | Langue |
|---|---|---|
| *aujourd'hui* | *l'absence* | Français |
| *va-et-vient* | *vas-tu* | Français |
| *vai-não-vai* | *ausentar-se* | Portugais |

TAB. 4.1 – Exemples de caractères délimiteurs et non délimiteurs

On remarquera que le problème de la segmentation en mots dépasse la notion de caractères délimiteurs dans le cadre des langues anglo-saxonnes. En effet, ces dernières présentent la particularité de créer des mots composés en agglomérant plusieurs substantifs ou radicaux

verbaux comme en Allemand.

| Mots composés Allemand | Traduction Française |
|---|---|
| Trinkwasser | eau potable |
| Zusammengehörigkeitsgefühl | sentiment d'appartenance à |

TAB. 4.2 – Exemples cités dans [1]

Dans le cadre des méthodes de la statistique textuelle, nous introduisons l'approche du découpage des énoncés en formes graphiques [55]. Pour réaliser une segmentation automatique du texte en occurrences de formes graphiques, il suffit de définir parmi l'ensemble des caractères un sous-ensemble que l'on désigne sous le nom d'ensemble des caractères délimiteurs et dont l'ensemble complémentaire s'appelle ensemble des caractères non-délimiteurs. Ainsi, une suite de caractères non-délimiteurs bornée à ses deux extrémités par des caractères délimiteurs forme une occurrence d'une forme graphique. La segmentation ainsi définie permet de considérer le texte comme une suite d'occurrences de formes graphiques séparées entre elles par un ou plusieurs caractères délimiteurs. L'ensemble des caractères délimiteurs est généralement défini par le caractère "espace", l'ensemble des caractères de ponctuation, les caractères invisibles — retour-chariot, tabulation etc. — et les caractères spéciaux — parenthèses, guillemets etc. Cette approche suppose qu'à chacun des caractères de l'énoncé corresponde un statut et un seul, c'est-à-dire que le texte ait été débarrassé de certaines ambiguïtés de codage que représentent par exemple les points de fin de phrase, d'abréviations et de nombres. Par exemple, dans l'abréviation *T.A.L.* ou dans le nombre *1,969*, le point et la virgule ne représentent pas une séparation bien qu'ils fassent respectivement partie de l'ensemble des caractères délimiteurs. Dans ce cadre, on citera comme référence les travaux réalisés par E. Brill [70] qui définit un certain nombre d'heuristiques afin d'effacer les ambiguïtés introduites par les caractères délimiteurs. Nous illustrons la segmentation en formes graphiques à partir de l'énoncé suivant dans lequel seuls les guillemets, le point de fin de phrase et le caractère "espace" servent de délimiteurs.

> *Le survol des problématiques linguistiques sur la composition nominale met en évidence l'absence de critères purement linguistiques permettant aux théoriciens de délimiter l'ensemble des "noms composés".*

Enoncé pré-traité

*Le survol des problématiques linguistiques sur la composition nominale met en évidence l'absence de critères purement linguistiques permettant aux théoriciens de délimiter l'ensemble des " noms composés " .*

Enoncé segmenté en formes graphiques

On remarquera que quand un article est défini à partir de sa forme condensée grâce à l'apostrophe, l'article n'est pas dissocié de son substantif. Cette règle est bien entendue sujette à débat, notre point de vue étant que l'association entre l'article et le substantif est si forte dans ce cas qu'il n'est pas possible de dissocier l'article de son substantif. Avant de poursuivre, nous tenons à citer L. Lebart et A. Salem [1] qui soulignent que "*[...] la forme graphique ne constitue en aucun cas une unité naturelle pour le dépouillement des textes ; l'avantage des décomptes en formes graphiques réside avant tout dans la facilité incomparable qu'il y a à les automatiser*". En effet, il est clair que les formes graphiques ne représentent en aucun cas une solution optimale. Cependant, les mêmes auteurs montrent que "*[...] les typologies réalisées à partir des décomptes textuels se révèlent peu sensibles aux variations de l'unité de décompte*". Cette affirmation confirme notre position initiale selon laquelle l'introduction de contraintes extérieures aux textes ne bénéficie pas forcément de l'analyse de l'énoncé.

### 4.1.4 Segmentation en étiquettes morpho-syntaxiques

Avec l'évolution des techniques dans le domaine du traitement automatique des langues, les corpora en format électronique ont évolué de la forme de simples suites de mots nus vers le statut de ressources linguistiquement riches. Au-delà du texte proprement dit, les corpora contiennent aujourd'hui un ensemble d'informations complexes — informations morphologique, syntaxique, sémantique etc. Ces corpora sont dits annotés — ou encore étiquetés ou enrichis. Parmi ceux-ci, les corpora étiquetés morpho-syntaxiquement sont les plus courants. Dans ce contexte, à chaque mot du texte est associée une étiquette — label — morpho-syntaxique qui représente le plus souvent sa catégorie grammaticale voire son lemme. A titre d'exemple, un ensemble d'étiquettes morpho-syntaxiques est défini dans le tableau 4.3.

La segmentation d'un corpus annoté correspond au découpage du texte enrichi en unités indivisibles que sont les étiquettes morpho-syntaxiques. L'ensemble des caractères de l'énoncé originel — avant annotation — est donc supprimé pour ne retenir que l'ensemble

| Etiquette | Définition |
|---|---|
| /ADJ | Adjectif |
| /ADV | Adverbe |
| /N | Nom |
| /PREP | Préposition |
| /V | Verbe |
| /SPEC | Catégorie Spéciale |

TAB. 4.3 – Exemple d'étiquettes morpho-syntaxiques

des étiquettes. Nous illustrons cette segmentation à partir de l'énoncé suivant.

*Le /ART survol /N des /PREP problématiques /N linguistiques /ADJ sur /PREP la /ART composition /N nominale /ADJ met /V en /PREP évidence /N l'absence /N de /PREP critères /N purement /ADV linguistiques /ADJ permettant /V aux /PREP théoriciens /N de / PREP délimiter /V l'ensemble /N des /PREP " /SPEC noms /N composés /ADJ " /SPEC . /SPEC*

Enoncé étiqueté

*/ART /N /PREP /N /ADJ /PREP /ART /N /ADJ /V /PREP /N /N /PREP /N /ADV /ADJ /V /PREP /N / PREP /V /N /PREP /SPEC /N /ADJ /SPEC /SPEC*

Enoncé segmenté en étiquettes morpho-syntaxiques

On remarquera que cette segmentation peut être facilement étendue à d'autres types d'étiquettes. Il suffit alors de découper le texte suivant l'ensemble des étiquettes à considérer — sémantiques, prosodiques etc.

Après avoir défini les unités minimales de traitement qui vont servir de données de base, nous introduisons dans un deuxième temps les unités minimales de mesure et de comptage — modèles $N$-gram. Dans la plupart des travaux du TALN, la présence d'une unité textuelle est déterminée uniquement par son contexte immédiat antérieur — l'ensemble des unités textuelles voisines qui la précèdent. Dans ce cadre, de nombreux travaux en statistique textuelle ont privilégié l'étude des modèles $N$-gram "classiques" — suites ininter-

rompues de $N$ unités textuelles. Cependant, il est du sens commun que l'occurrence d'une unité textuelle peut être dictée non seulement par son contexte immédiat antérieur mais aussi par son contexte postérieur. De même, toutes les unités textuelles qui constituent le contexte immédiat d'une unité n'influencent pas forcément sa présence. Afin d'appréhender ces affirmations, nous introduisons les modèles $N$-gram positionnels – suites ininterrompues ou interrompues de $N$ unités textuelles.

## 4.2 Modèles $N$-gram Classiques

Les modèles $N$-gram ont d'abord été introduits par Shannon [71] dans le cadre de la prédiction d'occurrence de caractères pour l'anglais – Jeu de Shannon. Shannon propose de suivre la règle de Markov selon laquelle la présence d'une unité textuelle — UT — est déterminée uniquement par l'ensemble de ses UTs précédentes. Dans ce cas précis, Shannon prédit la présence d'un caractère à partir de la séquence ininterrompue des $N-1$ caractères qui le précèdent. En regroupant toutes les séquences identiques de $N-1$ UTs dans une même classe d'équivalence, Shannon obtient un modèle de Markov d'ordre $N-1$ ou modèle $N$-gram d'UTs – l'UT de rang $N$ étant l'unité à prédire. Dans ce cadre, Shannon utilise les modèles $N$-gram pour des valeurs de $N = 2, 3, 4$ que l'on nomme classiquement digram[1], trigram et tetragram.

### 4.2.1 Définitions

Un modèle $N$-gram d'UTs est défini comme l'ensemble des suites ininterrompues de $N$ UTs — $N$-grams contigus — construites à partir d'un énoncé initial. Ainsi, nous proposons les définitions suivantes pour les concepts de $N$-gram contigu et de modèle $N$-gram.

> *Un* ***N-gram contigu*** *d'unités textuelles est une séquence ordonnée de N unités textuelles correspondant à une séquence continue d'un énoncé — séquence ininterrompue. L'ordre de la séquence est défini par l'ordre d'apparition des UTs dans l'énoncé.*

[1]Nous retiendrons cette terminologie plutôt que celle de "bigram" qui ne respecte pas l'adéquation entre préfixe et thème. En effet, la racine grecque *gram* doit recevoir un préfixe grecque — *di* — et non un préfixe latin — *bi*.

> *Un* ***modèle N-gram*** *d'unités textuelles est l'ensemble des N-grams contigus calculés à partir d'un énoncé initial d'unités textuelles.*

Un $N$-gram contigu est généralement représenté par un vecteur ordonné d'UTs. Nous adoptons donc la notation suivante pour un $N$-gram contigu générique où $u_i$ — $i$ variant de 1 à $N$ — représente une UT.

$$[u_1\ u_2\ u_3\ \ldots u_i\ \ldots u_N]$$

### 4.2.2 Implémentation

Dans le but d'observer — compter et mesurer — les cooccurrences d'UTs, les chercheurs en statistique textuelle ont longuement privilégié l'étude des modèles $N$-gram. Ceci s'explique principalement par la simplicité de l'implémentation des modèles. En effet, l'ensemble des $N$-grams contigus peut être facilement calculé à partir de la lecture séquentielle d'un énoncé. Chaque UT de l'énoncé est tour à tour UT pivot — unité de traitement — et une fenêtre de taille $N-1$ délimite son contexte immédiat postérieur. Dans ce cadre, l'UT pivot est l'UT de référence à partir de laquelle le $N$-gram est construit. Un $N$-gram contigu est alors construit pour chaque UT pivot et défini comme la concaténation de l'UT pivot et des $N-1$ UTs suivantes formant ainsi une suite ininterrompue de $N$ UTs.

> *Le* (*pivot*) *survol des problématiques linguistiques sur la composition nominale met en évidence l'absence de critères purement linguistiques permettant aux théoriciens de délimiter l'ensemble des " noms composés " .*

Situation initiale

> *Le survol* (*pivot*) *des problématiques linguistiques sur la composition nominale met en évidence l'absence de critères purement linguistiques permettant aux théoriciens de délimiter l'ensemble des " noms composés " .*

Situation après une lecture séquentielle

Cette situation est illustrée dans l'exemple précédent pour le modèle 3-gram, définissant respectivement les deux 3-grams contigus [*Le survol des*] et [*survol des problématiques*] après une lecture séquentielle.

### 4.2.3 Complexité

L'utilisation des modèles $N$-gram s'explique également par leur complexité linéaire de calcul en fonction de la taille de l'énoncé. En effet, si l'on note $T$ le nombre d'UTs contenues dans un énoncé — taille de l'énoncé, il est facile de vérifier que chaque modèle $N$-gram d'unités textuelles est l'ensemble de $T-(N-1)$ $N$-grams contigus où $N-1$ est la taille de la fenêtre considérée. Or, il est nécessaire de construire tous les modèles $N$-gram à partir de l'UT pivot. Par conséquent, si l'on considère que pour chaque UT de l'énoncé il est possible de construire $K$ $N$-grams contigus — un $N$-gram contigu pour chacun des $K$ modèles $N$-gram, la complexité de calcul de l'ensemble des modèles $N$-gram est $\mathcal{O}(K.T)$ c'est-à-dire linéaire en fonction de la taille de l'énoncé. En effet, exactement $K.T - \frac{K.(K-1)}{2}$ $N$-grams contigus sont construits pour l'ensemble des modèles comme il est possible de le vérifier à partir du tableau 4.4.

| Modèle $N$-gram | Nombre de $N$-grams |
|---|---|
| 1-gram | $T$ |
| 2-gram | $T-1$ |
| 3-gram | $T-2$ |
| 4-gram | $T-3$ |
| ... | ... |
| $K$-gram | $T-(K-1)$ |

TAB. 4.4 – Nombre de $N$-grams construits par $N$

### 4.2.4 Limites

Bien que largement utilisés, les modèles $N$-gram ne permettent pas de rendre compte de toutes les associations textuelles présentes dans les textes. En effet, l'occurrence d'une UT peut être dictée non seulement par son contexte immédiat antérieur mais aussi par son contexte postérieur. De même, toutes les UTs qui constituent le contexte immédiat d'une UT n'influencent pas forcément sa présence. Nous illustrons nos propos dans la prochaine partie dans laquelle nous introduisons les modèles $N$-gram positionnels.

## 4.3 Modèles $N$-gram Positionnels

L'organisation des unités textuelles en groupes cohérents fortement liés est un phénomène complexe qui ne peut être limité à la simple règle de Markov. En effet, tant le contexte immédiat antérieur que postérieur d'une UT influencent sa présence dans un énoncé. Par exemple, dans le contexte des associations lexicales, l'occurrence d'un mot n'est généralement pas déterminée par la seule séquence des UTs précédentes, mais plutôt par son environnement immédiat — contexte antérieur et postérieur. Dans l'exemple suivant, il est clair que l'occurrence de la préposition *en* est dictée par la présence simultanée du verbe *mettre* à la troisième personne du singulier — *met* — et du nom commun *évidence*.

> *Le survol des problématiques linguistiques sur la composition nominale* $\boxed{met}$ $\rightarrow$ *en* $\leftarrow$ $\boxed{évidence}$ *l'absence de critères purement linguistiques permettant aux théoriciens de délimiter l'ensemble des " noms composés " .*

Influence lexicale à contexte droit et gauche

Les modèles $N$-gram "classiques" mettent également en évidence des insuffisances de représentation. Leur caractère contigu implique que tous les éléments voisins qui précèdent une UT dictent sa présence. Or, il est facile de vérifier que cette règle ne s'applique pas dans tous les cas puisque l'occurrence d'une UT peut être déterminée à partir d'un simple sous-ensemble de son contexte. Dans l'exemple suivant, l'occurrence de la préposition *de* est déterminée par la seule occurrence du verbe *permettre* au participe présent — *permettant* — et non pas par une suite quelconque ininterrompue précédant la préposition.

> *Le survol des problématiques linguistiques sur la composition nominale met en évidence l'absence de critères purement linguistiques* $\boxed{permettant}$ $\cdots$ *aux* $\cdots$ *théoriciens* $\cdots \rightarrow$ $\boxed{de}$ *délimiter l'ensemble des " noms composés ".*

Influence lexicale non contiguë

La seule information de présence d'une UT dans le contexte antérieur d'une autre UT est clairement insuffisante pour appréhender les spécificités comportementales des associations textuelles. La définition de l'environnement immédiat d'une UT et l'introduction explicite de la notion de position entre constituants — UTs — sont abordées dans les parties suivantes.

### 4.3.1 Environnement immédiat

La plupart des associations textuelles regroupent des UTs séparées par au plus cinq autres UTs. En particulier, dans le cadre des cooccurrences lexicales, de nombreuses études lexicographiques [39] [3] [40] [41] montrent que la plupart des relations associent des mots séparés les uns des autres par au plus cinq autres mots. Dans le contexte des associations morphologiques et syntaxiques, A. Van den Bosch [72] et L. Ramshaw [18] avancent respectivement que l'ensemble des relations existantes associent des UTs distantes les unes par rapport aux autres d'au plus trois UTs. Par conséquent, une association textuelle peut être définie, en terme de structure, comme une séquence particulière — ininterrompue ou interrompue — d'UTs dans un environnement variable pouvant atteindre jusqu'à cinq UTs. Il est de noter que l'espace des UTs associées à une UT donnée — UT pivot — est ainsi délimité par l'environnement immédiat considéré. Nous définissons la notion d'environnement immédiat d'une UT comme l'ensemble de ses contextes antérieur et postérieur. Ainsi, les éléments associés à une UT devront être contenus dans son contexte antérieur — suite d'UTs précédant immédiatement l'UT — ou dans son contexte postérieur — suite d'UTs suivant immédiatement l'UT. Nous définissons formellement le concept d'environnement immédiat dans la définition suivante.

> *L'**environnement immédiat** de taille F d'une unité textuelle est défini comme l'ensemble des F unités textuelles qui précèdent immédiatement l'unité textuelle considérée — fenêtre gauche — et l'ensemble des F unités textuelles qui suivent immédiatement l'unité textuelle considérée — fenêtre droite.*

Dans l'énoncé suivant, nous représentons graphiquement l'environnement immédiat de taille 3 du nom commun *évidence* — UT pivot.

> *Le survol des problématiques linguistiques sur la composition* nominale met en *Fenêtre gauche* *évidence* *pivot* l'absence de critères *Fenêtre droite* *purement linguistiques permettant aux théoriciens de délimiter l'ensemble des " noms composés " .*

Environnement immédiat de taille 3

L'espace des UTs associées à l'UT pivot étant délimité par la taille de l'environnement immédiat considéré, le nombre de modèles $N$-gram positionnels pour un énoncé, est borné. Ainsi, la définition des modèles $N$-gram positionnels dépend intrinsèquement de la taille de l'environnement immédiat considéré. Par conséquent, nous définissons la notion de modèle $N$-gram positionnel en fonction du concept d'environnement immédiat.

### 4.3.2 Définition

Une association textuelle est définie, en terme de structure, comme une séquence ininterrompue ou interrompue de $N$ UTs dans un environnement immédiat de taille $F$. Ainsi, contrairement aux modèles $N$-gram "classiques", les modèles $N$-gram positionnels doivent prendre en compte les suites non continues de $N$ UTs dans l'environnement considéré et par conséquent rendre explicite la notion de position entre UTs[2]. Nous introduisons dans un premier temps les notions de $N$-gram positionnel et de modèle $N$-gram positionnel.

> *Un* ***N-gram positionnel*** *d'unités textuelles est une séquence ordonnée de N unités textuelles correspondant à une séquence continue ou non d'un énoncé — séquence ininterrompue ou interrompue — et délimitée par la taille de l'environnement immédiat associé. L'ordre de la séquence est défini par l'ensemble des positions des UTs calculées par rapport à l'UT pivot du N-gram positionnel.*

> *Un* ***modèle N-gram positionnel*** *d'unités textuelles est l'ensemble des N-grams positionnels calculés à partir d'un énoncé initial d'unités textuelles pour un environnement immédiat donné.*

### 4.3.3 Notation

Un $N$-gram positionnel est représenté sous la forme d'un vecteur d'UTs affectées de leur position par rapport à l'UT pivot du $N$-gram. Nous adoptons donc la notation suivante pour un $N$-gram positionnel générique d'UTs où $u_i$ représente une UT et $p_{1i}$ — $i$ variant de 1 à $N$ — définit la position de l'UT $u_i$ par rapport à l'UT pivot $u_1$ telle que pour tout $i < j$, $p_{1i} < p_{1j}$ c'est-à-dire que l'UT $u_i$ se trouve avant l'UT $u_j$ dans la séquence considérée. Par définition, l'UT pivot est la première UT du vecteur et sa position vaut zéro — plus généralement, $p_{jj}$ équivaut à zéro pour $j = 1..N$.

[2]Cette notion est implicite dans les modèles $N$-grams classiques.

$$[p_{11}\ u_1\ p_{12}\ u_2\ p_{13}\ u_3\ \ldots p_{1i}\ u_i\ \ldots p_{1N}\ u_N]$$

Par convention, les positions sont exprimées par des entiers négatifs quand une UT se trouve dans la fenêtre gauche de l'UT pivot — l'UT précède l'UT pivot dans l'énoncé — et par des entiers positifs lorsqu'une UT se trouve dans la fenêtre droite de l'UT pivot — l'UT suit l'UT pivot dans l'énoncé.

### 4.3.4 Implémentation

L'implémentation des modèles $N$-gram positionnels ne pose pas de problème majeur. En effet, l'ensemble des $N$-grams positionnels peut être facilement calculé à partir de la lecture séquentielle d'un énoncé. Chaque UT de l'énoncé est tour à tour UT pivot et une fenêtre gauche de taille $F$ et une fenêtre droite de même taille définissent son environnement immédiat.

Pour un $N$ donné, un ensemble de $C_{2.F}^{N-1}$ $N$-grams positionnels est alors construit pour chaque UT pivot. En effet, chaque $N$-gram positionnel correspond à une combinaison de $N-1$ UTs calculée dans l'environnement immédiat $F$ de l'UT pivot et combinée à l'UT pivot pour former un $N$-gram positionnel. Nous illustrons cette situation pour le modèle 2-gram positionnel dans un environnement immédiat de taille 2 à partir de l'énoncé suivant.

> *Le survol* $\underbrace{\textit{des}}_{pivot}$ *problématiques linguistiques sur la composition nominale met en évidence l'absence de critères purement linguistiques permettant aux théoriciens de délimiter l'ensemble des " noms composés " .*

Quatre — $C_4^1 = 4$ — 2-grams positionnels peuvent être construits. Ils sont énumérés dans le tableau 4.5.

Une fois réalisée la construction de tous les $N$-grams de l'UT pivot considérée, l'environnement immédiat se déplace séquentiellement pour encadrer la prochaine UT pivot. Ainsi, dans notre exemple, l'UT pivot passe à être l'UT *problématiques*. Cette situation est illustrée dans l'exemple suivant pour lequel le processus précédent de construction se répète.

| 2-grams positionnels |
|---|
| [*0 des -2 Le*] |
| [*0 des -1 survol*] |
| [*0 des 1 problématiques*] |
| [*0 des 2 linguistiques*] |

TAB. 4.5 – Exemple de 2-grams positionnels

*Le* $\boxed{\textit{survol des}}$ $\underbrace{\textit{problématiques}}_{\textit{pivot}}$ $\boxed{\textit{linguistiques sur}}$ *la composition nominale met en évidence l'absence de critères purement linguistiques permettant aux théoriciens de délimiter l'ensemble des " noms composés " .*

Quatre autres 2-grams positionnels sont alors construits. Ils sont énumérés dans le tableau 4.6[3].

| 2-grams positionnels |
|---|
| [*0 problématiques -2 survol*] |
| [*0 problématiques -1 des*] |
| [*0 problématiques 1 linguistiques*] |
| [*0 problématiques 2 sur*] |

TAB. 4.6 – Exemple de 2-grams positionnels

Afin d'éliminer tous les doutes sur l'implémentation des modèles $N$-gram positionnels, nous listons pour le modèle 3-gram positionnel dans un environnement immédiat de taille 2 les six — $C_4^2 = 6$ — 3-grams positionnels construits à partir de la situation de lecture initiale mentionnée ci-dessus.

### 4.3.5 Complexité

Contrairement aux modèles $N$-gram "classiques", les modèles $N$-gram positionnels mettent en évidence une complexité de calcul qui varie en fonction de la taille de l'énoncé et de l'environnement immédiat considéré. Avant de passer aux calculs proprement dits, nous

[3]Le lecteur averti aura remarqué que cette construction séquentielle implique la formation de doublons. Nous présenterons une solution à ce problème dans la suite de ce chapitre.

| 3-grams positionnels |
|---|
| [*0 des -2 Le -1 survol*] |
| [*0 des -2 Le 1 problématiques*] |
| [*0 des -2 Le 2 linguistiques*] |
| [*0 des -1 survol 1 problématiques*] |
| [*0 des -1 survol 2 linguistiques*] |
| [*0 des 1 problématiques 2 linguistiques*] |

TAB. 4.7 – Exemple de 3-grams positionnels

rappelons un certain nombre de définitions fondamentales de la théorie de la complexité ainsi que quelques résultats mathématiques utiles à la compréhension des démonstrations.

**Définitions**

Lorsque des programmes doivent être exécutés sur des données de grande taille, le temps d'exécution devient un critère vital. Celui-ci dépend de différents paramètres comme la puissance du processeur utilisé, la vitesse de lecture et écriture de la mémoire de l'ordinateur et bien entendu de la taille des données. Or, contrairement aux autres, ce dernier critère est le seul qui puisse être étudié sur le programme (ou l'algorithme) lui-même indépendamment de l'implémentation. On essaie alors d'évaluer le temps d'exécution par exemple en nombre d'instructions effectuées. Ce temps sera en général exprimé sous la forme d'une fonction qui dépendra du nombre des données. Le paramètre intéressant de cette fonction est son ordre qui permet d'estimer comment va évoluer le temps de calcul avec la taille des données. Cet ordre s'appelle complexité de l'algorithme. Afin de poser les fondements de la théorie de la complexité, nous introduisons trois définitions essentielles : complexité d'ordre $\mathcal{O}$, complexité d'ordre $\Omega$ et croissance exponentielle.

**Complexité d'ordre $\mathcal{O}$ :** On note $f(x) = \mathcal{O}(g(x))$ le fait qu'une fonction $f$ croisse plus lentement qu'une fonction $g$ c'est-à-dire :

$$\forall x, x > x_0 \quad \text{et} \quad x \to \infty, \exists (c, x_0) \quad \text{tels que} \quad f(x) \leq c.g(x).$$

**Complexité d'ordre $\Omega$ :** On note $f(x) = \Omega(g(x))$ le fait qu'une fonction $f$ croisse plus vite qu'une fonction $g$ c'est-à-dire :

$$f(x) = \Omega(g(x)) \Leftrightarrow f(x) = \mathcal{O}(g(x)) \quad \text{est faux.}$$

**Croissance exponentielle :** Une fonction $f$ est à croissance exponentielle s'il existe une constante $c > 1$ telle que $f(x) = \Omega(c^x)$ et s'il existe une constante $d$ telle que $f(x) = \mathcal{O}(d^x)$.

### Rappels mathématiques

Avant d'évaluer la complexité du calcul des modèles $N$-gram positionnels, nous rappelons un certain nombre de résultats mathématiques nécessaires à la compréhension des démonstrations.

$$\sum_{i=1}^{n} i = \frac{n.(n+1)}{2} \tag{4.1}$$

$$\sum_{i=1}^{n} x^i = \frac{x^{n+1} - x}{x - 1} \tag{4.2}$$

$$\sum_{i=0}^{n} \mathrm{C}_n^i = 2^n \tag{4.3}$$

Dans la partie suivante, nous présentons l'ensemble des calculs nécessaires à l'évaluation de la complexité de l'algorithme proposé pour la construction des modèles $N$-gram positionnels.

### Calculs

Si l'on note $T$ le nombre d'UTs contenues dans un énoncé — taille de l'énoncé — et $F$ la taille de l'environnement immédiat, nous vérifions que pour un modèle $N$-gram d'UTs — pour un $N$ donné, $(T - 2.F) \times \mathrm{C}_{2.F}^{N-1}$ $N$-grams positionnels sont construits[4]. Or, pour chaque UT de l'énoncé, il faut construire tous les $N$-grams positionnels pour chacun des $K$ modèles $N$-gram positionnels. Il faut donc évaluer la valeur de $K$.

Nous savons que le nombre de modèles $N$-gram positionnels est borné par l'environnement immédiat. En effet, il n'est pas possible de construire un modèle pour une valeur de $N$ excédant $2.F + 1$ puisque dans un environnement immédiat de taille $F$ nous comptons $2.F$ UTs qui peuvent être associées à l'UT pivot. Pour chaque UT pivot il est donc possible

[4]La complexité sera revue lors de l'optimisation des modèles $N$-gram positionnels dans la section 4.3.6.

de construire $K = 2.F + 1$ modèles $N$-gram. Par conséquent, l'ensemble $E$ — Espace de données — des $N$-grams positionnels calculés pour un environnement immédiat fixe $F$ est donné par la formule de l'équation suivante.

$$\begin{aligned} E &= (T - 2.F) \times \sum_{N=1}^{2.F+1} \mathrm{C}_{2.F}^{N-1} \\ &= (T - 2.F) \times \sum_{N'=0}^{2.F} \mathrm{C}_{2.F}^{N'} \\ &= (T - 2.F) \times 2^{2.F} \\ &= (T - 2.F) \times 4^{F} \end{aligned} \tag{4.4}$$

Dans ces conditions, sachant que l'on ne peut concevoir un modèle $N$-gram positionnel que pour un environnement immédiat donné et que par conséquent $F$ est une constante du système, la complexité de calcul de l'ensemble des $2.F + 1$ modèles $N$-gram positionnels est linéaire en fonction de la taille de l'énoncé. En effet, dans ce cas, les valeurs $4^F$ et $2.F$ sont des constantes et la seule variable restante est la taille de l'énoncé. Ainsi, si l'on note $c$ la constante $4^F$, la complexité de calcul des modèles $N$-gram positionnels est $\mathcal{O}(c.T)$. A titre d'exemple, nous présentons dans le tableau 4.8 la valeur de $c$ pour les valeurs de $F$ considérées dans le cadre des associations textuelles.

| Taille $F$ de l'environnement | Constante $c$ |
|---|---|
| 1 | 4 |
| 2 | 16 |
| 3 | 64 |
| 4 | 256 |
| 5 | 1024 |

TAB. 4.8 – Accélération de la constante $c$

Cependant, l'exercice qui consiste à comparer les modèles $N$-gram positionnels en fonction de la taille de l'environnement immédiat doit considérer $F$ comme une variable du système. Dans ce cas, pour chaque UT de l'énoncé, il faut calculer tous les $N$-grams positionnels pour chacun des environnements immédiats possibles. L'ensemble $E'$ de tous les $N$-grams positionnels possibles est déterminé par l'équation suivante où $F$ est la taille de l'environnement immédiat et $T$ est la taille de l'énoncé. Dans ces conditions, il est possible

de construire l'ensemble des $N$-grams positionnels pour $P$ environnements immédiats — $P$ variant de 1 à $F$.

$$\begin{aligned} E' &= \sum_{P=1}^{F}((T-2.P)\times 4^P) \\ &= \sum_{P=1}^{F}(T.4^P - 2.P.4^P) \\ &= T\times\sum_{P=1}^{F}4^P - 2\times\sum_{P=1}^{F}P.4^P \end{aligned} \tag{4.5}$$

A partir des fondements de la théorie de la complexité, nous calculons dans un premier temps la complexité d'ordre $\mathcal{O}$ de $E'$ qui peut être considérée comme une fonction $f$ dont les arguments sont $T$ et $F$ : $f(T,F) = T\times\sum_{P=1}^{F}4^P - 2\times\sum_{P=1}^{F}P.4^P$.

**Complexité d'ordre $\mathcal{O}$ :** Un majorant de $E'$ peut facilement être formulé par la suppression du terme négatif de la partie droite de l'équation 4.5. Il est déterminé dans l'inéquation suivante.

$$\begin{aligned} E' &< T\times\sum_{P=1}^{F}4^P \\ &< T\times\frac{4^{F+1}-4}{3} \\ &< T\times\frac{4^{F+1}}{3} \\ &< \frac{4.T}{3}\times 4^F \end{aligned} \tag{4.6}$$

Ainsi, si l'on considère $g(T,F) = \frac{4.T}{3}\times 4^F$, $f(T,F)$ est une fonction qui croît plus lentement que $g(T,F)$ c'est-à-dire $f(T,F) = \mathcal{O}(g(T,F))$. Cependant, $g$ démontre un facteur de croissance exponentielle si l'on prend en compte le terme $4^F$ qui la compose. Nous devons donc vérifier si la définition de croissance exponentielle s'applique au cas qui nous intéresse. Dans ce cadre, nous cherchons la complexité d'ordre $\Omega$ de la fonction $f$.

**Complexité d'ordre $\Omega$ :** Un minorant possible de $E'$ peut être formulé par la suppression du facteur de pondération $T-2.P$ de la somme initiale de l'équation 4.5. En effet, on considèrera sans problème que le terme $T-2.P$ est strictement supérieur à 1 dans le cadre de notre étude. Ainsi, nous proposons un possible minorant dans l'inéquation suivante.

$$\begin{aligned} E' &> \sum_{P=1}^{F} 4^P \\ &> \frac{4^{F+1}-4}{3} \\ &> \frac{4^{F+1}}{3} - \frac{4}{3} \\ &> \frac{4}{3} \times 4^F - \frac{4}{3} \\ &> \frac{4}{3} \times (4^F - 1) \end{aligned} \tag{4.7}$$

Donc, si l'on considère $g'(F) = \frac{4}{3} \times (4^F - 1)$, $g'$ est une fonction telle que $f(T, F) = \mathcal{O}(g'(F))$ est faux. Nous nous trouvons par conséquent dans une situation où la fonction $f$ croît exponentiellement en fonction de la taille de l'environnement et linéairement en fonction de la taille de l'énoncé. En théorie de la complexité, si le temps de calcul est au plus un polynôme des données d'entrée, le problème de calcul est un problème dit facile. Dans le cas contraire, il est difficile. Le calcul des modèles $N$-gram positionnels est évidemment difficile. Par conséquent, lors de l'implémentation des modèles $N$-gram positionnels, une attention toute particulière devra être portée à la définition de la taille de l'environnement immédiat afin d'éviter des problèmes insurmontables de calcul. Malgré tout, quelques améliorations peuvent être apportées à la construction des modèles $N$-gram positionnels. Nous les présentons dans la partie suivante.

### 4.3.6 Optimisation

Le lecteur attentif aura rapidement remarqué que la construction des modèles $N$-gram positionnels définie ci-dessus n'est pas optimale. En effet, un certain nombre de $N$-grams positionnels sont redondants. Nous illustrons cette affirmation à partir de l'énoncé suivant dans lequel *des* et *problématiques* sont tour à tour UT pivot.

*Le survol* $\underbrace{\textit{des}}_{\textit{pivot}}$ *problématiques linguistiques sur la composition nominale met en évidence l'absence de critères purement linguistiques permettant aux théoriciens de délimiter l'ensemble des " noms composés " .*

*Le survol des* $\underbrace{\textit{problématiques}}_{\textit{pivot}}$ *linguistiques sur la composition nominale*

*met en évidence l'absence de critères purement linguistiques permettant aux théoriciens de délimiter l'ensemble des " noms composés " .*

Après une lecture séquentielle de l'énoncé, les 2-grams positionnels suivants sont construits pour un environnement immédiat de taille 2.

| UT pivot *des* | UT pivot *problématiques* |
|---|---|
| [*0 des -2 Le*] | [*0 problématiques -2 survol*] |
| [*0 des 1 problématiques*] | [*0 problématiques -1 des*] |
| [*0 des -1 survol*] | [*0 problématiques 1 linguistiques*] |
| [*0 des 2 linguistiques*] | [*0 problématiques 2 sur*] |

TAB. 4.9 – Exemple de 2-grams positionnels répétés

Il apparaît clairement que les deux 2-grams positionnels de la deuxième ligne du tableau sont redondants puisqu'ils représentent la même séquence de l'énoncé — *des problématiques.*

### Contexte Droit

Dans le but d'éviter les redondances créées par l'algorithme proposé dans la partie précédente, nous avons défini un nouvel algorithme de construction des modèles $N$-gram positionnels. L'idée de base est de ne construire que les $N$-grams positionnels non redondants pour chaque UT de l'énoncé. Pour nous en assurer, les fenêtres gauche et droite de taille $F$ sont regroupées en une seule fenêtre droite — contexte droit — qui se déplace séquentiellement tout au long de l'énoncé et couvre le même espace que les deux fenêtres gauche et droite[5]. Nous illustrons cette situation dans l'exemple suivant.

*Le survol* (*Fenêtre gauche*) *des* (*pivot*) *problématiques linguistiques* (*Fenêtre droite*) *sur la composition nominale*
*met en évidence l'absence de critères purement linguistiques permettant aux théoriciens de délimiter l'ensemble des " noms composés " .*

Environnement à fenêtre droite et gauche

[5]Seuls les cas limites — début et fin d'énoncé, ne seront pas traités de la même façon par les deux méthodes. Nous négligerons ces cas limites pour des énoncés de grande taille.

$\underbrace{Le}_{pivot}$ $\underbrace{\boxed{survol\ des\ problématiques\ linguistiques}}_{Contexte\ \ droit}$ *sur la composition nominale met en évidence l'absence de critères purement linguistiques permettant aux théoriciens de délimiter l'ensemble des " noms composés " .*

Environnement à contexte droit

La taille du contexte droit vaudra donc $F$ dans le cas du modèle 2-gram positionnel et $2.F$ dans tous les autres cas — $F$ étant la taille de l'environnement immédiat considéré. L'exemple suivant montre cette situation pour un environnement immédiat de taille 2. Pour le calcul des 2-grams positionnels, le contexte droit de chaque UT pivot regroupera donc 2 UTs et 4 UTs pour le calcul des $N$-grams positionnels tels que $N > 2$.

$\underbrace{Le}_{pivot}$ $\boxed{survol\ des}$ *problématiques linguistiques sur la composition nominale met en évidence l'absence de critères purement linguistiques permettant aux théoriciens de délimiter l'ensemble des " noms composés " .*

Contexte droit de taille 2 pour les 2-grams positionnels

*Le* $\underbrace{survol}_{pivot}$ $\boxed{des\ problématiques\ linguistiques\ sur}$ *la composition nominale met en évidence l'absence de critères purement linguistiques permettant aux théoriciens de délimiter l'ensemble des " noms composés " .*

Contexte droit de taille 4 pour les $N$-grams positionnels, $N > 2$

Du fait de la définition du contexte droit qui s'étend uniquement en avant de l'UT pivot considérée, le changement séquentiel de l'UT pivot assure l'unicité de chaque $N$-gram positionnel construit. En effet, lors du changement d'UT pivot, l'UT pivot antérieure sort du contexte immédiat de la nouvelle UT pivot. Ainsi, aucun $N$-gram contenant l'ancienne UT pivot ne peut être construit. Notre attention doit maintenant se tourner vers la construction de l'ensemble des $N$-grams positionnels qu'il est possible de construire dans ce nouveau contexte.

**Implémentation**

Toutes les combinaisons de $N-1$ UTs du contexte droit ne forment pas forcément un $N$-gram positionnel lorsqu'elles sont associées à l'UT pivot. En effet, seules les combinaisons qui seraient possibles dans un environnement à fenêtre gauche et droite sont calculées. Afin d'éclaircir nos propos, nous appliquons cette règle sur l'exemple suivant.

> *Le survol* $\underbrace{\textit{des}}_{pivot}$ $\boxed{\textit{problématiques linguistiques sur la}}$ *composition nominale met en évidence l'absence de critères purement linguistiques permettant aux théoriciens de délimiter l'ensemble des " noms composés " .*

Contexte droit de taille $2.F = 4$

Si nous considérons le calcul d'un 3-gram positionnel pour l'UT pivot *des* dans un contexte droit de taille $2.F = 4$, une combinaison possible de 2 UTs dans le contexte droit peut amener à la construction du 3-gram positionnel [*0 des 3 sur 4 la*] qui n'est pas un $N$-gram positionnel valide. En effet, sa construction est impossible à partir de deux fenêtres de taille $F = 2$, l'une gauche et l'autre droite. De fait, aucune des trois UTs — *des*, *sur*, *la* — ne peut jouer le rôle d'UT pivot et compter dans son environnement immédiat les deux autres UTs. Les trois cas possibles sont énumérés ci-après.

> $\boxed{\textit{Le survol}}$ $\underbrace{\textit{des}}_{pivot}$ $\boxed{\textit{problématiques linguistiques}}$ *sur la composition nominale met en évidence l'absence de critères purement linguistiques permettant aux théoriciens de délimiter l'ensemble des " noms composés " .*

Les UTs *sur* et *la* ne sont pas comprises dans l'environnement immédiat

> *Le survol des* $\boxed{\textit{problématiques linguistiques}}$ $\underbrace{\textit{sur}}_{pivot}$ $\boxed{\textit{la composition}}$ *nominale met en évidence l'absence de critères purement linguistiques permettant aux théoriciens de délimiter l'ensemble des " noms composés " .*

L'UT *des* n'est pas comprise dans l'environnement immédiat

> *Le survol des problématiques* $\boxed{\textit{linguistiques sur}}$ $\underbrace{\textit{la}}_{pivot}$ $\boxed{\textit{composition nominale}}$ *met en évidence l'absence de critères purement linguis-*

*tiques permettant aux théoriciens de délimiter l'ensemble des " noms composés " .*

L'UT *des* n'est pas comprise dans l'environnement immédiat

Plus généralement, à partir de l'espace constitué par l'UT pivot et son contexte droit, seules les combinaisons de $N$ UTs qui peuvent être calculées pour un environnement immédiat à deux fenêtres sont retenues comme $N$-grams positionnels valides.

Dans ce cadre, nous devons considérer chaque UT du contexte droit comme étant une UT pivot potentielle pour un environnement à deux fenêtres. Chaque UT du contexte droit est donc tour à tour utilisée comme UT de traitement pour le calcul des $N$-grams positionnels possibles dans un environnement à contexte droit. Cette UT de traitement représente symboliquement une UT pivot potentielle pour un environnement à deux fenêtres.

Par exemple, si la première UT du contexte droit est l'UT de traitement de l'environnement à deux fenêtres, les $N$-grams positionnels valides sont les combinaisons de $N$ UTs contenant obligatoirement l'UT pivot et l'UT de traitement dans un contexte de $F$ UTs à droite et un contexte d'une seule UT à gauche de l'UT de traitement — dans ce cas l'UT pivot. Exemplifions cette situation à partir du cas suivant pour un contexte droit de taille 4 — i.e. $F = 2$ — et pour le calcul des 3-grams positionnels.

*Le* $\underbrace{\textit{survol}}_{\textit{pivot}}$ $\boxed{\textit{des problématiques linguistiques sur}}$ *la composition nominale met en évidence l'absence de critères purement linguistiques permettant aux théoriciens de délimiter l'ensemble des " noms composés " .*

Si *des* est l'unité de traitement, alors seuls les 3-grams positionnels pouvant être calculés dans l'environnement immédiat suivant et contenant *survol* et *des* sont des 3-grams valides.

*Le* $\boxed{\textit{survol}}$ $\underbrace{\textit{des}}_{\textit{traitement}}$ $\boxed{\textit{problématiques linguistiques}}$ *la composition nominale met en évidence l'absence de critères purement linguistiques permettant aux théoriciens de délimiter l'ensemble des " noms composés " .*

Nous définissons génériquement cette situation ci-après. Considérons un énoncé de $T$ occurrences $A_i$ — $i$ variant de 1 à $T$, un contexte droit de taille $2.F$ et l'UT pivot $A_1$.

$$A_1 \boxed{A_2\ A_3\ A_4\ \ldots A_{F+2}\ A_{F+3}\ A_{F+4}\ \ldots A_{2.F+1}}\ A_{2.F+2}\ \ldots A_T$$

Enoncé et Contexte droit de taille $2.F$

Le tableau 4.10 montre l'ensemble des UTs de traitement potentielles ainsi que les fenêtres droite et gauche associées à chaque UT de traitement. Chaque couche correspond alors, pour une UT de traitement donnée, au contexte à prendre en compte pour le calcul des $N$-grams positionnels. On rappelle que l'UT pivot — $A_1$— doit toujours être comprise dans l'environnement considéré. On notera également que le trait "$-$" représente une UT possible pour le calcul combinatoire et que l'UT limite correspond à l'UT à partir de laquelle il n'est plus possible de calculer un $N$-gram positionnel pour l'UT pivot $A_1$ et l'UT de traitement $A_j$ — $j$ variant de 2 à $F+1$.

| Couche | Contexte à fenêtre gauche et droite | UT limite |
|---|---|---|
| 1 | $\underbrace{A_1}_{1}\ A_2\ \underbrace{- \cdots -}_{F}$ | $A_{F+2}$ |
| 2 | $\underbrace{A_1\ -}_{2}\ A_3\ \underbrace{- \cdots -}_{F}$ | $A_{F+3}$ |
| 3 | $\underbrace{A_1\ -\ -}_{3}\ A_4\ \underbrace{- \cdots -}_{F}$ | $A_{F+4}$ |
| ... | ... ... ... | |
| F | $\underbrace{A_1\ -\ - \cdots -}_{F}\ A_{F+1}\ \underbrace{- \cdots -}_{F}$ | $A_{2F+1}$ |

TAB. 4.10 – Espace des combinaisons de $N$ UTs

L'ensemble des $N$-grams positionnels peut donc être calculé de la forme suivante à partir de la lecture séquentielle d'un énoncé. Chaque UT de l'énoncé est tour à tour UT pivot et le contexte droit — de taille $F$ pour le modèle 2-gram positionnel et $2.F$ pour les autres modèles — définit son environnement immédiat. Chaque $N$-gram positionnel correspond alors à une combinaison spécifique de $N-1$ UTs calculée dans le contexte droit de l'UT pivot et combinée à l'UT pivot pour former un $N$-gram positionnel. Une combinaison

n'est valide qu'à la condition d'être une combinaison possible de l'une des $F$ couches mises en évidence dans la figure 4.10 pour un environnement à fenêtre gauche et droite de taille $F$.

Une remarque importante à souligner est le fait que les positions qui indexent chaque UT à l'UT pivot seront toutes positives. En effet, le calcul des $N$-grams positionnels se fait uniquement par la droite de l'UT pivot.

**Complexité**

Pour chaque UT pivot, pour un $N$ et un environnement immédiat $F$ donnés, il est possible de construire $K$ $N$-grams positionnels. $K$ peut être calculé grâce à l'équation 4.8 où l'indice $i$ correspond à l'ensemble des UTs de traitement possibles et l'indice $j$ à l'ensemble des UTs qu'il est possible de combiner avec l'UT pivot et l'UT de traitement. En fait, $K$ représente la somme de tous les $N$-grams positionnels qu'il est possible de calculer à partir de chacune des $F$ couches définies antérieurement et ceci en éliminant les doublons qui surgissent d'une couche à l'autre.

$$K = \begin{cases} N = 1, & 1 \\ N = 2, & F \\ 2 < N \leq 2.F + 1, & \sum_{i=1}^{F} \sum_{j=1}^{F} \mathrm{C}_{j-1}^{i-1} \times \mathrm{C}_{j}^{N-i-1} \end{cases} \tag{4.8}$$

Du fait de la définition constructive de cette équation, il est nécessaire de pauser un ensemble de conditions $R$ qui assure la correction des calculs pour le cas de $2 < N \leq 2.F+1$. En effet, dans ce cas, tous les calculs ne sont pas toujours possibles. Ainsi, nous définissons $R$ de la forme suivante où $\sigma$ est un terme spécifique de la somme $\sum_{i=1}^{F} \sum_{j=1}^{F} \mathrm{C}_{j-1}^{i-1} \times \mathrm{C}_{j}^{N-i-1}$, pour un $i$ et un $j$ donnés. L'objectif est ainsi de garantir la faisabilité de l'expression $\mathrm{C}_{j-1}^{i-1} \times \mathrm{C}_{j}^{N-i-1}$ qui correspond au $\sigma$.

$$R = \begin{cases} si & \begin{cases} i > j \quad \vee \\ N > j + i + 1 \quad \vee \\ N < i + 1, \end{cases} & \sigma = 0 \\ \text{sinon} & & \sigma = \mathrm{C}_{j-1}^{i-1} \times \mathrm{C}_{j}^{N-i-1} \end{cases} \tag{4.9}$$

Ainsi, les valeurs de $K$ doivent toujours être considérées en tenant compte de l'ensemble des contraintes $R$. Dans la suite des calculs, nous considérerons que l'ensemble $R$ doit

toujours être satisfait pour la correction des démonstrations.

A partir des résultats précédents, l'espace des données $E''$ qui représente l'ensemble des $N$-grams positionnels qu'il est possible de construire pour un environnement immédiat $F$ et une taille d'énoncé $T$ donnés, peut être formulé par l'équation 4.10.

$$\begin{aligned} E'' &= (T - 2.F) \times 1 + \\ &\quad (T - 2.F) \times F + \\ &\quad (T - 2.F) \times \sum_{N=3}^{2.F+1} \sum_{i=1}^{F} \sum_{j=1}^{F} \mathrm{C}_{j-1}^{i-1} \times \mathrm{C}_{j}^{N-i-1} \end{aligned} \tag{4.10}$$

Ainsi, si l'on note $a$ la somme suivante,

$$a = 1 + F + \sum_{N=3}^{2.F+1} \sum_{i=1}^{F} \sum_{j=1}^{F} \mathrm{C}_{j-1}^{i-1} \times \mathrm{C}_{j}^{N-i-1} \tag{4.11}$$

$E''$ peut s'exprimer de la forme suivante.

$$E'' = (T - 2.F) \times a \tag{4.12}$$

En ce qui concerne la complexité des calculs, si l'on considère $T$ comme l'unique variable de notre système, celle-ci est linéaire en fonction de $T$. En effet, $F$ étant alors une constante, le nombre de $N$-grams positionnels qui peuvent être construits à partir du nouvel algorithme est proportionnel à la taille de l'énoncé. Dans ces conditions, si $a$ est défini comme le coefficient d'accélération de $E''$, la complexité de calcul des modèles $N$-gram positionnels est $\mathcal{O}(a.T)$. A titre d'exemple, les valeurs de $a$ correspondant aux différentes valeurs de $F$ envisagées sont énumérées dans le tableau 4.11.

En comparaison avec la première implémentation des modèles $N$-gram positionnels, le nouvel algorithme démontre un facteur d'accélération plus faible que celui de l'algorithme précédent. En effet, le nouvel algorithme réduit d'environ un tiers le nombre de $N$-grams positionnels construits selon l'environnement considéré. Le tableau 4.12, dans lequel $c$

| Taille $F$ | Constante $a$ |
|---|---|
| 1 | 4 |
| 2 | 11 |
| 3 | 43 |
| 4 | 171 |
| 5 | 683 |

TAB. 4.11 – Accélération de la constante $a$

et $a$ sont respectivement les coefficients d'accélération de la première et la deuxième implémentation, illustre cette affirmation.

| Taille $F$ | Réduction (%) |
|---|---|
| 1 | 0 |
| 2 | 31 |
| 3 | 32 |
| 4 | 33 |
| 5 | 33 |

TAB. 4.12 – Réduction du coefficient d'accélération

Cependant, si l'on considère la taille de l'environnement immédiat $F$ comme une variable du système, il est nécessaire de prendre en compte ses variations. Dans ce cas, pour chaque UT de l'énoncé, il est nécessaire de calculer tous les $N$-grams positionnels pour chacun des environnements immédiats possibles. Par conséquent, l'ensemble $E'''$ de tous les $N$-grams positionnels possibles, ceci quelles que soient les tailles de l'énoncé et de l'environnement immédiat considéré, est donné par l'équation 4.13.

A partir des fondements de la théorie de la complexité, nous calculons dans un premier temps la complexité d'ordre $\mathcal{O}$ de $E'''$ qui peut être considérée comme une fonction $f'$ dont les arguments sont $T$ et $F$.

$$E''' = \sum_{P=1}^{F} ((T - 2.P) \times 1) +$$

$$\sum_{P=1}^{F}((T-2.P)\times P)+$$
$$\sum_{P=1}^{F}\left((T-2.P)\times\sum_{N=3}^{2.P+1}\sum_{i=1}^{P}\sum_{j=1}^{P}\mathrm{C}_{j-1}^{i-1}\times\mathrm{C}_{j}^{N-i-1}\right) \tag{4.13}$$

**Complexité d'ordre $\mathcal{O}$ :** Nous allons formuler une borne supérieure de $E'''$ à partir de la notion de "couche" qui est inhérente à la construction des $N$-grams positionnels à partir du nouvel algorithme. Un majorant évident de $E'''$ peut être déterminé à partir de l'ensemble des $N$-grams positionnels qu'il est possible de calculer pour chacune des $F$ couches mises en évidence dans le tableau 4.10. Dans ce cadre, nous allons d'abord proposer un majorant de $K$ qui nous permettra de majorer à son tour le coefficient d'accélération $a$ qui lui-même servira de majorant de $E''$ et par voie de conséquence de $E'''$.

Nous définissons dans un premier temps un majorant de $K$ qui correspond au nombre de $N$-grams positionnels qu'il est possible de calculer pour un environnement immédiat $F$ et un $N$ donnés. Si l'on note $S$ la somme correspondant à l'ensemble des $N$-grams positionnels qu'il est possible de construire à partir des $F$ couches d'un contexte droit, $S$ peut être définie par l'équation 4.14.

$$S = \begin{cases} N=1, & 1 \\ N=2, & F \\ \forall N, 2<N\leq 2.F+1, & \begin{cases} F\geq N-2, & \sum_{i=F}^{2.F-1}\mathrm{C}_{i}^{N-2} \\ F<N-2, & \sum_{i=N-2}^{2.F-1}\mathrm{C}_{i}^{N-2} \end{cases} \end{cases} \tag{4.14}$$

Dans ces conditions, plusieurs combinaisons de $N$ UTs calculées pour chacune des $F$ couches d'une UT pivot textuelle sont redondantes. En effet, par exemple, entre la couche 1 et la couche 2 du tableau 4.10, toutes les combinaisons qui contiennent à la fois $A_1$, $A_2$ et $A_3$ et qui ne contiennent pas d'UTs se trouvant dans le contexte droit de $A_{F+2}$ — UT limite de la couche 1 — sont des doublons. Par conséquent, $S$ est nécessairement supérieure à $K$, le nombre de $N$-grams positionnels réellement construits.

A partir de ce résultat, nous déterminons dans un deuxième temps un majorant de $E''$ qui correspond au nombre de $N$-grams positionnels qui sont construits pour chaque UT

pour un environnement immédiat $F$ donné et pour un $N$ variant de 1 à 2.F+1. Dans le cas où $2 < N \leq 2.F + 1$, $S$ est une somme dont la borne supérieure vaut $2.F - 1$ et par conséquent peut être majorée par le terme $(2.F-1) \times \mathrm{C}_{2.F-1}^{N-2}$. Ainsi, une borne supérieure du coefficient d'accélération $a$ de $E''$ peut être formulée de la forme suivante.

$$a \leq 1 + F + \sum_{N=3}^{2.F+1} ((2.F-1) \times \mathrm{C}_{2.F-1}^{N-2}) \tag{4.15}$$

A partir de cette inéquation, $E''$ peut donc être majoré comme l'inéquation 4.16 le démontre.

$$\begin{aligned} E'' &\leq (T-2.F) \times \left( \sum_{N=3}^{2.F+1} ((2.F-1) \times \mathrm{C}_{2.F-1}^{N-2}) + F + 1 \right) \\ &\leq (T-2.F) \times \left( (2.F-1) \times \sum_{N=3}^{2.F+1} \mathrm{C}_{2.F-1}^{N-2} + F + 1 \right) \\ &\leq (T-2.F) \times \left( (2.F-1) \times \sum_{N'=1}^{2.F-1} \mathrm{C}_{2.F-1}^{N'} + F + 1 \right) \\ &\leq (T-2.F) \times \left( (2.F-1) \times (2^{2.F-1} - 1) + F + 1 \right) \end{aligned} \tag{4.16}$$

A partir de cette inéquation, nous définissons finalement un majorant de $E'''$. En effet, il suffit d'introduire l'argument variable $F$ dans l'inéquation 4.16 pour obtenir le résultat suivant.

$$\begin{aligned} E''' &\leq \sum_{P=1}^{F} \left( (T-2.P)((2.P-1) \times (2^{2.P-1} - 1) + P + 1) \right) \\ &\leq \sum_{P=1}^{F} \left( (T-2.P)(2.P \times 2^{2.P-1} + P + 2) \right) \\ &\leq \sum_{P=1}^{F} \left( T \times (2.P \times 2^{2.P-1} + P + 2) \right) \\ &\leq T \times \sum_{P=1}^{F} (2.P \times 2^{2.P-1} + P + 2) \\ &\leq T \times \left( \sum_{P=1}^{F} (2.P \times 2^{2.P-1}) + \sum_{P=1}^{F} P + \sum_{P=1}^{F} 2 \right) \\ &\leq T \times \sum_{P=1}^{F} (2.P \times 2^{2.P-1}) + \frac{(F+1).F.T}{2} + 2.F.T \end{aligned}$$

$$\begin{aligned} &\leq T \times (F.(2.F).2^{2.F-1}) + \frac{(F+1).F.T}{2} + 2.F.T \\ &\leq \frac{T}{2} \times (2.F^2.2^{2.F}) + \frac{T}{2} \times ((F+1).F) + \frac{T}{2} \times (4.F) \\ &\leq \frac{T}{2} \times \big((2.F^2.2^{2.F}) + ((F+1).F) + (4.F)\big) \end{aligned} \tag{4.17}$$

Ainsi, si l'on considère $g''(T,F) = \frac{T}{2} \times \big((2.F^2.2^{2.F}) + ((F+1).F) + (4.F)\big)$, $f'(T,F)$ est une fonction qui croît plus lentement que $g''(T,F)$ c'est-à-dire $f'(T,F) = \mathcal{O}(g''(T,F))$. Cependant, $g''(T,F)$ démontre un facteur de croissance exponentielle si l'on prend en compte le terme $2^{2.F}$ qui la compose. Nous devons donc vérifier si la définition de croissance exponentielle s'applique au cas qui nous intéresse. Dans ce cadre, nous cherchons la complexité d'ordre $\Omega$ de la fonction $f'(T,F)$.

**Complexité d'ordre $\Omega$ :** Parallèlement à ce qui a été fait dans le cadre de la borne supérieure de $E'''$, nous proposons dans un premier temps un minorant évident de $E''$ pour ensuite calculer un minorant de $E'''$. Pour se faire, nous proposerons d'abord un minorant de $K$, puis un de $a$ pour conclure avec un minorant de $E''$ qui nous permettra de mettre en évidence le minorant de $E''''$ attendu.

Si l'on considère la somme $S'$ qui correspond à l'ensemble des $N$-grams positionnels qui peuvent être construits à partir de la seule couche 1 — seule la première UT du contexte droit est prise en compte pour le calcul des $N$-grams positionnels possibles —, $S'$ peut être formulée de la forme suivante.

$$S' = \begin{cases} N = 1, & 1 \\ N = 2, & F \\ \forall N, 2 < N \leq F+2, & \mathrm{C}_F^{N-2} \end{cases} \tag{4.18}$$

Dans ces conditions, les combinaisons de $N$ UTs qu'il est possible de construire à partir de la couche 1 ne forment évidemment qu'un sous-ensemble de toutes les combinaisons possibles. En effet, à partir du tableau 4.10, il est clair que toutes les combinaisons qui contiennent les UTs de $A_{F+3}$ à $A_{2.F+1}$ ne sont pas contenues dans $S'$. De plus, aucune combinaison n'est répétée. La somme $S'$ est donc strictement inférieure à K, le nombre de $N$-grams positionnels réellement construits. Ainsi, on propose une borne inférieure du coefficient d'accélération $a$.

$$a > 1 + F + \sum_{N=3}^{F+2} \mathrm{C}_F^{N-2} \tag{4.19}$$

A partir de cette inéquation, nous définissons facilement une borne inférieure de $E''$ pour ensuite déduire le minorant de $E'''$. Ainsi, $E''$ peut être minoré de la forme suivante.

$$\begin{aligned} E'' &> (T - 2.F) \times \left(1 + F + \sum_{N=3}^{F+2} \mathrm{C}_F^{N-2}\right) \\ &> (T - 2.F) \times \left(1 + F + \sum_{N'=1}^{F} \mathrm{C}_F^{N'}\right) \\ &> (T - 2.F) \times \left(1 + F + 2^F - 1\right) \\ &> (T - 2.F) \times \left(F + 2^F\right) \end{aligned} \tag{4.20}$$

Sachant, que $E''$ peut être minoré de cette forme, nous déduisons facilement un minorant pour $E'''$. En effet, à partir de l'inéquation 4.20, nous introduisons dans $E''$ l'argument variable $F$. On obtient ainsi l'inéquation 4.21 pour laquelle on notera sans problème que le facteur de pondération $T - 2.P$ est strictement supérieur à 1.

$$\begin{aligned} E''' &> \sum_{P=1}^{F} \left((T - 2.P) \times (P + 2^P)\right) \\ &> \sum_{P=1}^{F} (P + 2^P) \\ &> \sum_{P=1}^{F} P + \sum_{P=1}^{F} 2^P \\ &> \frac{F.(F+1)}{2} + \sum_{P=1}^{F} 2^P \\ &> \frac{F.(F+1)}{2} + 2^{F+1} - 2 \\ &> \frac{F^2}{2} + \frac{F}{2} + 2^{F+1} - 2 \end{aligned} \tag{4.21}$$

Par conséquent, si l'on considère $E'''$ une fonction $f'$ ayant comme arguments $T$ et $F$, $g'''(F) = \frac{F^2}{2} + \frac{F}{2} + 2^{F+1} - 2$ est une fonction telle que $f'(T, F) = \mathcal{O}(g'''(F))$ est faux. Dans ces conditions, on notera $f'(T, F) = \Omega(g'''(F))$. Nous nous trouvons par conséquent

dans une situation où la fonction $f'(T, F)$ croît exponentiellement en fonction de la taille de l'environnement et linéairement en fonction de la taille de l'énoncé.

La nouvelle implémentation proposée ne modifie donc pas les caractéristiques de complexité mises en évidence par la première implémentation. Cependant, comme nous l'avons déjà vu, les coefficients d'accélération des calculs sont largement réduits par le nouvel algorithme.

Afin de rendre compte de la tâche que représente le calcul des modèles $N$-gram positionnels, nous proposons dans le tableau 4.13 le nombre de $N$-grams positionnels qui sont construits à partir d'un énoncé de 343 992 UTs pour un environnement immédiat de taille 3 dans le cas des deux implémentations. Dans ce contexte, nous définissons la notion de gain par le quotient entre le nombre de $N$-grams positionnels économisés et le nombre maximum de $N$-grams positionnels construits.

| $N$-gram | Méthode simple | Méthode Optimisée | Gain (%) |
|---|---|---|---|
| 1-gram | 343 992 | 343 992 | 0 |
| 2-gram | 2 063 916 | 1 031 967 | 50 |
| 3-gram | 5 159 790 | 3 095 874 | 40 |
| 4-gram | 6 879 720 | 4 471 818 | 35 |
| 5-gram | 5 159 790 | 3 783 846 | 27 |
| 6-gram | 2 063 916 | 1 719 930 | 17 |
| 7-gram | 343 986 | 343 986 | 0 |

TAB. 4.13 – Comparaison des méthodes de calcul

L'implémentation des modèles $N$-gram positionnels n'est donc pas une tâche facile en ce qui concerne la gestion du nombre de données construites. Comme nous le verrons dans l'une des parties suivantes, plusieurs impératifs devront être pris en compte pour simplifier cette tâche. Mais avant d'aller plus loin dans nos propos, nous définissons dans la section suivante un certain nombre de propriétés qui nous permettent en particulier d'exprimer l'ensemble des $N$-grams positionnels construits à partir de la nouvelle implémentation de la forme générique déjà définie.

### 4.3.7 Propriétés

L'indexation de chaque UT à une UT pivot permet de déduire un ensemble de propriétés intéressantes. Parmi celles-ci, nous nous attardons dans un premier temps sur la notion de changement d'UT référentielle — UT pivot — dans un $N$-gram positionnel.

#### Unité Textuelle de Référence

Du fait de l'utilisation explicite des positions, il est facile d'exprimer un $N$-gram positionnel en fonction d'une nouvelle unité de référence. Les positions doivent simplement être ajustées à la nouvelle UT pivot. En particulier, cette propriété permet de transformer l'ensemble des $N$-grams positionnels construits à partir du nouvel algorithme en l'ensemble des $N$-grams positionnels qui seraient calculés à partir de la première implémentation.

Supposons un $N$-gram positionnel générique noté de la forme suivante

$$[p_{11}\ u_1\ p_{12}\ u_2\ \ldots p_{1i}\ u_i\ \ldots p_{1N}\ u_N]$$

et $u_t$ une UT quelconque du $N$-gram positionnel — $t$ variant de 2 à $N$, chaque $u_t$ peut être UT pivot d'un nouvel $N$-gram positionnel noté de la forme suivante où $p_{tt}$ vaut zéro et pour $\forall u_i$, telle que $i = 1..N$, $p_{ti} = p_{1i} - p_{1t}$.

$$[p_{tt}\ u_t\ p_{t1}\ u_1\ p_{t2}\ u_2\ \ldots p_{t(t-1)}\ u_{t-1}\ p_{t(t+1)}\ u_{t+1}\ \ldots p_{tN}\ u_N]$$

Nous illustrons cette situation à partir du 3-gram positionnel [*0 des 1 problématiques 2 linguistiques*]. Tous ses 3-grams positionnels équivalents — chacune des deux UTs non pivot étant tour à tour pivot — sont énumérés dans le tableau 4.14.

| 3-grams positionnels équivalents |
|---|
| [*0 problématiques -1 des 1 linguistiques*] |
| [*0 linguistiques -2 des -1 problématiques*] |

TAB. 4.14 – Changement d'UT pivot

Cette propriété est essentielle à la compréhension d'un certain nombre de notations que nous utiliserons au long de notre exposé. Dans la partie suivante, nous introduisons une autre propriété importante pour la notion de sous-groupe complémentaire que nous verrons dans la suite de ce chapitre : le changement de position de référence.

**Position de Référence**

Il est également facile d'exprimer les positions d'un $N$-gram positionnel en fonction d'une nouvelle position de référence. Les positions doivent simplement être ajustées à la nouvelle position de référence. En effet, nous avons vu qu'il convient d'exprimer la position de l'UT pivot comme valant zéro. Or, les opérations que l'on effectue sur les $N$-grams positionnels peuvent amener à considérer la position de l'UT pivot comme étant non nulle. Dans ce cas, il est souhaitable de fixer la position de l'UT de référence à zéro et d'ajuster toutes les autres positions du $N$-gram positionnel en fonction de ce changement.

Un $N$-gram positionnel calculé pour une UT pivot dont la position de référence n'est pas nulle peut être exprimé en fonction de la même UT pivot pour une position nulle en soustrayant à chacune des positions initiales la valeur de la position non nulle de l'UT pivot. Supposons un $N$-gram positionnel générique noté de la forme suivante.

$$[p_{11}\ u_1\ p_{12}\ u_2\ \ldots p_{1i}\ u_i\ \ldots p_{1N}\ u_N]$$

Si, la position de l'UT pivot $u_1$ notée $p_{11}$ n'est pas nulle, le $N$-gram positionnel suivant est équivalent au précédent au changement de position de référence près.

$$[p'_{11}\ u_1\ p'_{12}\ u_2\ \ldots p'_{1i}\ u_i\ \ldots p'_{1N}\ u_N], \forall j, j = 1..N, p'_{1j} = p_{1j} - p_{11}$$

Nous illustrons cette situation à partir des deux 3-grams positionnels suivants qui peuvent être considérés comme équivalents au changement de position de référence près.

[ *-2 des -1 problématiques 0 linguistiques*]

[*0 des 1 problématiques 2 linguistiques*]

Dans la partie suivante, nous introduisons respectivement les notions de sous-groupes, sous-groupes complémentaires et sur-groupes de $N$-grams positionnels qui utilisent ces deux dernières propriétés. Ces notions seront particulièrement importantes pour la définition des mesures d'association que nous verrons dans le prochain chapitre.

**Sous-groupes**

Dans les parties précédentes, nous avons défini que l'occurrence d'une UT est dictée par un ensemble d'UTs présentes dans son environnement immédiat. Afin de mesurer ces cooccurrences d'UTs, il est nécessaire de connaître les comptages de tous les sous-ensembles d'UTs d'un $N$-gram positionnel. Chaque ensemble sera appelé sous-groupe et suivant le nombre d'UTs qui le constitueront, il dénotera un ordre spécifique. Nous expliquons la notion de sous-groupe d'UTs dans la définition suivante.

> *Un* ***sous-groupe*** *d'un $N$-gram positionnel est un $K$-gram positionnel — $K$ variant de 1 à $N-1$ — contenu dans le $N$-gram positionnel. On appelle ce $K$-gram, sous-groupe de rang $K$.*

Un sous-groupe de rang $K$ d'un $N$-gram positionnel est donc une combinaison particulière de $K$ UTs parmi les $N$ UTs du $N$-gram positionnel considéré — $K$ variant de 1 à $N-1$. Formellement, un $K$-gram positionnel noté $[p'_{11}\ u'_1\ p'_{12}\ u'_2\ \ldots p'_{1K}\ u'_K]$ est un sous-groupe de rang $K$ d'un $N$-gram positionnel noté $[p_{11}\ u_1\ p_{12}\ u_2\ \ldots p_{1N}\ u_N]$ si les trois conditions $C_1$, $C_2$ et $C_3$ sont respectées.

$$C_1 \equiv K < N$$

$$C_2 \equiv \{u'_j | \forall j, j = 1..K\} \subset \{u_i | \forall i, i = 1..N\}$$

$$C_3 \equiv \begin{cases} u'_1 = u_1 & \begin{cases} \forall u'_j, j = 1..K \exists i, u'_j = u_i \Rightarrow p'_{1j} = p_{1i} \end{cases} \\ u'_1 \neq u_1 & \begin{cases} \exists i, u'_1 = u_i \quad \text{tel que} \\ \forall u'_j, j = 1..K \exists l, u'_j = u_l \Rightarrow p'_{1j} = p_{1l} - p_{1i} \end{cases} \end{cases}$$

Nous illustrons cette situation à partir du 3-gram positionnel suivant : [*0 des 1 problématiques 2 linguistiques*]. L'ensemble de ses sous-groupes de rang 1 et 2 sont

énumérés dans le tableau 4.15 où le 2-gram [*0 problématiques 1 linguistiques*] met en évidence la propriété de changement d'UT de référence énoncée précédemment.

| Rang 1 | Rang 2 |
|---|---|
| [*0 des*] | [*0 des 1 problématiques*] |
| [*0 problématiques*] | [*0 des 2 linguistiques*] |
| [*0 linguistiques*] | [*0 problématiques 1 linguistiques*] |

TAB. 4.15 – Ensemble de sous-groupes de rang $K$

**Sous-groupes complémentaires**

Dans le cadre du calcul des cooccurrences entre UTs, il est souvent nécessaire de considérer les sous-groupes d'un $N$-gram positionnel qui, lorsqu'ils sont concaténés, permettent de reconstruire le $N$-gram positionnel de référence. Ces sous-groupes sont alors appelés sous-groupes complémentaires. En particulier, nous nous intéresserons aux groupes de deux sous-groupes qui permettent cette opération. Formellement, considérons deux sous-groupes, l'un de rang $K$ et l'autre de rang $L$, d'un même $N$-gram positionnel noté $[p_{11}\ u_1\ p_{12}\ u_2\ \ldots p_{1N}\ u_N]$. Les deux sous-groupes sont notés de la forme suivante.

$$[p'_{11}\ u'_1\ p'_{12}\ u'_2\ \ldots p'_{1K}\ u'_K]$$

$$[p''_{11}\ u''_1\ p''_{12}\ u''_2\ \ldots p''_{1L}\ u''_L]$$

Ces deux sous-groupes sont dits complémentaires si l'ensemble des conditions $C_1$, $C_2$ et $C_3$ sont respectées.

$$C_1 \equiv K + L = N \wedge K \neq 0 \wedge L \neq 0$$

$$C_2 \equiv \{u'_p | \forall p, p = 1..K\} \cup \{u''_j | \forall j, j = 1..L\} = \{u_i | \forall i, i = 1..N\}$$

$$C_3 \equiv \begin{cases} u'_1 = u_1 & \begin{cases} \forall u'_j, j = 1..K, \quad \exists i, u'_j = u_i \Rightarrow p'_{1j} = p_{1i} \\ \wedge \\ \exists i, u''_1 = u_i \quad \text{tel que} \quad \forall u''_q, q = 1..L \\ \exists j, u''_q = u_j \Rightarrow p''_{1q} = p_{1j} - p_{1i} \end{cases} \\ u''_1 = u_1 & \begin{cases} \forall u''_j, j = 1..L, \quad \exists i, u''_j = u_i \Rightarrow p''_{1j} = p_{1i} \\ \wedge \\ \exists i, u'_1 = u_i \quad \text{tel que} \quad \forall u'_q, q = 1..K \\ \exists j, u'_q = u_j \Rightarrow p'_{1q} = p_{1j} - p_{1i} \end{cases} \end{cases}$$

Nous illustrons cette notion à partir du 3-gram positionnel suivant : [*0 des 1 problématiques 2 linguistiques*]. Dans le tableau 4.16, chaque ligne correspond à un couple de sous-groupes complémentaires de ce 3-gram.

| Sous-groupe de rang 1 | Sous-groupe de rang 2 |
|---|---|
| [*-1 des*] | [*0 problématiques 1 linguistiques*] |
| [*1 problématiques*] | [*0 des 2 linguistiques*] |
| [*2 linguistiques*] | [*0 des 1 problématiques*] |

TAB. 4.16 – Ensemble de sous-groupes complémentaires

Encore une fois, on remarquera l'utilisation de la propriété de changement d'UT de référence dans le cas du premier sous-groupe de rang 2 du tableau 4.16.

### Sur-groupes

De la même façon que nous avons défini le concept de sous-groupe, nous définissons la notion opposée de sur-groupe.

> *Un* ***sur-groupe*** *d'un N-gram positionnel est un K'-gram positionnel qui contient le N-gram positionnel — K' variant de* $N + 1$ *à* $2.F + 1$ *où F est la taille de l'environnement immédiat considéré. On appelle ce K'-gram sur-groupe de rang K'.*

Ainsi, nous déduisons une étroite relation entre les notions de sous-groupe et sur-groupe. En effet, un sous-groupe étant un $K$-gram positionnel — $K$ variant de 1 à $N-1$ — contenu

dans un $N$-gram positionnel, ce dernier constitue alors un sur-groupe de rang $N$ du $K$-gram considéré. La définition formelle d'un sur-groupe peut donc être déduite directement à partir de la définition de sous-groupe énoncée précédemment. Nous illustrons dans le tableau 4.17 quelques-uns des sur-groupes du 2-gram positionnel [*0 des 1 problématiques*] calculés à partir de la situation suivante.

*Le survol* $\underbrace{\textit{des}}_{pivot}$ *problématiques linguistiques sur la composition nominale met en évidence l'absence de critères purement linguistiques permettant aux théoriciens de délimiter l'ensemble des " noms composés " .*

| Sur-groupes dans un environnement de taille 2 |
|---|
| [*0 des 1 problématiques 2 linguistiques*] |
| [*0 des 1 problématiques 3 sur*] |
| [*0 des 1 problématiques 2 linguistiques 3 sur*] |

TAB. 4.17 – Sur-groupes d'un $N$-gram

On remarquera qu'un sur-groupe n'a pas forcément la même UT pivot que son $N$-gram de référence. Dans ce cas, la propriété de changement d'unité de référence devra être appliquée. Par exemple, à partir de la situation suivante, il est possible d'énumérer un certain nombre de sur-groupes du $N$-gram positionnel considéré précédemment, [*0 des 1 problématiques*].

*Le* $\underbrace{\textit{survol}}_{pivot}$ *des problématiques linguistiques sur la composition nominale met en évidence l'absence de critères purement linguistiques permettant aux théoriciens de délimiter l'ensemble des " noms composés " .*

| Sur-groupes dans un environnement de taille 2 |
|---|
| [*0 survol 1 des 2 problématiques*] |
| [*0 survol 1 des 2 problématiques 3 linguistiques*] |

TAB. 4.18 – Sur-groupes d'un $N$-gram avec changement d'UT pivot

L'ensemble des propriétés que nous avons énoncées est particulièrement important pour la compréhension des opérations de normalisation que nous verrons dans les deux cha-

pitres suivants. Mais avant d'aller plus loin dans notre argumentation, nous abordons le concept de codification du matériel textuel qui joue une rôle extrêmement important dans le développement d'applications en statistique textuelle. Notamment, la codification permet de diminuer l'ampleur de la tâche qui consiste à gérer l'ensemble des données engendrées par les modèles $N$-gram positionnels.

## 4.4 Codification

La mise en oeuvre des traitements informatisés revêt une importance de plus en plus grande dans le domaine du traitement automatique des langues naturelles. D'une part, le volume des textes disponibles en format électronique ne cesse de croître avec l'avènement de la globalisation de l'usage de l'outil informatique et de la toile — réseau internet. D'autre part, les travaux récents en statistique textuelle louent les avantages de l'utilisation de corpora de grandes tailles[6]. Barkema [73] illustre bien ces propos en prodiguant cette affirmation.

> "*[...] un corpus d'un million de mots est bien trop restreint pour étudier la flexibilité [des expressions toutes faites] et [...] un corpus de 20 millions de mots est trop petit pour trouver un nombre suffisant d'occurrences de toutes les expressions [idiomatiques].*"

Dans le tableau 4.19, nous mentionnons à titre d'exemple un ensemble de corpora de référence associés à leur taille, langue et année de compilation.

| Nom | Taille (en $\overline{\text{M}}$ de mots) | Langue | Année |
|---|---|---|---|
| Brown | 1 | Anglais | 1979 |
| BNC | 100 | Anglais | 1996 |
| PAROLE | 20 | Français | 2000 |
| Le Monde | 80 | Français | 1995 |
| CETEMPúblico | 180 | Portugais | 2000 |

TAB. 4.19 – Corpora de référence

[6]Nous utilisons la notion de corpus dans sa plus grande acception c'est-à-dire regroupant les concepts de collections de textes organisées [53] et collections qui ne nécessitent pas de sélection.

Cette course à l'utilisation de collections de textes toujours plus volumineuses pose des problèmes évidents d'ingénierie des logiciels — génie logiciel. L'utilisation d'ordinateurs toujours plus puissants et de mémoires à capacités toujours plus grandes ne suffit pas seule à traiter les problèmes de stockage des données textuelles. De nombreux efforts doivent être fournis dans le but de mieux codifier et organiser l'information disponible. Par exemple, le calcul de tous les modèles $N$-gram positionnels pour un corpus de 100 millions d'UTs est une tâche dantesque si aucun effort n'est employé pour réduire la taille des UTs.

La codification des données est une forme de compression informatique qui consiste à faire abstraction, pendant l'étape des calculs, de la représentation des UTs décomptées pour ne retenir qu'un code qui sera associé à toutes les occurrences d'une même UT. Ces codes sont stockés dans un dictionnaire de formes qui permet à l'issue des calculs de reconstituer le graphisme des formes du texte. Le but de la codification est donc de retenir des codes faciles à manipuler pour le traitement informatique et optimaux pour le stockage informatique. La plupart des méthodes de codification privilégient l'aspect manipulation au détriment du stockage. Dans le cadre des modèles $N$-gram positionnels, nous définissons une méthode de codification alphanumérique qui permet d'optimiser l'espace nécessaire à la représentation des UTs. En effet, du fait de la complexité de calcul des modèles $N$-gram positionnels, il est primordial de réduire la taille de l'énoncé par la compression des UTs afin de diminuer l'espace nécessaire au stockage de l'ensemble des $N$-grams positionnels. Avant d'aller au-delà dans nos propos, nous remarquons que la codification des données textuelles ne s'applique qu'aux formes graphiques — mots. En effet, les caractères de l'alphabet sont naturellement exprimés dans leur forme la plus simple. Parallèlement, les étiquettes morpho-syntaxiques sont des UTs introduites par intervention extérieure et par conséquent leur taille peut être facilement optimisée lors de leur définition.

### 4.4.1 Compression Alphanumérique

Un compresseur permet de réduire la taille des données si celles-ci mettent en évidence des régularités qui peuvent être exploitées. Formellement, si un fichier informatique $F$ est comprimé par un compresseur $C$, détectant un ensemble de régularités $R$, pour donner le fichier comprimé $F'$ alors $F'$ ne pourra plus être comprimé par $C$ car l'ensemble des régularités $R$ ne seront plus présentes. Par conséquent, l'étape fondamentale de la concep-

tion d'un compresseur est la définition de l'ensemble des régularités mises en évidence par les données. Dans le cadre des données textuelles, de nombreuses études lexicométriques ont tenté de rendre compte de l'ensemble de ces régularités [74] [75] [54] [76]. Parmi celles-ci, nous nous intéressons plus particulièrement aux résultats obtenus sur le taux de couverture du vocabulaire. Si l'on définit le vocabulaire d'un énoncé comme l'ensemble des mots distincts qui le composent, le taux de couverture d'une forme du vocabulaire est déterminé dans l'équation 4.17 comme étant le quotient entre le nombre d'occurrences du mot considéré $N_{mot}$ et la taille $T$ du texte — le nombre d'occurrences de tous les mots.

$$\text{Taux de couverture} \quad = \quad \frac{N_{mot}}{T} \tag{4.22}$$

Les études sur le taux de couverture montrent que pour un texte donné un petit nombre de formes de son vocabulaire couvre une grande proportion de sa surface. Ainsi, l'idée de base de la compression alphanumérique est définie sur le principe du codage de Huffman [77] qui traduit chaque symbole d'un texte par une suite de bits d'autant plus courte que la fréquence d'apparition du symbole est grande. Du fait des contraintes imposées par l'implémentation des modèles $N$-gram positionnels, nous proposons que chaque UT soit codée par une suite de caractères alphanumériques d'autant plus courte que la fréquence d'apparition d'un mot est grande.

Afin d'implémenter le compresseur, les formes de l'énoncé sont d'abord triées en fonction de leur fréquence dans l'ensemble du corpus. Les mots de même fréquence sont départagés par leur taille en nombre de caractères, les formes de taille supérieure étant d'abord listées. L'ordre alphabétique est appliqué pour déterminer l'ordre final. Chaque forme ainsi listée se voit attribuer un code alphanumérique d'autant plus court que sa fréquence est élevée. Un code alphanumérique est une combinaison quelconque de $n$ caractères élémentaires — caractères visibles de la table ASCII. Ainsi, les premières formes sont codées avec un seul caractère. Une fois épuisés tous les codes d'un caractère, les formes sont comprimées avec une séquence de deux caractères. Ce processus se répète jusqu'à ce que soit attribué un code à la dernière forme de la liste ordonnée. A titre d'exemple, si l'on limite l'ensemble des caractères visibles de la table ASCII à l'ensemble $C = \{a, b, c\}$, les codes attribués à chacune des formes pour notre énoncé de travail sont énumérés dans le tableau 4.20. Afin d'effectuer les calculs des modèles $N$-gram positionnels, l'énoncé sera donc comprimé de

| $N^o$ | Forme | Fréq. | Code | $N^o$ | Forme | Fréq. | Code |
|---|---|---|---|---|---|---|---|
| 1 | linguistiques | 2 | a | 14 | nominale | 1 | aab |
| 2 | des | 2 | b | 15 | purement | 1 | aac |
| 3 | de | 2 | c | 16 | survol | 1 | aba |
| 4 | problématiques | 1 | aa | 17 | noms | 1 | abb |
| 5 | composition | 1 | ab | 18 | aux | 1 | abc |
| 6 | théoriciens | 1 | ac | 19 | met | 1 | aca |
| 7 | l'ensemble | 1 | ba | 20 | sur | 1 | acb |
| 8 | permettant | 1 | bb | 21 | la | 1 | acc |
| 9 | délimiter | 1 | bc | 22 | Le | 1 | baa |
| 10 | l'absence | 1 | ca | 23 | en | 1 | bab |
| 11 | composés | 1 | cb | 24 | “ | 1 | bac |
| 12 | critères | 1 | cc | 25 | ” | 1 | bba |
| 13 | évidence | 1 | aaa | 26 | . | 1 | bbb |

TAB. 4.20 – Liste ordonnée des formes du vocabulaire

la façon suivante[7].

> *Le survol des problématiques linguistiques sur la composition nominale met en évidence l'absence de critères purement linguistiques permettant aux théoriciens de délimiter l'ensemble des “ noms composés ” .*

Enoncé non codifié

[7] On notera que dans cet exemple plusieurs formes ne sont pas comprimées mais au contraire expansées. C'est le cas pour toutes les formes à partir du rang 21 jusqu'au rang 26. Cette caractéristique ne s'applique pas dans des conditions normales de codification — pour des énoncés de tailles suffisamment grandes — pour deux raisons principales. Premièrement, l'ensemble des caractères élémentaires est un ensemble de 147 caractères et pas seulement 3 comme dans cet exemple. Ainsi, il existe $147^1$ codes d'un caractère, $147^2$ codes de deux caractères, $147^3$ codes de trois caractères etc.. Deuxièmement, selon la loi de Zipf [74], les formes les plus fréquentes sont les formes de plus petites tailles, stipulant ainsi que le propre langage est déjà comprimé et que la situation illustrée dans cet exemple ne se vérifie jamais dans des conditions d'énoncés de tailles suffisamment grandes.

*baa aba b aa a acb acc ab aab aca bab aaa ca c cc aac a bb abc ac c bc ba b bac abb cb bba bbb*

Enoncé codifié

Afin de motiver l'intérêt de la conception de ce compresseur, une évaluation s'impose.

### 4.4.2 Evaluation

Dans le but d'évaluer notre compresseur, nous le comparons sur la base du taux de compression à la technique de numérisation communément utilisée en statistique textuelle [1].

#### Numérisation

La numérisation est une technique largement utilisée en statistique textuelle qui consiste à attribuer à chaque forme du texte un numéro d'ordre qui sera associé à chacune des occurrences du mot. Les mots sont d'abord triés en fonction de leur fréquence dans l'ensemble du corpus et l'ordre alphabétique départage les formes de même fréquence. A partir de l'énoncé précédent, le code de chaque forme correspond à son rang dans la liste ordonnée présentée dans le tableau 4.21.

Afin d'effectuer les traitements nécessaires, l'énoncé serait donc comprimé de la façon suivante.

*14 25 2 22 3 24 17 9 19 18 12 13 15 1 10 23 3 21 7 26 1 11 16 2 4 20 8 5 6*

#### Taux de compression

Afin de pouvoir évaluer les performances d'un compresseur sur un ensemble de données, nous introduisons le concept de taux de compression. Le taux de compression est défini par le quotient entre le nombre de symboles économisés par le codage — gain — et le nombre de symboles des données initiales. Dans le cadre des formes graphiques, nous définissons le taux de compression comme étant le quotient entre le nombre d'octets[8] économisés par le codage et le nombre d'octets de l'énoncé initial. La formule du taux de compression

[8]Il faut trois octets pour coder un entier entre 0 et 16 777 215 et un octet pour coder un caractère de la table ASCII

| Rang | Forme | Fréq. | Rang | Forme | Fréq. |
|---|---|---|---|---|---|
| 1 | de | 2 | 14 | Le | 1 |
| 2 | des | 2 | 15 | l'absence | 1 |
| 3 | linguistiques | 2 | 16 | l'ensemble | 1 |
| 4 | “ | 1 | 17 | la | 1 |
| 5 | ” | 1 | 18 | met | 1 |
| 6 | . | 1 | 19 | nominale | 1 |
| 7 | aux | 1 | 20 | noms | 1 |
| 8 | composés | 1 | 21 | permettant | 1 |
| 9 | composition | 1 | 22 | problématiques | 1 |
| 10 | critères | 1 | 23 | purement | 1 |
| 11 | délimiter | 1 | 24 | sur | 1 |
| 12 | en | 1 | 25 | survol | 1 |
| 13 | évidence | 1 | 26 | théoriciens | 1 |

TAB. 4.21 – Numérisation des formes du vocabulaire

est définie par l'équation 4.23 où $N_{cod}$ et $N_{ini}$ sont respectivement le nombre d'octets de l'énoncé codifié et le nombre d'octets de l'énoncé initial.

$$\text{Taux de compression} = \frac{N_{ini} - N_{cod}}{N_{ini}} \tag{4.23}$$

Le taux de compression mesure donc la quantité d'informations économisée. Ainsi, plus le taux de compression est grand, meilleure est la compression.

### Résultats comparatifs

Nous avons appliqué notre compresseur sur trois corpora d'environ 30 000 mots chacun, extraits d'une collection de débats politiques du Parlement Européen. Les trois corpora sont écrits dans trois langues différentes. Dans tous les cas, notre compresseur s'est montré plus performant que la technique de numérisation[9]. Dans le tableau 4.22, nous présentons les taux de compression — en pourcentage — de chacune des deux méthodes ainsi que

[9]Chaque code entier de la numérisation a été codé sur trois octets.

celui de l'utilitaire *gzip* — version 1.2.4 — qui servira de valeur de référence[10] [11].

| Langue | Codification Alphanumérique | Numérisation | *gzip* |
|---|---|---|---|
| Français | 71.21 | 38.64 | 66.05 |
| Portugais | 70.41 | 37.28 | 65.87 |
| Anglais | 70.27 | 36.16 | 65.52 |

TAB. 4.22 – Taux de compression en pourcentage par langue

Afin de réaliser une évaluation consistante, nous devons mesurer l'efficacité de notre compresseur par rapport à la taille de l'énoncé considéré. En effet, dans l'évaluation précédente seules les langues différaient. Dans le tableau 4.23, nous proposons donc les mêmes comparaisons pour un ensemble d'énoncés de différentes tailles extraits d'une collection de textes juridiques écrits en portugais.

| Taille (mots) | Codification Alphanumérique | Numérisation | *gzip* |
|---|---|---|---|
| 100 000 | 79.49 | 59.41 | 73.44 |
| 500 000 | 78.68 | 59.11 | 73.99 |
| 1 000 000 | 78.56 | 59.19 | 74.05 |

TAB. 4.23 – Taux de compression en pourcentage par taille

Les résultats montrent un gain de compression d'environ 20% en faveur de la codification alphanumérique par rapport à la numérisation — ceci indépendamment de la langue utilisée et la taille de l'énoncé. Ces chiffres doivent néanmoins être relativisés. En effet, la numérisation privilégie l'aspect manipulation des données au détriment du stockage. Cependant, dans le cadre de la construction des modèles $N$-gram positionnels, le stockage des données doit être considéré comme l'élément fondamental d'optimisation.

[10] La compression réalisée par *gzip* donne lieu à un énoncé codifié qui n'est pas exploitable pour le calcul des modèles $N$-gram positionnels du fait de la perte des séparations entre formes graphiques. Cependant, il nous paraît utile de mentionner ses résultats comme véritables valeurs de référence.

[11] On notera que dans l'absolu il faudrait compter la taille du dictionnaire pour le calcul du taux de compression. En effet, la codification alphanumérique et la numérisation nécessitent d'un dictionnaire de formes codées alors que la méthode *gzip* n'en utilise pas. Dans notre cas, nous négligerons cette variable puisque la méthode *gzip* sert uniquement de valeur de référence.

## 4.5 Conclusion

Dans ce chapitre, nous avons défini les unités minimales de traitement — unités textuelles — ainsi que les unités de mesure et de comptage — modèles $N$-gram positionnels — sur lesquelles nous allons nous appuyer pour évaluer le degré de cohésion qui lie entre elles les différentes UTs d'un $N$-gram positionnel. En effet, les associations textuelles peuvent être définies en terme de structure par un $N$-gram positionnel quelconque. Cependant, tous les $N$-grams positionnels construits ne sont pas des associations textuelles. Par conséquent, dans le chapitre suivant, nous introduisons une nouvelle mesure d'association qui permet d'associer à chaque $N$-gram positionnel — $\forall N, N \geq 2$ — une valeur probabiliste de cohésion qui mesure les forces d'attraction qui existent entre les UTs d'un $N$-gram positionnel. Cette mesure s'appelle l'Expectative Mutuelle ■

# Chapitre 5

# Extraction d'Associations Textuelles

Classiquement, l'extraction d'associations textuelles se divise en deux phases. Dans un premier temps, une mesure d'association évalue le degré de cohésion intrinsèque à une séquence d'UTs. Dans un deuxième temps, un ensemble d'heuristiques permet d'identifier les associations textuelles candidates parmi tous les $N$-grams pondérés selon leur degré d'attraction. Dans ce cadre, plusieurs problèmes restent encore irrésolus. C'est le cas de la définition de mesures d'association normalisées et de la conception d'heuristiques de sélection qui ne dépendent pas de valeurs seuil définies de façon *ad hoc*. Devant ce constat, nous proposons une nouvelle méthodologie pour l'extraction d'associations textuelles qui combine une nouvelle mesure d'association — l'Expectative Mutuelle — à un algorithme de sélection basé sur la notion de maximaux locaux — le GenLocalMaxs[1].

## 5.1 Généralités des Mesures d'Association

La méthode statistique s'appuie sur des mesures et des comptages réalisés à partir d'objets que l'on veut comparer. Ainsi, après avoir défini la structure des objets considérés dans le chapitre précédent — unités textuelles et modèles $N$-gram positionnels, nous introduisons le concept de mesure d'association. Les mesures d'association sont des modèles statistiques, probabilistes ou numériques qui permettent de mesurer les forces qui lient entre elles plusieurs UTs. En effet, une association textuelle est une combinaison récurrente

[1]Nous verrons que le GenLocalMaxs constitue une généralisation de l'algorithme LocalMaxs proposé par J. Silva *et al.* [30]

d'UTs qui se trouvent ensemble plus souvent que par le simple fait du hasard et dénote ainsi un degré de cohésion implicite. Dans le cadre de la statistique textuelle, plusieurs mesures d'association ont été proposées et leur analyse distingue deux approches principales. D'une part, des mesures d'association mesurant les forces d'attraction entre deux UTs ont été élaborées. Nous les nommons mesures d'association binaires. D'autre part, un certain nombre de recherches se sont fixées pour objet le repérage direct de cooccurrences de deux ou plusieurs UTs. Nous les mentionnons mesures d'association $N$-aires. Sans prétendre à l'exhaustivité, nous illustrons quelques-unes des mesures élaborées selon les deux méthodes proposées dans les deux paragraphes suivants.

### 5.1.1 Mesures d'Association Binaires

On peut regrouper en trois catégories les mesures d'association binaires qui calculent les attirances entre couples d'UTs au sein d'un contexte donné. Ainsi, les chercheurs ont proposé des mesures d'association basées soit sur des tests statistiques [28] [29] soit sur des estimateurs de la théorie de l'information [27] [32] soit sur des heuristiques probabilistes [31] [56] [57]. Pour une UT pivot donnée, ces méthodes sélectionnent un ensemble d'UTs qui ont tendance à se trouver dans le voisinage de cette forme. Dans ce cadre, il faut commencer par définir une unité de contexte, ou voisinage, à l'intérieur de laquelle on considèrera que deux formes sont cooccurrentes. Cette unité peut ressembler à la phrase ou encore être constituée par un contexte de longueur fixe — environnement immédiat. Ensuite, une mesure d'association est élaborée ayant pour objectif de repérer les couples d'UTs les plus pertinents à l'intérieur de ce contexte.

#### Heuristiques Probabilistes

Dans le cadre des associations lexicales binaires, D. Labbé [56] propose une méthode particulièrement simple destinée à mettre en évidence ce qu'il appelle l'univers lexical d'une UT donnée. Pour chaque UT du corpus notée $u_i$, l'ensemble des phrases de l'énoncé peut être divisé en deux sous-ensembles : $P_1$, le sous-ensemble de celles qui contiennent l'UT $u_i$ et $P_0$, le sous-ensemble dont $u_i$ est absente. Pour chacune des autres UTs du corpus, on applique ensuite le test de l'écart-réduit aux sous-fréquences dans les deux ensembles $P_0$ et $P_1$ en tenant compte de leurs longueurs respectives. Si les fréquences des UTs considérées ne sont pas trop faibles, cette méthode permet de sélectionner, pour

chaque UT pivot donnée, un ensemble de formes qui se trouvent situées de manière privilégiée dans les mêmes phrases.

Suivant cette même approche, P. Lafon et M. Tournier [57] diffèrent cependant de la méthode précédente sur deux points principaux. Premièrement, ils distinguent les positions relatives par rapport à l'UT pivot séparant de ce fait des cooccurrences *avant* et des cooccurrences *après*. Deuxièmement, ils font intervenir à la fois la cofréquence des deux formes considérées et leur distance moyenne mesurée en nombre d'occurrences. A l'aide de valeurs seuil en probabilité, cette méthode permet d'extraire des paires d'UTs présentant des affinités dans la collection des textes étudiés.

Toujours selon la même ligne, F. Smadja [31] propose une méthode similaire à celle de P. Lafon et M. Tournier pour le calcul des paires d'UTs. Dans un premier temps, il mesure l'attirance entre deux UTs par le biais du score centré réduit[2] appliqué à la fréquence des UTs qui se trouvent dans le voisinage droit et gauche de l'UT pivot considérée. Dans un deuxième temps, une analyse des positions permet de ne sélectionner que les UTs dont la position est raisonnablement fixe par rapport à l'UT pivot. Le système de F. Smadja se distingue cependant des travaux de P. Lafon et M. tournier par l'élaboration d'une simple heuristique qui permet d'extraire des unités lexicales complexes de taille supérieure à deux UTs. Ainsi, à partir des associations binaires préalablement extraites, une nouvelle étude du voisinage du couple d'UTs est effectuée. Les UTs dont la probabilité d'apparition dépasse une certaine valeur seuil pour une position donnée sont concaténées aux deux UTs de l'association binaire pour former une unité complexe de taille supérieure à deux. On appellera méthode d'amorçage cette forme d'extraction.

**Tests Statistiques**

Parallèlement aux approches présentées précédemment, W. Gale [29] et T. Dunning [28] proposent respectivement deux mesures statistiques supportées par la théorie du test d'hypothèse pour des séries statistiques bivariées. Leur objectif est de mesurer les liens qui unissent deux UTs notées $u_1$ et $u_2$ à partir de l'analyse d'un tableau de contingence construit de la forme suivante en 5.1.

[2]Le score centré réduit est la traduction française du z-score utilisé par les anglo-saxons.

| | $u_2$ | $\neg u_2$ |
|---|---|---|
| $u_1$ | $k(u_1, u_2)$ | $k(u_1, \neg u_2)$ |
| $\neg u_1$ | $k(\neg u_1, u_2)$ | $k(\neg u_1, \neg u_2)$ |

TAB. 5.1 – Tableau de Contingence

Ainsi, parmi l'ensemble des digrams contigus construits à partir d'un énoncé, un tableau de contingence met en évidence les comptages de ceux qui contiennent à la fois $u_1$ et $u_2$ c'est-à-dire $k(u_1, u_2)$, ainsi que ceux qui ne contiennent ni l'une ni l'autre des deux unités c'est-à-dire $k(\neg u_1, \neg u_2)$. Parallèlement, $k(u_1, \neg u_2)$ et $k(\neg u_1, u_2)$ représentent les comptages des digrams qui contiennent $u_1$ (resp. $u_2$) et qui ne contiennent pas $u_2$ (resp. $u_1$).

Dans ce cadre, W. Gale [29] propose de tester la validité de l'hypothèse d'indépendance qui consiste à considérer deux UTs indépendantes si leur probabilité de cooccurrence est égale au produit de leurs probabilités marginales. L'hypothèse d'indépendance — ou hypothèse nulle $H_0$ — est exprimée par l'équation de probabilité suivante entre deux UTs notées $u_1$ et $u_2$.

$$H_0 \quad : \quad p(u_1, u_2) = p(u_1).p(u_2)$$

Afin de mesurer l'écart d'indépendance entre deux UTs, W. Gale utilise le coefficient d'association de Pearson $\Phi^2$ basé sur le test statistique $\chi^2$ [78] et défini dans l'équation suivante où $N$ est le nombre d'UTs de l'énoncé.

$$\Phi^2\,(u_1, u_2) = \frac{(N \times k(u_1,u_2) - k(u_1) \times k(u_2))^2}{k(u_1) \times (N - k(u_1)) \times k(u_2) \times (N - k(u_2))} \tag{5.1}$$

Ainsi, si le $\Phi^2$ est minimum, l'hypothèse d'indépendance est vraie et par conséquent les deux UTs peuvent être considérées indépendantes. Dans le cas contraire, W. Gale assume que les deux UTs forment une unité lexicale complexe.

De son côté, T. Dunning [28] attaque la supposition de normalité qui est sous-jacente à un grand nombre de tests statistiques. En effet, les tests tels que le score centré réduit ou le $\chi^2$ qui se basent sur des observations de grande taille, trébuchent face à la réalité du matériel textuel qui se distingue par la prédominance d'événements peu fréquents. Ainsi, l'idée qui consiste à estimer que les résultats des observations effectuées sur les textes suivent une distribution normale — ou approximativement normale — est caduque pour des énoncés de taille réduite. Dans ce cadre, T. Dunning utilise le test de vraisemblance maximum qui permet d'analyser avec succès les tableaux de contingence dont les comptages ne sont pas forcément élevés. Ainsi, il propose de tester l'hypothèse nulle $H_0$ — hypothèse d'indépendance — selon laquelle la probabilité d'apparition d'une UT dans un énoncé est indépendante de la cooccurrence d'une autre UT quelconque dans son voisinage. Dans ce cadre, il est nécessaire de déterminer l'hypothèse alternative à $H_0$ notée $H_1$. Ces deux hypothèses peuvent être formulées par les deux expressions suivantes pour deux UTs notées $u_1$ et $u_2$.

$$\begin{aligned} H_0 &: p(u_1|u_2) = p(u_1|\neg u_2) = p(u_1) = \theta \\ H_1 &: p(u_1|u_2) = \theta_1 \neq p(u_1|\neg u_2) = \theta_2 \end{aligned}$$

Si l'on considère une série d'expériences de Bernoulli qui consistent à observer l'occurrence ou la non apparition d'une UT donnée dans un digram, le test de validité de l'hypothèse $H_0$ par rapport à l'hypothèse alternative $H_1$ est mesuré grâce à la valeur $-2log\lambda$ qui a une distribution $\chi^2$. Ainsi, plus la valeur $-2log\lambda$ est élevée, plus l'hypothèse d'indépendance $H_0$ est fausse et plus il est probable que les deux UTs forment une unité lexicale complexe. Nous donnons la formule de la valeur $-2log\lambda$ dans l'équation suivante où $N$ correspond au nombre d'UTs de l'énoncé.

$$-2\log\lambda = Loglike\,(u_1, u_2) =$$

$$\begin{aligned} 2 \times (\ & \log\theta_1^{s1}\,(1-\theta_1)^{n1-s1} + \log\theta_2^{s2}\,(1-\theta_2)^{n2-s2} \\ & - \log\theta^{s1}\,(1-\theta)^{n1-s1} - \log\theta^{s2}\,(1-\theta)^{n2-s2}) \end{aligned}$$

où

$$
\begin{array}{ll}
s1 = k\left(u_1, u_2\right) & s2 = k\left(u_2\right) - k\left(u_1, u_2\right) \\
n1 = k\left(u_1\right) & n2 = N - k\left(u_1\right) \\
\theta_1 = \frac{s1}{n1} & \theta_2 = \frac{s2}{n2} \\
\theta = \frac{k(u_2)}{N} &
\end{array}
$$

Nous aborderons plus en détail l'ensemble de ces mesures pour lesquelles nous présenterons une normalisation possible dans le chapitre suivant.

**Théorie de l'information**

Finalement, à contre courant des deux approches déjà énoncées, un certain nombre de travaux se sont intéressés à l'application de mesures de la théorie de l'information pour découvrir certaines paires d'UTs pertinentes [27] [32].

Suivant cette intuition, K. Church et P. Hanks [27] sont les premiers à proposer l'application d'une mesure basée sur l'Information Mutuelle [79] pour identifier un ensemble de paires d'UTs cooccurrentes. Ils l'appellent le coefficient d'association. Dans ce cadre, cette nouvelle mesure permet d'évaluer la quantité d'informations contenue dans une UT sur l'occurrence d'une autre UT dans son voisinage. Ainsi, pour une UT donnée, plus l'information qu'elle contient sur l'occurrence d'une autre UT est grande, plus il est probable que le couple d'UTs forme une unité lexicale complexe. Le coefficient d'association entre deux UTs notées $u_1$ et $u_2$ est défini dans l'équation suivante où $p(u_1, u_2)$ correspond à la probabilité d'occurrence conjointe des deux UTs $u_1$ et $u_2$ et $p(u_1)$ (resp. $p(u_2)$) à la probabilité marginale d'occurrence de la forme $u_1$ (resp. $u_2$).

$$I(u_1, u_2) \equiv \log_2 \frac{p(u_1, u_2)}{p(u_1).p(u_2)} \tag{5.2}$$

A la lumière de cette formule, il est clair que plus les deux UTs considérées sont indépendantes l'une de l'autre, plus la valeur du coefficient d'association est proche de 0. Inversement, plus les deux UTs montrent un fort degré d'attraction, plus le coefficient d'association est élevé.

Dans le même ordre d'idée, S. Shimohata [32] propose l'utilisation de la mesure d'entropie [59] pour l'identification de cooccurrences de caractères. Là encore, l'objectif principal est de mesurer la quantité d'information contenue dans une variable aléatoire donnée. Ainsi, S. Shimohata propose de mesurer l'entropie de l'ensemble des UTs qui se trouvent immédiatement *devant* et immédiatement *après* une séquence de caractères notée *str* et de ne retenir comme cooccurrents de la suite *str* que les UTs démontrant une faible valeur d'entropie c'est-à-dire une forte cohésion. Dans ces conditions, la mesure d'entropie peut être définie par l'expression suivante où $u_i$ est l'une des $n$ UTs adjacentes à la séquence *str* et $p(u_i)$ correspond au quotient $\frac{k(u_i)}{k(str)}$.

$$H(str) = -\sum_{i=1}^{n} p(u_i).\log_2 p(u_i) \tag{5.3}$$

Dans ces conditions, plus les UTs qui précèdent ou suivent la suite de caractères considérée sont dispersées, plus la stabilité de la chaîne *str* est mise en cause et plus l'entropie est élevée. En guise de remarque, on notera que la mesure proposée par S. Shimohata est une mesure d'association binaire bien qu'elle permette l'identification d'unités complexes $N$-aires. En effet, l'entropie est calculée entre deux groupes d'UTs et les unités complexes de plus de deux UTs sont identifiées par amorçage.

Plusieurs autres mesures d'association binaires ont été proposées dans la littérature. Cependant, il n'est pas concevable dans le cadre de notre travail de toutes les détailler individuellement. Parmi celles-ci, nous rappellerons le coefficient Dice introduit par F. Smadja [45] et l'association sélective proposée par P. Resnik [80]. Le lecteur pourra également trouver un récapitulatif exhaustif d'un ensemble de mesures d'association statistiques dans [81] ainsi que dans [82].

La plupart de ces modèles n'ont cependant été définis que pour les modèles digrams d'unités textuelles — modèles $N$-gram pour $N = 2$ — et ne permettent donc pas de mesurer les forces d'attraction qui existent entre tous les constituants d'un $N$-gram générique — $\forall N, N \geq 2$. De ce fait, l'acquisition d'associations de plus de deux UTs requiert un travail complémentaire où les paires d'associations acquises initialement jouent le rôle d'amorce [31] [43]. Parallèlement, la plupart de ces modèles sont sensibles à l'occurrence d'UTs fréquentes et par conséquent sous-évaluent généralement les associations qui les

contiennent. Ainsi, dans le cadre des associations lexicales, B. Daille [21] et C. Enguehard [23] ne peuvent considérer que les occurrences des mots pleins pour évaluer les forces de cohésion et évitent l'intégration des fragments fonctionnels souvent fréquents dans l'application des mesures d'association. Dans le but de proposer une alternative aux méthodes d'amorçage, un certain nombre de mesures qui évaluent les attirances entre toutes les UTs d'un $N$-gram générique ont été définies. Nous en introduisons quelques-unes dans le paragraphe suivant.

### 5.1.2 Mesures d'Association $N$-aires

Un autre groupe de mesures d'association a pour objectif le repérage direct des cooccurrences de deux ou plusieurs UTs du texte. Dans ce cadre, plusieurs méthodes ont été proposées suivant des concepts numériques [61] [33] [34] [60] [78] ou probabilistes [30].

#### Heuristiques Numériques

La méthode la plus simple a été introduite par A. Salem [34] dont l'idée principale est formulée par la citation suivante : *"[...] le repérage des segments les plus répétés dans un corpus de textes, en plus de l'éclairage quantitatif qu'il apporte sur ces problèmes, permet de mettre en évidence des unités qui constituent souvent des syntagmes autonomes"*. Ainsi, A. Salem propose de repérer toutes les séquences d'UTs qui se répètent dans les énoncés. Dans ce cadre, il préconise une analyse détaillée de chaque segment répété par le biais des tableaux des segments répétés — TSR — et de leurs inventaires alphabétiques, hiérarchiques et distributionnels. Plus tard, M. Bécue [78] présentera un algorithme permettant de repérer les quasi-segments c'est-à-dire les segments répétés qui supportent quelques modifications d'ordre.

Dans le même ordre d'idée, K. Frantzi [33] adopte la *C-value* pour calculer la pertinence d'une séquence d'UTs notée $a$. Cette mesure se base sur trois concepts fondamentaux : la fréquence d'occurrence de chaque séquence d'UTs — $n(a)$, la fréquence d'occurrence à l'intérieur de séquences plus longues — $t(a)$ et le nombre de ces séquences qui contiennent d'autres séquences — $c(a)$. La *C-value* est définie de la façon suivante selon trois conditions 1, 2 et 3.

$$\text{C-value(a)} = \begin{cases} \textit{Condition} & 1, \quad 0 \\ \textit{Condition} & 2, \quad n(a) \\ \textit{Condition} & 3, \quad n(a) - \frac{t(a)}{c(a)} \end{cases} \tag{5.4}$$

Ainsi, si la séquence $a$ apparaît exactement le même nombre de fois qu'une autre séquence qui la contient, la condition 1 doit être appliquée — $a$ n'est pas un terme. Si la séquence n'est pas un sous-groupe d'une autre séquence, la condition 2 est mise en évidence. Finalement, si la séquence apparaît dans d'autres séquences plus longues et la condition 1 n'est pas assurée, la condition 3 définit la *C-value*. L'idée de K. Frantzi est comparable à celle du repérage des segments répétés proposés par A. Salem à la différence qu'elle utilise les comptages des sous-segments — quasi-segments — pour identifier les segments les plus pertinents.

Parallèlement à ces deux approches, les travaux de G. Chartron [61] et R. Schneider [60] se distinguent par l'utilisation des fréquences marginales des UTs. Dans le domaine de l'indexation automatique, G. Chartron [61] propose de repérer les cooccurrences de plusieurs termes à partir du coefficient d'implication réciproque. L'objectif est de comparer la fréquence d'apparition d'une séquence de $n$ UTs par rapport au produit des fréquences de chacune des UTs constituant la séquence. Ainsi, pour une suite composée de $n$ UTs, le coefficient d'implication réciproque noté $E$ est calculé de la façon suivante où $k(u_1u_2...u_n)$ correspond au nombre de cooccurrences des $n$ formes dans le corpus et $k(u_1)$, $k(u_2)$, ..., $k(u_n)$ au nombre d'occurrences de chacune des $n$ UTs notées $u_i$.

$$E = \frac{k(u_1u_2...u_n)}{k(u_1)k(u_2)...k(u_n)} \tag{5.5}$$

L'idée sous-jacente proposée par G. Chartron est de déterminer l'ensemble des séquences qui contiennent le plus petit nombre de fragments fonctionnels — les UTs fréquentes. Ainsi, moins il y a d'UTs fréquentes dans la séquence, plus le coefficient d'implication réciproque $E$ est élevé, et plus il est possible que la suite soit pertinente.

Suivant la même intuition, R. Schneider [60] propose une mesure basée sur les listes de fréquences d'UTs. Dans un premier temps, chaque UT de l'énoncé reçoit un poids qui

est fonction de son "importance relative"[3]. Ensuite, les UTs sont rangées dans l'ordre décroissant de leur poids et sont associées à un rang noté $r$ dans la liste — les UTs les plus "informatives" ont un rang moins élevé. A partir de cette liste, R. Schneider définit la pertinence de chaque séquence d'UTs par le rapport entre la fréquence des rangs de chaque UT constituant la séquence et la fréquence de la suite d'UTs. Ce rapport est noté $\tilde{C}$ et défini de la forme suivante où $N$ correspond au nombre d'UTs de l'énoncé.

$$\tilde{C}_{(u_1 u_2 ... u_n)} = \frac{\sum_{i=1}^{n} \tilde{r}_{u_i}}{N.k(u_1 u_2 ... u_n)} \tag{5.6}$$

Ainsi, plus le coefficient est proche de zéro, plus il est probable que la suite d'UTs soit une unité complexe. L'idée de R. Schneider est étroitement comparable à celle de G. Chartron à la différence de l'introduction d'une variable supplémentaire, le poids de chaque UT. En effet, son intention est de diminuer l'influence des mots fonctionnels en faveur des unités spécifiques du domaine.

### Heuristiques Probabilistes

Parallèlement aux approches purement numériques, J. Silva *et al.* [30] proposent de normaliser les mesures probabilistes définies pour deux UTs. En particulier, J. Silva propose la normalisation d'une nouvelle mesure d'association basée sur le concept de probabilité conditionnelle réciproque et définie préalablement pour deux UTs : la probabilité conditionnelle symétrique. L'idée sous-jacente à cette mesure est de calculer une valeur moyenne d'attraction entre les éléments d'un $N$-gram contigu. Ainsi, la probabilité conditionnelle symétrique est définie dans l'équation suivante pour deux UTs notées $u_1$ et $u_2$ où $k(u_1, u_2)$ correspond à la fréquence d'occurrence du digram contigu $[u_1, u_2]$ et $k(u_1)$ (resp. $k(u_2)$) à la fréquence d'occurrence de la forme $u_1$ (resp. $u_2$).

$$SCP(u_1, u_2) = \frac{k(u_1, u_2)^2}{k(u_1).k(u_2)} \tag{5.7}$$

Afin de généraliser cette mesure pour le cas des $N$-grams contigus — $\forall N, N \geq 2$, J. Silva propose d'évaluer un ensemble d'attractions possibles entre différents sous-groupes d'un $N$-

[3]Nous éviterons l'explication de cette donnée qui nous semble peu utile pour la compréhension de la méthodologie utilisée.

gram. La mesure d'association normalisée peut alors être considérée comme une moyenne des attractions implicites entre les éléments d'un $N$-gram contigu. Ainsi, la normalisation proposée permettant d'évaluer le degré de cohésion de tout $N$-gram contigu donne lieu à l'équation suivante.

$$SCP(u_1, ..., u_n) = \frac{k(u_1...u_n)^2}{\frac{1}{n-1}\sum_{i=1}^{i=n-1} k(u_1...u_i).k(u_{i+1}...u_n)} \tag{5.8}$$

Contrairement aux mesures précédentes, la probabilité conditionnelle symétrique se distingue par son contexte probabiliste qui permet d'étudier de façon rigoureuse le calcul des forces d'attraction. Dans ce cadre, on remarquera que d'autres auteurs ont proposé de généraliser certaines mesures d'association bien connues. En particulier, on citera les travaux de P.K. Kim [62] qui propose une nouvelle mesure basée sur la normalisation de l'Information Mutuelle dans le cadre des 3-grams d'UTs. On regrettera cependant qu'aucune méthode de généralisation exhaustive — i.e. pour tout $N$ — ne soit avancée.

En guise d'évaluation, la principale remarque qui se doit d'être énoncée, est le fait que les mesures d'association $N$-aires proposées jusqu'à présent ne s'appliquent qu'aux séquences continues d'UTs. En effet, il n'est jamais fait allusion aux associations textuelles non-contiguës qui comme nous l'avons vu précédemment sont nombreuses. Plus grave encore, le problème que posent les associations distantes n'est jamais abordé et laissé aux mains des mesures d'association binaires. La deuxième remarque s'applique plus particulièrement aux mesures numériques qui dans la pratique impliquent l'utilisation de listes de mots vides qui permettent d'éliminer un ensemble d'associations textuelles non pertinentes. Or, cette approche n'est pas souhaitable. En effet, elle engendre la contrainte de sélection des unités de décompte qui n'est pas un problème simple comme de nombreux auteurs l'ont défini. Finalement, l'idée de normalisation proposée par J. Silva *et al.* peut être définie comme le premier pas vers une uniformisation des mesures d'association. Dans ce cadre, tous nos efforts seront basés sur la définition, la plus exacte possible, des forces qui unissent tous les éléments d'une séquence d'UTs. Dans ce contexte, nous verrons que la méthode de normalisation proposée par J. Silva n'est pas complète. En effet, elle ne prend pas en compte toutes les associations possibles entre UTs et par conséquent ne propose qu'une solution intermédiaire. Ainsi, afin d'éviter l'intervention sur le choix des unités de décompte, le recours aux méthodes d'amorçage et l'absence de définition de mesures

prenant en compte les phénomènes non continus du langage, nous proposons une nouvelle mesure d'association appelée Expectative Mutuelle. A l'instar des mesures d'association proposées antérieurement, l'Expectative Mutuelle permet d'évaluer l'ensemble des attractions présentes entre les UTs de séquences continues ou non. Ainsi, l'Expectative Mutuelle est définie de façon générique pour tout $N$-gram positionnel c'est-à-dire pour tout $N$ tel que $N \geq 2$. De plus, comme on le verra dans la partie suivante de ce rapport, l'Expectative Mutuelle démontre la bonne propriété de ne pas sous-évaluer les associations qui contiennent des UTs fréquentes.

## 5.2 Une Nouvelle Mesure d'Association

Dans cette partie, nous définissons une nouvelle mesure d'association appelée Expectative Mutuelle. L'Expectative Mutuelle est une mesure probabiliste de l'approche non paramétrique qui permet de calculer les forces d'attraction qui lient entre elles les UTs d'un $N$-gram positionnel sans dépendre démesurément des fréquences marginales élevées. Dans un premier temps, nous définirons les bases de la théorie des probabilités par le biais de l'espace probabilisé sur lequel nous appliquerons l'ensemble de nos dénombrements. Ensuite, nous définirons formellement l'Expectative Mutuelle à partir des notions d'Expectative Normalisée et de fréquence relative qui peuvent être comparées respectivement aux notions de *support* et de *confiance* formulées par R. Agrawal [44] dans le cadre de la définition de règles d'association à partir de bases de données non textuelles.

### 5.2.1 Espace Probabilisé

Dans les applications de la statistique textuelle, il est nécessaire de porter une attention particulière à la définition de l'espace probabilisé sur lequel les modèles mathématiques sont appliqués. Dans le cas contraire, les décomptes utilisés ne sont que des facteurs de pondération *ad hoc* sans fondement théorique. Un espace probabilisé bien fondé noté ($\Omega$, $\mathcal{A}$, $P(.)$) est défini par un ensemble fondamental $\Omega$, un espace des événements $\mathcal{A}$ et une fonction de probabilité $P(.)$.

#### Ensemble fondamental

Les phénomènes auxquels s'applique la théorie des probabilités sont très variés. On les appelle habituellement des expériences aléatoires. Dans ce cadre, une expérience est

toute action ou processus qui engendre des observations ou des résultats[4]. On dira qu'une expérience est aléatoire s'il n'est pas possible de prédire avec certitude son résultat et si l'on peut décrire l'ensemble de tous les résultats possibles, appelé ensemble fondamental et noté $\Omega$. Un exemple classique d'expérience est l'observation du lancer de dé où les résultats sont des valeurs discrètes variant de 1 à 6 — $\Omega = \{1, 2, 3, 4, 5, 6\}$. Un ensemble fondamental peut être discret ou continu suivant les valeurs prises par les résultats de l'expérience considérée. L'expérience du lancer de dé met en évidence un ensemble fondamental discret puisque les observations sont dénombrables et finies. Au contraire, si l'on considère l'expérience qui consiste à calculer la durée d'attente d'une rame de métro, l'ensemble $\Omega$ est continu puisque les résultats ne sont pas dénombrables et par conséquent infinis. Du fait des décomptes réalisés sur les occurrences d'UTs, les ensembles fondamentaux en statistique textuelle sont généralement discrets.

Dans le cadre de notre recherche, l'objectif des mesures d'association est d'évaluer les forces d'attraction qui existent entre les UTs d'un $N$-gram positionnel. Pour ce faire, elles utilisent les observations réalisées sur les $N$-grams positionnels. Dans ce contexte, nous définissons un $N$-gram positionnel comme le résultat observé d'une expérience qui consiste à tirer $N$ UTs d'un sac d'UTs — l'énoncé — suivant un ensemble de positions déterminées. Ainsi, l'ensemble fondamental $\Omega$ discret peut être défini comme l'ensemble des $N$-grams positionnels construits à partir d'un énoncé initial tel que la taille du $N$-gram positionnel $N$, l'ensemble des $N$ positions $\wp$ et l'environnement immédiat de taille $F$ sont donnés. $\Omega$ est défini génériquement à partir de l'égalité suivante.

$$\Omega \quad = \quad \{\text{N-gram positionnel} \quad | \quad \text{N, } \wp\text{, F} \quad \text{sont donnés}\} \tag{5.9}$$

On notera que chaque $N$-gram positionnel de $\Omega$ est une instance de l'expérience. Or, si l'on considère $T$ la taille de l'énoncé et $F$ la taille de l'environnement immédiat, $\Omega$ est constitué de $T - 2.F$ instances — sauf si $N = 2$ où dans ce cas il existe $T - F$ instances. En effet, pour un $N$, un $F$ et un $\wp$ donnés, il n'existe qu'un seul $N$-gram possible pour chaque UT pivot de l'énoncé. Nous reviendrons sur ce résultat plus loin dans nos propos[5].

[4]Pour éviter une confusion possible, certains auteurs préfèrent le nom d'épreuve.

[5]Nous considèrerons $T - 2.F$ instances comme le cas général dans la suite de nos explications sans oublier bien sür que ceci n'est vrai que pour $N > 2$.

**Espace des Evénements**

Nous avons associé à une expérience aléatoire un ensemble fondamental fini $\Omega$ dont les éléments $\omega_1, \omega_2, \ldots, \omega_p$ désignent tous les résultats possibles de cette expérience. Or, chacun de ces éléments peut être défini par une propriété du résultat qui n'est possédé que par lui-même. On dit alors que chaque élément $\omega_i$ définit un événement élémentaire. Par exemple, dans le cas du lancer de dé, l'ensemble fondamental $\Omega = \{1, 2, 3, 4, 5, 6\}$ permet de définir six événements élémentaires correspondant chacun à la possibilité d'obtenir un résultat du jet égal à l'un des six nombres 1,2,3,4,5 ou 6.

Il est cependant possible de considérer des propriétés du résultat d'une expérience aléatoire qui sont possédées par plusieurs éléments de $\Omega$. On parle alors d'événement composé ou d'événement tout court. En reprenant l'exemple du jet de dé, l'apparition d'un nombre pair peut se réaliser par l'obtention du 2, du 4 ou du 6. Il est alors commode de représenter un événement défini par une propriété du résultat d'une expérience comme le sous-ensemble des instances de l'ensemble fondamental $\Omega$ qui possèdent cette propriété. Par exemple, dans l'expérience du lancer de dé, il est possible de considérer l'événement pour lequel la face supérieure du dé est un nombre pair. Dans ce cas, l'ensemble $D = \{2, 4, 6\}$ associé à l'événement est un sous-ensemble de $\Omega = \{1, 2, 3, 4, 5, 6\}$, l'ensemble des instances. On notera alors $D \subseteq \Omega$.

Dans le cadre des modèles mathématiques qui calculent le degré de cohésion des $N$-grams positionnels, un événement peut être considéré comme l'ensemble des $N$-grams positionnels de $\Omega$ qui partagent une certaine propriété. Par exemple, un événement peut consister à déterminer l'ensemble des $N$-grams positionnels de $\Omega$ qui partagent le même ensemble de $N$ UTs données. Ainsi, il est possible de restreindre l'espace des instances pour un cas particulier — un événement. Nous verrons dans les parties suivantes comment nous utiliserons cette propriété dans le cadre de l'Expectative Mutuelle.

Formellement, les bases de la théorie des probabilités dépendent d'un ensemble particulier noté $\mathcal{A}$. $\mathcal{A}$ est l'ensemble de tous les événements que l'on peut associer à une expérience aléatoire. Il est donc défini par l'ensemble des parties de $\Omega$ et noté de la forme suivante en 5.10.

$$\mathcal{A} \quad = \quad \mathcal{P}(\Omega) \tag{5.10}$$

Nous remarquons que deux événements particuliers sont associés à toute expérience aléatoire : l'événement certain représenté par l'ensemble $\Omega$ lui-même et l'événement impossible déterminé par la partie vide $\emptyset$.

**Fonction de Probabilité**

Assigner une probabilité à chaque événement possible consiste à mesurer la vraisemblance ou la chance qu'a cet événement de se produire. Les probabilités sont donc des nombres qui varient entre 0 et 1, 0 indiquant l'impossibilité d'un événement — l'événement correspond à $\emptyset$ — et 1 la certitude d'un événement — l'événement correspond à $\Omega$. Prenons l'exemple du jet d'une pièce de monnaie. L'ensemble fondamental est $\Omega = \{face, pile\}$ et l'ensemble des événements possibles est $\mathcal{A} = \mathcal{P}(\Omega) = \{\emptyset, \{face\}, \{pile\}, \Omega\}$. Dans le cadre d'une pièce bien équilibrée, la probabilité de voir le côté pile ou le côté face de la pièce — deux événements — sont équivalents et valent $\frac{1}{2}$. D'autre part, on accorde à la certitude que l'un ou l'autre des résultats se présente dans tous les cas — événement certain — une probabilité égale à 1. Enfin, si l'on demande la probabilité pour que la pièce retombe sur sa tranche, cet événement est considéré comme impossible et sa probabilité est nulle. Formellement, une fonction discrète de probabilité est une fonction $P(.)$ définie sur l'espace des événements $\mathcal{A}$ défini lui-même à partir de $\Omega$ comme suit.

$$P : \mathcal{A} \quad \rightarrow \quad [0, 1] \tag{5.11}$$

De plus, la fonction de probabilité $P(.)$ se doit de respecter les deux propriétés élémentaires suivantes.

$P(\Omega) = 1$

$\forall D_j \in \mathcal{A} \wedge \forall j \neq k, D_j \cap D_k = \emptyset, P\left(\bigcup_{j=1}^{\infty} D_j\right) = \sum_{j=1}^{\infty} P(D_j)$

En ce qui concerne la valeur de la fonction $P(.)$, elle est définie par le quotient entre le nombre de succès de l'événement considéré — le nombre d'instances qui satisfont les conditions de l'événement — et le nombre d'instances qui supportent l'expérience. Ainsi, si l'on note $D$ un événement et $|D|$ le nombre d'éléments de $D$ c'est-à-dire le nombre de succès pour l'événement considéré, la probabilité de l'événement $D$ notée $P(D)$ est définie dans l'équation 5.12.

$$\forall D \in \mathcal{A}, P(D) \quad = \quad \frac{|D|}{|\Omega|} \tag{5.12}$$

Reprenons le cas du lancer de dé[6]. Si l'on note $D$ l'événement qui consiste à voir un nombre pair sur la face supérieure du dé — $D = \{2, 4, 6\}$, $P(D)$ vaut $\frac{3}{6}$. Dans le cadre des modèles $N$-gram positionnels, nous pouvons définir de la même façon la probabilité de l'événement $D$ qui consiste à voir $N$ UTs données dans un $N$-gram positionnel. Cette probabilité est le résultat du quotient entre le nombre de $N$-grams positionnels qui contiennent l'ensemble des $N$ UTs et le nombre d'instances de $\Omega$ c'est-à-dire l'ensemble des $T - 2.F$ $N$-grams positionnels possibles — $T$ étant la taille de l'énoncé et $F$ la taille de l'environnement immédiat. La probabilité de l'événement $D$ vaut donc $\frac{|D|}{T-2.F}$.

### Probabilité Conditionnelle

La probabilité affectée à un événement dépend de l'information fournie par l'ensemble fondamental $\Omega$. Cependant, il se peut que des informations supplémentaires viennent modifier notre connaissance du problème étudié, et par voie de conséquence, les probabilités associées aux événements de $\Omega$. Si l'on revient à l'expérience du lancer de dé, l'événement $D$ représentant l'obtention d'un nombre pair donne origine à la probabilité $P(D) = \frac{3}{6} = \frac{1}{2}$. Or, si l'on indique que le résultat du lancer de dé est un nombre inférieur ou égal à 3, cette information va modifier l'ensemble des résultats possibles. En effet, ce dernier n'est plus l'ensemble fondamental $\Omega$ mais le sous-ensemble $\mathcal{B} = \{1, 2, 3\}$. Il y aura donc une probabilité égale à $\frac{1}{3}$ d'obtenir un nombre pair étant donnée l'information selon laquelle le résultat est inférieur ou égal à 3. Cette probabilité est dite conditionnelle. Dans le cadre des modèles $N$-gram positionnels et de l'Expectative Mutuelle, l'apparition d'une

[6]Dans ces conditions d'expérience, nous considérons que le dé n'est pas pipé et que chaque résultat de l'expérience est équiprobable.

UT peut être conditionnée par l'ensemble des autres $N-1$ UTs qui constituent un $N$-gram positionnel considéré. Cette situation sera dénotée **hypothèse conditionnelle**. En effet, l'ensemble fondamental $\Omega$ est ainsi réduit au sous-ensemble des instances qui respectent cette condition. Suivant la définition de fonction de probabilité donnée dans la partie précédente, la probabilité de l'événement $D$ qui consiste à voir une UT donnée dans un $N$-gram positionnel est notée $P(D)$ et déterminée par $P(D) = \frac{|D|}{|\Omega|}$. Cependant, dans le cadre de l'hypothèse conditionnelle, ce même événement est conditionné par l'occurrence du $N-1$-gram constitué par les $N-1$ UTs qui se trouvent dans l'environnement immédiat de l'UT considérée. Ainsi, l'ensemble fondamental $\Omega$ est réduit à l'ensemble des $N$-grams positionnels qui contiennent les $N-1$ UTs définies dans la condition. Si l'on note $D'$ l'événement qui consiste à vérifier la présence dans un $N$-gram positionnel du $N-1$-gram qui constitue le contexte de l'UT considérée, la probabilité de voir apparaître l'UT conditionnée par l'occurrence de son $N-1$-gram complémentaire est notée $P(D|D')$ et déterminée par l'équation suivante.

$$P(D|D') = \frac{P(D \cap D')}{P(D')} \tag{5.13}$$

Dans ce cas, l'ensemble $D \cap D'$ correspond à l'ensemble des $N$-grams positionnels qui contiennent à la fois l'UT considérée de l'événement initial et le $N-1$-gram positionnel qui forme son contexte pour l'ensemble des positions définies — son $N-1$-gram complémentaire.

**Expectative Mutuelle**

A partir de la définition de l'espace probabilisé ($\Omega$, $\mathcal{A}$, $P(.)$) et de la notion de probabilité conditionnelle, la mesure d'Expectative Mutuelle peut être théoriquement bien formulée. Dans ce cadre, l'Expectative Mutuelle est définie en fonction de l'approche non paramétrique qui comme le souligne I. Dagan [83] semble être la meilleure solution pour attaquer les problèmes soulevés par le langage.

L'idée de l'Expectative Mutuelle se rapproche des notions de *support* et de *confiance* qui ont été introduites par R. Agrawal [44] dans le cadre de l'extraction de règles d'association à partir de grandes bases de données non textuelles. Dans ce contexte, la notion de *support*

peut être vue comme une simple fréquence relative et la notion de *confiance* comme une probabilité conditionnelle moyenne qu'il est nécessaire d'estimer. Plus particulièrement, le *support* d'un $N$-gram positionnel sera sa fréquence relative et sa *confiance* sera son Expectative Normalisée. Nous définissons donc dans un premier temps le concept d'Expectative Normalisée qui utilisera l'estimation non paramétrique de la fonction de probabilité $P(.)$.

### 5.2.2 Expectative Normalisée

L'idée de base de l'Expectative Normalisée — EN — est d'évaluer le coût de la perte d'une UT dans un $N$-gram positionnel. Ainsi, plus une suite d'UTs est figée et témoigne d'une forte cohésion, moins cette séquence accepte la perte d'un de ses constituants et plus la valeur de l'EN doit être élevée. Dans ce cadre, nous définissons l'EN d'un $N$-gram positionnel d'UTs comme étant l'expectative moyenne de voir apparaître une UT dans une position donnée sachant que les $N-1$ autres UTs du $N$-gram positionnel apparaissent dans son environnement immédiat contraintes par leur position.

#### Exemple

Afin d'éclairer le lecteur sur la notion d'EN, nous proposons d'illustrer notre discours à partir d'un exemple. Prenons l'exemple du trigram positionnel [0 *Traité* 1 *de* 2 *Maastricht*]. Son Expectative Normalisée doit représenter l'expectative moyenne de voir apparaître :

- le nom *Traité* sachant que la séquence *de Maastricht* conditionne son apparition,
- la préposition *de* entre les deux noms *Traité* et *Maastricht* et
- le nom *Maastricht* sachant que la suite *Traité de* contraint son occurrence.

Nous illustrons cette situation dans le tableau 5.2 dans lequel chaque ligne correspond à une expectative particulière et le trait "—" à l'UT à prédire.

| Expectative d'apparition de l'UT | Sachant l'occurrence de |
|---|---|
| Traité | [0 — 1 de 2 Maasstricht] |
| de | [0 Traité 1 — 2 Maasstricht] |
| Maastricht | [0 Traité 1 de 2 — ] |

TAB. 5.2 – Expectatives d'un 3-gram

Cette situation est facilement généralisable. En effet, chaque $N$-gram positionnel implique la définition de $N$ prédictions — une pour chaque UT du $N$-gram positionnel considéré. Ainsi, l'idée de l'EN est de mesurer le coût de la perte d'une UT d'un $N$-gram positionnel donné. Nous approfondissons cette notion dans la partie suivante.

**Idée de Base**

L'Expectative Normalisée d'un $N$-gram positionnel peut être définie comme un ensemble de $N$ prédictions conditionnelles. Nous allons exemplifier cette situation.

Selon la notion de prédiction de Markov [42], la présence d'une UT est déterminée par l'ensemble des UTs voisines qui la précédent. Cette règle peut être formulée par la tentative d'estimer la fonction de probabilité $P(.)$ suivante[7].

$$P([u_N]|[u_1 u_2 \dots u_{N-1}]) \tag{5.14}$$

Cette probabilité conditionnelle représente la prédiction de Markov. Elle mesure la probabilité d'apparition de l'UT $u_N$ sachant que la séquence continue $[u_1 u_2 \dots u_{N-1}]$ la précède dans l'énoncé. Ce résultat peut être interprété comme la force qui lie $u_N$ à sa séquence précédente voisine. Ainsi, si la probabilité correspondante est proche de 1, les UTs sont parfaitement liées. Dans le cas contraire — la probabilité vaut 0 — l'UT $u_N$ est indépendante de la suite voisine. De même, plus la probabilité est forte et plus la perte de l'UT pénalise la cohésion totale du $N$-gram positionnel.

Cependant, la prédiction de Markov selon laquelle l'apparition d'une UT dépend uniquement des $N-1$ UTs qui la précédent, n'est pas suffisante pour représenter toutes les formes d'associations textuelles. Par conséquent, sa généralisation s'impose. Afin de mesurer les forces de cohésion qui lient entre elles toutes les UTs d'un $N$-gram positionnel, il est nécessaire d'évaluer toutes les prédictions possibles à l'intérieur d'un $N$-gram positionnel

[7] En théorie, nous devrions noter cette probabilité de la forme suivante : $P([u_1 u_2 \dots u_{N-1} u_N]|[u_1 u_2 \dots u_{N-1}])$. Mais pour des raisons de légibilité, nous avons préféré simplifier cette notation n'utilisant ainsi dans la partie non conditionnelle de la probabilité que l'élément à prédire.

considéré. C'est ce que l'on appellera **l'hypothèse conditionnelle**. Dans ce cadre, chaque prédiction sera une **prédiction conditionnelle** donnée. Ainsi, une prédiction revient à calculer la chance d'apparition d'une UT sachant que les autres $N-1$ UTs du $N$-gram positionnel considéré apparaissent dans son environnement immédiat. L'élargissement de la prédiction de Markov implique par conséquent le calcul de $N$ prédictions conditionnelles et l'estimation de la fonction de probabilité $P(.)$ suivante où chaque $u_i$ représente une UT et l'indice $i$ varie de 1 à $N$ définissant l'ordre des UTs.

$$NPrédictions \begin{cases} P([u_1]|[u_2u_3u_4 \dots u_N]) \\ P([u_2]|[u_1u_3u_4 \dots u_N]) \\ P([u_3]|[u_1u_2u_4 \dots u_N]) \\ \dots \\ P([u_N]|[u_1u_2u_3 \dots u_{N-1}]) \end{cases}$$

Or, l'introduction de l'hypothèse conditionnelle implique la nécessité de représenter des suites interrompues d'UTs et par conséquent l'introduction des modèles $N$-gram positionnels. En effet, les séquences $[u_1\ u_3\ u_4\ ...\ u_N]$ et $[u_1\ u_2\ u_4\ ...\ u_N]$ sont des suites d'UTs non continues puisque dans le premier cas l'UT $u_2$ qui se trouve entre les UTs $u_1$ et $u_3$ dans l'énoncé n'est pas retenue dans la séquence et dans le deuxième cas, le même phénomène s'applique à l'UT $u_3$ qui se trouve entre les UTs $u_2$ et $u_4$.

Avant de continuer notre exposé, nous déterminons avec précision les notations que nous utiliserons dans la suite de ce rapport. Ainsi, nous avons l'égalité notationnelle suivante où la deuxième ligne est la notation théoriquement correcte et la première la notation utilisée.

$$\begin{gathered} P([p_{j_1j_k}u_{j_k}]|[p_{j_1j_1}u_{j_1} \dots p_{j_1j_{k-1}}u_{j_{k-1}}p_{j_1j_{k+1}}u_{j_{k+1}} \dots p_{j_1j_n}u_{j_n}]) \\ \equiv \\ P([p_{j_1j_1}u_{j_1}p_{j_1j_2}u_{j_2} \dots p_{j_1j_k}u_{j_k} \dots p_{j_1j_n}u_{j_n}]|[p_{j_1j_1}u_{j_1} \dots p_{j_1j_{k-1}}u_{j_{k-1}}p_{j_1j_{k+1}}u_{j_{k+1}} \dots p_{j_1j_n}u_{j_n}]) \end{gathered} \tag{5.15}$$

Suivant cette notation, nous définissons donc les $N$ prédictions conditionnelles dans le cas des modèles $N$-gram positionnels à partir des $N$ probabilités conditionnelles suivantes.

$$NPrédictions\ conditionnelles \begin{cases} P([p_{21}u_1]|[p_{22}u_2p_{23}u_3p_{24}u_4\ldots p_{2N}u_N]) \\ P([p_{12}u_2]|[p_{11}u_1p_{13}u_3p_{14}u_4\ldots p_{1N}u_N]) \\ P([p_{13}u_3]|[p_{11}u_1p_{12}u_2p_{14}u_4\ldots p_{1N}u_N]) \\ \ldots \\ P([p_{1N}u_N]|[p_{11}u_1p_{12}u_2p_{13}u_3\ldots p_{1(N-1)}u_{N-1}]) \end{cases} \quad (5.16)$$

Ainsi, chacune des probabilités conditionnelles représente une prédiction conditionnelle particulière à l'intérieur d'un $N$-gram positionnel comme nous l'illustrons dans le tableau 5.3 où la marque "$-$" correspond à l'UT extraite du $N$-gram — UT prédite.

| UT extraite | Contexte de $N-1$ UTs |
|---|---|
| $u_1$ | $[p_{11} - p_{12}u_2p_{13}u_3p_{14}u_4\ldots p_{1N}u_N]$ |
| $u_2$ | $[p_{11}u_1p_{12} - p_{13}u_3p_{14}u_4\ldots p_{1N}u_N]$ |
| $u_3$ | $[p_{11}u_1p_{12}u_2p_{13} - p_{14}u_4\ldots p_{1N}u_N]$ |
| $\ldots$ | $\ldots$ |
| $u_N$ | $[p_{11}u_1p_{12}u_2p_{13}u_3\ldots p_{1(N-1)}u_{N-1}p_{1N}-]$ |

TAB. 5.3 – $N$ prédictions conditionnelles

Par définition, l'EN doit rendre compte des $N$ prédictions conditionnelles existantes dans un $N$-gram positionnel à partir d'une seule mesure normalisée. Formellement, l'EN doit normaliser l'ensemble des $N$ probabilités conditionnelles définies pour chacune des $N$ prédictions.

### Probabilité Conditionnelle

Suivant la définition de la probabilité conditionnelle énoncée précédemment, les deux équations suivantes définissent les $N$ prédictions conditionnelles qui correspondent aux $N$ probabilités conditionnelles mentionnées dans la partie précédente. Nous rappelons pour des raisons de compréhension que le $N$-gram positionnel $[p_{11}\ u_1\ p_{12}\ \ldots p_{1N}\ u_N]$ est équivalent au $N$-gram positionnel $[p_{22}\ u_2\ p_{21}\ u_1\ \ldots p_{2N}\ u_N]$ au changement d'unité de référence près.

$$P([p_{21}u_1]|[p_{22}u_2\ldots p_{2N}u_N]) = \frac{P([p_{22}u_2p_{21}u_1\ldots p_{2N}u_N])}{P([p_{22}u_2\ldots p_{2N}u_N])} \quad (5.17)$$

$$\forall i, i = 2..1, P([p_{1i}u_i]|[p_{11}u_1 \dots p_{1(i-1)}u_{(i-1)}p_{1(i+1)}u_{(i+1)} \cdots p_{1N}u_N]) =$$

$$\frac{P([p_{11}u_1p_{12}u_2 \dots p_{1N}u_N])}{P([p_{11}u_1 \dots p_{1(i-1)}u_{(i-1)}p_{1(i+1)}u_{(i+1)} \cdots p_{1N}u_N])} \tag{5.18}$$

Du fait des caractéristiques du matériel textuel, nous devons maintenant définir l'estimateur de la fonction de probabilité $P(.)$ pour calculer les valeurs de chacune des $N$ prédictions conditionnelles. A partir de la définition de fréquence relative, les $N$ probabilités conditionnelles formulées précédemment peuvent être facilement estimées. Nous considérons d'abord un cas particulier pour ensuite définir génériquement les $N$ prédictions conditionnelles. La probabilité $P([p_{12}u_2]|[p_{11}u_1p_{13}u_3 \dots p_{1N}u_N])$ équivaut au quotient entre les deux probabilités du $N$-gram positionnel $[p_{11}u_1p_{12}u_2p_{13}u_3 \dots p_{1N}u_N]$ et du $N-1$-gram $[p_{11}u_1p_{13}u_3 \dots p_{1N}u_N]$ — sous-groupe de rang $N-1$ du $N$-gram positionnel précédent. Ce ratio est déterminé et développé en fonction de la fréquence relative dans l'équation suivante.

$$\begin{aligned} P([p_{12}u_2]|[p_{11}u_1p_{13}u_3 \dots p_{1N}u_N]) &= \frac{P([p_{11}u_1p_{12}u_2p_{13}u_3 \dots p_{1N}u_N])}{P([p_{11}u_1p_{13}u_3 \dots p_{1N}u_N])} \\ &= \frac{\frac{k([p_{11}u_1p_{12}u_2p_{13}u_3 \dots p_{1N}u_N])}{T-2.F}}{\frac{k([p_{11}u_1p_{13}u_3 \dots p_{1N}u_N])}{T-2.F}} \\ &= \frac{k([p_{11}u_1p_{12}u_2p_{13}u_3 \dots p_{1N}u_N])}{k([p_{11}u_1p_{13}u_3 \dots p_{1N}u_N])} \end{aligned} \tag{5.19}$$

En répétant le même procédé pour chacune des $N$ probabilités conditionnelles, il est possible d'estimer la fonction de probabilité $P(.)$ de la forme suivante.

$$P([p_{21}u_1]|[p_{22}u_2 \dots p_{2N}u_N]) = \frac{k([p_{22}u_2p_{21}u_1 \dots p_{2N}u_N])}{k([p_{22}u_2 \dots p_{2N}u_N])} \tag{5.20}$$

$$\forall i, i = 2..1, P([p_{1i}u_i]|[p_{11}u_1 \dots p_{1(i-1)}u_{(i-1)}p_{1(i+1)}u_{(i+1)} \cdots p_{1N}u_N]) =$$

$$\frac{k([p_{11}u_1p_{12}u_2 \dots p_{1N}u_N])}{k([p_{11}u_1 \dots p_{1(i-1)}u_{(i-1)}p_{1(i+1)}u_{(i+1)} \cdots p_{1N}u_N])} \tag{5.21}$$

A partir de l'estimation de la fonction de probabilité $P(.)$ et par conséquent des $N$ prédictions conditionnelles, nous abordons à ce stade de notre étude, l'étape de normalisation proprement dite qui permet de considérer l'ensemble des probabilités conditionnelles en une seule mesure d'association, l'EN. Dans ce but, nous introduisons dans un premier temps le concept d'argument moyen conditionnel qui va nous permettre de considérer un et un seul dénominateur moyen pour tout $N$-gram positionnel.

**Argument Moyen Conditionnel**

L'analyse des $N$ probabilités conditionnelles des équations précédentes montre que les numérateurs restent inchangés d'une probabilité à l'autre et que seuls les dénominateurs changent. En effet, les deux numérateurs représentent la fréquence du même $N$-gram positionnel noté de forme différente par un changement d'unité de référence. Par conséquent, la normalisation qui consiste à considérer les $N$ prédictions conditionnelles en une seule mesure d'association doit prendre en compte le calcul de l'argument moyen conditionnel — AMC — représentant en un seul terme les $N$ dénominateurs considérés. Ainsi, l'AMC est défini comme étant la moyenne arithmétique des $N$ événements conditionnels des $N$ probabilités considérées[8]. Chaque dénominateur représente la fréquence d'un sous-groupe de rang $N-1$ du $N$-gram positionnel considéré en numérateur. L'AMC est donc la moyenne arithmétique de tous les sous-groupes de rang $N-1$ du $N$-gram positionnel sujet aux $N$ prédictions conditionnelles. Ainsi, l'AMC est défini dans l'équation 5.22 pour un $N$-gram positionnel générique.

$$AMC([p_{11}u_1 \dots p_{1N}u_N]) = \frac{1}{N}\sum_{i1=1}^{2}\sum_{i2=i1+1}^{3}\cdots\sum_{\substack{i(N-1)=\\ i(N-2)+1}}^{N} k([p_{i1i1}u_{i1}p_{i1i2}u_{i2}\dots p_{i1i(N-1)}u_{i(N-1)}]) \quad (5.22)$$

A partir de la définition de l'argument moyen conditionnel, nous définissons donc l'Expectative Normalisée.

[8] D'où le terme d'argument moyen conditionnel.

**Définition**

Par définition, l'EN d'un $N$-gram positionnel est l'expectative moyenne de voir apparaître une UT dans une position donnée sachant que les $N-1$ autres UTs du $N$-gram positionnel apparaissent dans son environnement immédiat contraintes par leur position. Par conséquent, l'EN doit rendre compte en une seule mesure normalisée des $N$ prédictions conditionnelles qu'implique un $N$-gram positionnel. L'EN peut ainsi être définie par l'équation 5.23 grâce à l'introduction de l'AMC dans la définition générale de la probabilité conditionnelle.

$$EN([p_{11}u_1 \dots p_{1N}u_N]) = \frac{k([p_{11}u_1 \dots p_{1N}u_N])}{AMC([p_{11}u_1 \dots p_{1N}u_N])} \tag{5.23}$$

Du fait de sa définition bien fondée sur l'espace probabilisé ($\Omega$, $\mathcal{A}$, $P(.)$), l'Expectative Normalisée démontre un certain nombre de propriétés intéressantes dont la récursivité. En effet, cette propriété peut s'avérer être un atout fondamental pour son implémentation informatique bénéficiant de l'étude d'algorithmes performants. Cependant, nous n'avons pas encore étudié cette caractéristique au niveau technique. Néanmoins, les calculs théoriques sont proposés en annexe A. Dans la partie suivante, nous définissons finalement la mesure d'Expectative Mutuelle basée sur les concepts d'Expectative Normalisée et de fréquence relative correspondant respectivement aux notions de *support* et de *confiance* introduites par R. Agrawal [44].

### 5.2.3 Expectative Mutuelle

L'Expectative Mutuelle — EM — peut être considérée comme une mesure qui allie les concepts de *confiance* et de *support* introduits par Agrawal [44] dans le cadre de l'extraction de règles d'association à partir de grandes bases de données non textuelles. Nous verrons dans cette partie comment la notion de support — fréquence relative — est combinée à l'Expectative Normalisée — mesure de confiance — pour formuler l'EM. Dans un premier temps, nous introduisons la notion de règle d'association qui sert de motivation à l'introduction de cette nouvelle mesure de l'approche non paramétrique.

**Règle d'Association**

Dans le cadre de l'extraction de classes de régularité à partir de bases de données de grande taille, R. Agrawal [44] introduit le concept de règle d'association. Une règle d'association n'est autre qu'une expression $X \Rightarrow Y$ où $X$ et $Y$ sont deux ensembles d'éléments et dont le sens peut être exprimé par la proposition suivante : *"l'ensemble des transactions d'une base de données qui contiennent $X$ tendent à contenir $Y$"*. Un exemple d'une telle règle peut être de stipuler que 98% des clients qui achètent des pneus et des accessoires automobiles ont également besoin de services de garage. Dans le cadre de notre recherche, on peut facilement concevoir une règle d'association qui rend compte du degré d'attraction liant entre eux les éléments d'un $N$-gram positionnel. Dans ce cas, on dira que l'ensemble des $N$-grams positionnels qui contiennent un certain $N-1$-gram positionnel tendent à contenir une certaine UT — une prédiction conditionnelle est ainsi mise en évidence.

Dans le but de certifier la pertinence d'une règle d'association, R. Agrawal affirme qu'une règle d'association est solide — c'est-à-dire pertinente — si son support est au moins égal à un support minimal donné et si sa confiance est au moins égale à une confiance minimale donnée. Nous nous baserons sur ces deux concepts pour formuler la mesure d'EM. Par conséquent, nous définissons formellement les concepts de règle d'association, de support et de confiance.

**Règle d'association :** Soit $I$ un ensemble de littéraux appelés items tel que $I = \{i_1, i_2, ..., i_t\}$. Une règle d'association $R$ est une implication de la forme suivante :

$$R \quad = \quad X \Rightarrow Y \quad \text{telle que} \quad X \subseteq I, Y \subseteq I, X \cap Y = \emptyset \tag{5.24}$$

Si l'on note un ensemble de transactions $D = \{t_1, t_2 \ldots t_m\}$ tel que $\forall i = 1..m, t_i \subseteq I$ pour lesquelles la règle d'association $R$ est valide, on dira que les transactions de $D$ contenant $X$ tendent à contenir $Y$. Ainsi, si l'on définit $I$ comme l'ensemble des UTs distinctes de l'énoncé initial — le vocabulaire — et $D$ l'ensemble des $N$-grams positionnels pour une distribution donnée, une règle d'association revient à formuler les forces d'attraction qui lient entre elles les UTs d'un $N$-gram positionnel. En effet, si $X$ est l'ensemble des sous-groupes de rang $N-1$ d'un $N$-gram positionnel donné et $Y$ le singleton qui représente l'UT extraite du $N$-gram positionnel, la règle d'association $R$ revient à formuler la dépendance

qui existe entre l'UT extraite et les autres $N-1$ UTs du $N$-gram positionnel. On notera tout de même que la contrainte imposée par R. Agrawal telle que $X \cap Y = \emptyset$ n'est pas imposée par les caractéristiques de notre étude.

**Confiance :** La mesure de confiance d'une règle d'association $R = X \Rightarrow Y$ est le pourcentage de transactions de $D$ qui contiennent $Y$ parmi celles qui contiennent $X$ et est déterminée par le quotient suivant.

$$Confiance(X \Rightarrow Y) = \frac{Fréquence(X \cup Y)}{Fréquence(X)} \tag{5.25}$$

A la lumière de cette définition, il est clair que dans le cadre de notre recherche, la mesure de confiance n'est autre qu'une prédiction conditionnelle particulière. Dans ces conditions, l'EN peut donc être considérée comme une mesure de confiance normalisée qui évalue les forces de cohésion qui sont implicites dans les $N$-grams positionnels.

**Support :** Le support d'une règle d'association $R = X \Rightarrow Y$ est le nombre de transactions de $D$ contenant $X$ et $Y$ par rapport au nombre total de transactions de $D$. Le support est donc déterminé par le quotient suivant où $|D|$ représente le nombre d'éléments de l'ensemble $D$ — son cardinal.

$$Support(X \Rightarrow Y) = \frac{Fréquence(X \cup Y)}{|D|} \tag{5.26}$$

Dans le cadre des mesures d'association, le support n'est autre que la fréquence relative d'un $N$-gram positionnel générique. Or, plusieurs études montrent que la fréquence est un facteur de cohésion primordial pour l'identification d'associations textuelles pertinentes. Par exemple, J. Justeson [22] et B. Daille [21] ont mis en évidence dans leurs travaux sur les associations lexicales que la fréquence est l'un des critères fondamentaux pour l'identification d'unités polylexicales. G. Gross [51] corrobore cette opinion et affirme que la compréhension d'une unité lexicale complexe est un processus itératif, étant nécessaire qu'une unité soit prononcée plus d'une fois pour que sa compréhension soit possible. Dans le contexte des associations syntaxiques, Argamon-Engelson [84] argumente que la répétition d'une séquence d'étiquettes morpho-syntaxiques est un bon indicateur

de pertinence. Ainsi, la prise en compte de la fréquence relative d'un $N$-gram positionnel lors de la quantification de son degré de cohésion nous paraît être un point fondamental d'une analyse correcte. Dans ces conditions, la combinaison entre mesure de confiance et support s'adapte particulièrement au cas spécifique du matériel textuel[9]. Nous définissons donc l'EM à partir de ces deux concepts. En effet, on dira qu'un $N$-gram positionnel démontre un degré de cohésion d'autant plus fort que sa confiance — EM — et son support — fréquence relative — sont élevés.

**Définition**

Selon R. Agrawal, une règle d'association est dite solide si sa confiance et son support sont au moins égaux à une valeur seuil donnée dans chacun des cas. Il est clair que plus une règle d'association est fréquente — son support est élevé — et plus la force qui lie ses éléments $X$ et $Y$ est forte — sa confiance est élevée, plus la règle est solide. Cependant, l'une de nos motivations principales met en évidence le refus de la définition de valeurs seuil. En effet, il n'existe pas de base théorique qui nous permet de formuler les valeurs des seuils envisagés. Ainsi, nous proposons une nouvelle définition pour le terme de règle d'association solide. On dira qu'une règle d'association est d'autant plus solide qu'elle est fréquente et dénote un taux de cohésion élevé. Dans le contexte de notre recherche, une règle d'association solide correspond à un $N$-gram positionnel qui démontre une forte cohésion par le biais de son Expectative Normalisée et dont la fréquence relative est élevée. Ainsi, nous définissons l'Expectative Mutuelle d'un $N$-gram positionnel comme étant le produit entre son Expectative Normalisée et sa fréquence relative. L'EM est déterminée dans l'équation suivante où $T$ est la taille de l'énoncé initial et $F$ la taille de l'environnement immédiat considéré.

$$EM([p_{11}u_1 \dots p_{1N}u_N]) = \frac{k([p_{11}u_1 \dots p_{1N}u_N])}{T - \lambda} \times EN([p_{11}u_1 \dots p_{1N}u_N]) \tag{5.27}$$

$$tel \quad que \quad \lambda = \begin{cases} F, N = 2 \\ 2.F, N > 2 \end{cases}$$

L'Expectative Mutuelle permet donc de mesurer le degré de cohésion implicite des $N$-grams positionnels sans être limitée aux associations binaires. Ainsi, il est possible de

[9] Agrawal ne mentionne pas le matériel textuel comme champ d'application des règles d'association.

classer n'importe quel $N$-gram positionnel — $\forall N, N \geq 2$ — suivant son degré de pertinence. Une fois mesurées les attirances entre UTs à l'intérieur d'un $N$-gram positionnel, la phase proprement dite d'extraction des associations textuelles candidates doit être menée à bien. Dans ce cadre, nous introduisons l'algorithme GenLocalMaxs qui propose une solution innovatrice pour l'identification d'associations textuelles candidates à partir de l'ensemble des $N$-grams positionnels pondérés selon leur degré de cohésion.

## 5.3 Identification d'Associations Textuelles

L'identification d'associations textuelles candidates peut être définie comme étant la tâche qui consiste à détecter, parmi l'ensemble des $N$-grams positionnels pondérés selon leur degré de cohésion, un sous-ensemble d'éléments qui partagent certaines caractéristiques propres au concept d'association textuelle. En particulier, on dira que ces $N$-grams positionnels sont pertinents. Par conséquent, l'extraction d'associations textuelles dépend intrinsèquement de la définition de l'ensemble de ces caractéristiques et donc de la notion de pertinence : Qu'est-ce qu'un $N$-gram positionnel pertinent ?

### 5.3.1 Concept de Pertinence

A partir de l'ensemble des séquences d'UTs pondérées selon leur degré de cohésion, la phase proprement dite d'extraction revient à définir un sous-ensemble de séquences pertinentes qui constituent l'ensemble des associations textuelles candidates[10]. La difficulté réside alors dans la définition du concept de pertinence d'une suite d'UTs. Paradoxalement, peu de travaux se sont attaqués à ce problème. En effet, la plupart des approches préconisent l'utilisation de valeurs seuil de fréquence ou de mesure d'association qui, comme nous allons le voir, sont loin d'être satisfaisantes.

Dans le cadre des associations lexicales, la plupart des approches se basent sur la définition de valeurs seuil qui permettent de diviser en deux groupes distincts l'ensemble des séquences d'UTs [22] [33] [23] [45] [31] [32] [34] [23] [28] [27] [21]. D'une part, les séquences d'UTs qui dépassent les valeurs seuil stipulées sont cataloguées dans la catégorie des unités complexes pertinentes alors que les autres suites d'UTs sont rejetées par les processus de sélection. Par exemple, K. Church et P. Hanks [27] proposent que tout

[10] Les suites d'UTs extraites ne forment pas forcément des associations textuelles "correctes". Par conséquent, on les appellera associations textuelles candidates.

couple d'UTs démontrant une valeur de coefficient d'association supérieure à 3 puisse être considéré association lexicale candidate. Dans le même ordre d'idée, F. Smadja [31] défend que la valeur seuil du score centré réduit doit être définie par l'utilisateur selon l'objectif prétendu du processus d'extraction. De son côté, S. Shimohata [32] définit une valeur seuil de fréquence plus une valeur seuil d'entropie imposant ainsi deux restrictions au processus de sélection des unités complexes. Cependant, il n'existe pas de fondement théorique permettant de définir ces valeurs limites. De ce fait, elles sont généralement imposées par l'expérimentation. Dans ce cadre, le genre, la longueur, le domaine ou la langue de l'énoncé sont autant de paramètres à prendre en compte lors de leur définition. Ainsi, le changement de l'une de ces données implique nécessairement le réajustement des valeurs seuil. Dans ces conditions, la définition de pertinence proposée par les valeurs seuil s'éloigne radicalement d'une solution "universelle". En effet, d'une part, les résultats du processus de sélection sont étroitement liés aux conditions de l'expérience considérée. D'autre part, la définition de l'ensemble des valeurs seuil — de fréquence et/ou de mesure d'association — nécessaires au processus d'acquisition reste une question ouverte formulée à la communauté scientifique.

La définition de valeurs seuil comme solution du processus d'extraction d'unités complexes pertinentes pose également un certain nombre de problèmes pratiques d'implémentation. Dans le cadre des mesures d'association binaires, l'utilisation de valeurs limites introduit un ensemble de contraintes qui réduit la couverture du processus d'acquisition par amorçage. En effet, dans ce contexte, l'acquisition d'associations de plus de deux UTs requiert un travail complémentaire où les paires d'association acquises initialement jouent le rôle d'amorce à l'identification d'associations de tailles supérieures. Ainsi, les résultats de la méthode d'amorçage dépendent fondamentalement des unités complexes binaires retenues lors de la première étape du processus de sélection. Or, cette sélection pose un certain nombre de problèmes du fait de l'utilisation de mesures d'association sensibles aux UTs fréquentes[11]. En particulier, certaines associations binaires qui contiennent des formes fréquentes ne seront pas sélectionnées pour l'initialisation du traitement itératif d'amorçage. Nous proposons un exemple afin d'illustrer nos propos. L'identification du terme complexe *Traité de Maastricht* dépend de l'identification préalable de l'association binaire *Traité de*. Or, le degré d'attraction entre *Traité* et *de* est généralement sous-évalué

[11] Nous le verrons en détail dans le chapitre 8 de ce rapport.

par les mesures d'association du fait de la forte fréquence de la préposition *de* dans l'ensemble du corpus. Ainsi, l'unité complexe *Traité de Maastricht* ne sera vraisemblablement pas identifiée du fait de la non identification du digram [*Traité de*]. Par conséquent, un certain nombre d'associations textuelles ne seront pas extraites diminuant ainsi le taux de couverture de la méthode de sélection par amorçage.

Parallèlement, dans le cadre des mesures d'association N-aires, l'utilisation de valeurs seuil impose le recours au post-traitement des résultats d'extraction comme le soulignent K. Frantzi et S. Ananiadou [33]. En effet, un certain nombre des séquences d'UTs extraites ne peuvent pas être considérées associations textuelles candidates. Ce sont généralement des sous-séquences d'unités complexes — ensembles d'UTs contenues dans des séquences d'UTs plus longues correspondant à des unités complexes. Afin d'illustrer nos dires, nous proposons un exemple caractéristique de post-traitement des résultats de sélection. Le nom propre *Valéry Giscard d'Estaing* peut être considéré sans trop de problèmes comme une association lexicale de trois UTs[12]. Dans ce contexte, les mesures d'association N-aires permettent de calculer son degré de cohésion indépendamment de l'identification du digram [*Valéry Giscard*]. Parallèlement, ces mesures permettent d'évaluer les forces d'attraction contenues dans la sous-séquence *Valéry Giscard*. Or, il est clair que la mesure de cohésion du trigram [*Valéry Giscard d'Estaing*] est très proche voire égale à celle du digram [*Valéry Giscard*]. Par conséquent, si le trigram est retenu par le processus de sélection, il est pratiquement certain que le digram sera lui aussi élu comme unité complexe pertinente[13]. Ainsi, le post-traitement des résultats s'impose comme une étape fondamentale du processus d'extraction mettant en évidence la nécessité de sélectionner les "vraies" des "fausses" unités complexes à l'intérieur d'un espace de séquences d'UTs réduit. Finalement, la propre nécessité du post-traitement réduit à néant la définition de pertinence basée sur les valeurs seuil.

Afin de correspondre à l'ensemble de nos motivations initiales, les valeurs seuil doivent être évitées et une nouvelle définition de pertinence doit être proposée. C'est pourquoi J. Silva *et al.* [30] proposent d'étudier l'environnement immédiat des séquences d'UTs afin de déterminer leur pertinence. Ainsi, une suite d'UTs sera définie pertinente ou non suivant les variations observées des mesures d'association de ses sous-groupes de rang $N-1$ et

[12]Nous suivons la définition de segmentation en formes graphiques donnée dans le chapitre précédent.
[13]Nous illustrerons ce phénomène dans le chapitre 8 de ce rapport.

de ses sur-groupes de rang $N+1$. Pour les modèles $N$-gram positionnels, nous proposons l'algorithme GenLocalMaxs qui définit un $N$-gram positionnel comme une association textuelle candidate s'il démontre un degré de cohésion plus fort que l'ensemble de ses sous-groupes de rang $N-1$ et de ses sur-groupes de rang $N+1$. L'idée sous-jacente du GenLocalMaxs est celle de déterminer les suites d'UTs qui perdent en cohésion lorsqu'une UT leur est ajoutée ou lorsqu'une UT leur est soustraite. Deux hypothèses fondamentales supportent cette définition de pertinence. D'une part, les mesures d'association montrent que plus une suite d'UTs est figée et cohésive, plus sa valeur de mesure d'association est forte[14]. D'autre part, les associations textuelles témoignent d'une forte cohésion dans leur environnement immédiat et sont par conséquent localement motivées. En effet, les associations textuelles sont des phénomèmes étroitement localisés.

Dans ce contexte, le GenLocalMaxs permet d'identifier les séquences d'UTs pertinentes sans recourir à la définition *ad hoc* de valeurs seuil qui mettent en évidence la nécessité de post-traitement des résultats ou bien diminuent le taux de couverture des unités complexes extraites. Le GenLocalMaxs peut donc être appliqué indépendamment de quelconque expérience considérée, ce qui le distingue fondamentalement des méthodes de valeurs seuil.

### 5.3.2 Algorithme GenLocalMaxs

L'algorithme GenLocalMaxs est une généralisation de l'algorithme LocalMaxs proposé par J. Silva dans le cadre des modèles $N$-gram classiques. L'idée de base est fondée sur la nouvelle définition de pertinence proposée précédemment et par conséquent sur la détection de maxima locaux de mesures d'association. Ainsi, un $N$-gram est une association textuelle candidate s'il démontre un degré de cohésion plus fort que l'ensemble de ses sous-groupes de rang $N-1$ et de ses sur-groupes de rang $N+1$.

Dans le cadre des $N$-grams contigus, l'ensemble des sous-groupes de rang $N-1$ d'un $N$-gram correspond à deux $N-1$-grams contigus construits à partir du $N$-gram de base considéré. Les deux $N-1$-grams sont respectivement obtenus par l'élimination de la première et de la dernière UT du $N$-gram. Ainsi, si l'on note $[u_1\ u_2\ u_3\ \ldots u_N]$ le $N$-gram contigu de base, les deux $N-1$-grams suivants définissent l'ensemble de ses sous-groupes

[14] La mesure d'entropie est l'une des exceptions à cette règle.

de rang $N-1$.

$$[u_2 u_3 \cdots u_N]$$
$$[u_1 u_2 \cdots u_{N-1}]$$

Parallèlement, l'ensemble des sur-groupes de rang $N+1$ d'un $N$-gram contigu correspond à l'ensemble des $N+1$-grams contigus qui contiennent le $N$-gram de base. Par conséquent, un sur-groupe de rang $N+1$ correspond au $N$-gram contigu de base auquel une UT a été ajoutée en début ou en fin de séquence. Ainsi, l'ensemble des sur-groupes de rang $N+1$ d'un $N$-gram contigu générique noté $[u_1\ u_2\ u_3\ \ldots u_N]$ peut être défini de la forme suivante à partir de l'ensemble de tous les $N+1$-grams contigus construits à partir de l'énoncé initial.

$$\forall i,\ [u_i\ u_1\ u_2\ u_3\ \ldots u_N]$$
$$\forall j,\ [u_1\ u_2\ u_3\ \ldots u_N\ u_j]$$

Cependant, il est clair que les notions de sous-groupe de rang $N-1$ et de sur-groupe de rang $N+1$ proposées par J. Silva sont incomplètes puisqu'elles ne prennent en compte que les variations d'UTs en extrémité des $N$-grams contigus. Dans le cadre des modèles $N$-gram positionnels, nous pouvons élargir la notion de sous-groupe de rang $N-1$ et de sur-groupe de rang $N+1$ à l'ensemble de tous les $N-1$-grams positionnels qui sont contenus dans le $N$-gram positionnel de base et l'ensemble de tous les $N+1$-grams positionnels qui contiennent le $N$-gram positionnel de base considéré[15]. Sur les bases du principe de pertinence énoncé précédemment et des modèles $N$-grams positionnels définis dans le chapitre précédent, nous définissons donc formellement l'algorithme GenLocalMaxs de la forme suivante. Nous rappelons que le GenLocalMaxs identifie comme pertinents les $N$-grams positionnels dont la mesure d'association est un maximum local.

Soient

- $W$ un $N$-gram positionnel construit pour un environnement immédiat de taille $F$,

[15]Nous avons défini ces notions dans le chapitre précédent.

- $\Omega^F_{N-1}$ l'ensemble de tous les sous-groupes de rang $N-1$ de $W$ pour l'environnement $F$,

- $\Omega^F_{N+1}$ l'ensemble de tous les sur-groupes de rang $N+1$ de $W$ dans ce même environnement,

- $assoc(.)$ une mesure d'association correspondant aux hypothèses formulées dans la définition de pertinence et

- $taille(.)$ une fonction qui renvoie le nombre d'UTs d'un $N$-gram $W$.

$W$ est une association textuelle candidate si les conditions suivantes sont respectées :

$$\forall W_{N-1} \in \Omega^F_{N-1}, \forall W_{N+1} \in \Omega^F_{N+1}$$

$$(taille(W) = 2 \wedge assoc(W) > assoc(W_{N+1})) \quad \vee$$

$$(taille(W) \neq 2 \wedge assoc(W) \geq assoc(W_{N-1}) \wedge assoc(W) > assoc(W_{N+1}))$$

Algorithme GenLocalMaxs

Nous illustrons le fonctionnement du GenLocalMaxs dans le cadre des associations lexicales à partir du 3-gram positionnel [*0 Traité 1 de 2 Maastricht*]. Celui-ci peut être défini comme pertinent si tous ses sous-groupes de rang 2 et tous ses sur-groupes de rang 4 mettent en évidence une valeur d'association inférieure à son degré de cohésion. L'ensemble des sous-groupes de rang 2 est représenté par les trois 2-grams positionnels suivants.

[*0 Traité 1 de*] [*0 Traité 2 Maastricht*] [*0 de 1 Maastricht*]

Or, il est clair que le degré de cohésion de chacun des 2-grams positionnels précédents est dans le pire des cas inférieur ou égal[16], mais jamais supérieur, à la mesure d'association

[16]On peut considérer que le degré de cohésion du 2-gram positionnel [*0 Traité 2 Maastricht*] est très proche de celui du 3-gram positionnel [*0 Traité 1 de 2 Maastricht*].

du 3-gram positionnel. En effet, la perte d'un des constituants du 3-gram positionnel diminue évidemment son intégrité et en aucun cas augmente sa cohésion. Dans le cas des sur-groupes de rang 4, le lecteur conviendra facilement qu'il n'existe aucune forme graphique qui puisse renforcer la cohésion implicite de la séquence *Traité de Maastricht*. Ainsi, tous les 4-grams positionnels contenant [*0 Traité 1 de 2 Maastricht*] devront mettre en évidence une mesure d'association strictement inférieure à celle du 3-gram positionnel. Dans ces conditions, le 3-gram positionnel [*0 Traité 1 de 2 Maastricht*] serait défini comme pertinent par l'algorithme GenLocalMaxs et serait élu comme potentielle association lexicale.

L'algorithme GenLocalMaxs propose ainsi une solution au problème de l'extraction d'unités pertinentes sans dépendre de la définition de valeurs seuil déterminées par expérimentation. En particulier, il se démarque par sa totale flexibilité d'application puisqu'il ne dépend ni de la longueur, ni de la langue, ni du domaine, ni du genre des énoncés considérés. Ainsi, n'importe quel énoncé d'UTs — caractères, formes graphiques ou étiquettes morpho-syntaxiques — peut être traité par le GenLocalMaxs. Parmi l'ensemble des propriétés du GenLocalMaxs, nous nous attarderons dans la partie suivante sur deux points fondamentaux qui le démarquent encore plus des autres méthodes de sélection : l'évaluation sur une unique plateforme de différentes mesures d'association et l'extraction d'associations textuelles obtenues par composition.

Cependant, cet algorithme n'est somme toute pas une solution optimale. En effet, le lecteur familiarisé avec le problème de l'extraction automatique d'unités polylexicales aura certainement remarqué une faiblesse de l'Algorithme GenLocalMaxs. En effet, il est possible qu'une séquence de $N$ UTs soit une unité polylexicale ainsi qu'un de ses sur-groupe de rang $N+1$. Par exemple, il est clair que les deux séquences *algorithme génétique* et *algorithme génétique binaire* sont deux termes composés. Néanmoins, le GenLocalMaxs n'est pas en mesure de les identifier tous les deux. Un seul pourra être extrait par l'algorithme de maximaux locaux. Cette situation n'est pas fâcheuse pour les textes de la Communauté Européenne où ce genre de phénomène est relativement rare. Mais elle l'est certainement pour les textes scientifiques qui utilisent cette facilité pour décrire des sous-concepts de termes. Dans ce cadre, d'autres recherches devraient être menées pour prendre en compte ce genre de phénomènes qui ne sont pas identifiés par le GenLocalMaxs.

### 5.3.3 Propriétés du GenLocalMaxs

Grâce à l'étude des variations locales des mesures d'association, le GenLocalMaxs permet d'évaluer toute mesure d'association qui respecte l'hypothèse de pertinence c'est-à-dire qui croît parallèlement au degré de cohésion. En effet, puisque l'algorithme ne dépend pas d'une mesure d'association particulière, il permet de comparer directement les résultats obtenus à partir de différentes mesures d'association. D'autre part, le GenLocalMaxs permet l'extraction directe d'associations textuelles obtenues par composition. Ainsi, sans processus computationnel supplémentaire, il élit des associations textuelles qui contiennent d'autres associations textuelles. Nous présentons chacun de ces deux aspects dans les deux prochaines parties.

#### Evaluation

Comme nous l'avons vu précédemment, les méthodes d'extraction qui consistent à définir des valeurs seuil pour séparer les séquences d'UTs pertinentes des autres sont peu flexibles et peu fiables. En effet, à chaque nouvelle expérience, il est nécessaire de réajuster ces valeurs limites. Du point de vue de l'évaluation des résultats, cette méthodologie pose un certain nombre de problèmes.

D'une part, il est difficile de répéter une expérience pour les cas où la langue, le domaine, la longueur ou le genre de l'énoncé changent. A chaque fois, l'utilisateur doit trouver la valeur seuil idéale qui maximise le taux de rappel et le taux de précision dans le cadre de son expérience. Les résultats de l'extraction sont donc biaisés par la définition plus ou moins fine des valeurs seuil. La difficulté de déterminer des valeurs seuil idéales est tellement grande que certains auteurs [85] [86] ont proposé l'application d'algorithmes génétiques pour définir les valeurs de fréquence et de mesure d'association qui maximisent la précision du processus d'extraction. En effet, il est pratiquement impossible de prétendre évaluer manuellement les valeurs idéales d'extraction.

D'autre part, la comparaison entre mesures d'association est rendue difficile par l'application de ces méthodes puisque pour chaque mesure il est nécessaire de définir une valeur seuil particulière. Or, s'il est difficile de définir les valeurs seuil pour une même mesure suivant les conditions d'expérience, il est bien plus compliqué de calculer les valeurs seuil optimales pour différentes mesures d'association. Dans ce cas, l'utilisateur

doit être familier à l'ensemble des mesures statistiques qu'il veut comparer. En effet, d'après notre expérience, chaque mesure d'association met en évidence un certain nombre de caractéristiques particulières qui compliquent la définition de valeurs seuil idéales.

Contrairement aux méthodes qui mettent en avant la définition de valeurs seuil, l'algorithme GenLocalMaxs propose une plateforme unique pour l'évaluation des résultats de l'extraction, ceci quelles que soient les conditions de l'expérience considérée. En effet, le lecteur attentif aura facilement remarqué que nous n'avons pas fait allusion directe à la mesure d'Expectative Mutuelle lors de l'introduction du GenLocalMaxs. Nous n'avons mentionné que la notion générale de mesure d'association par le biais de la fonction $assoc(.)$. En effet, le GenLocalMaxs est défini de forme générique pour n'importe quelle mesure d'association qui respecte l'hypothèse de pertinence. Ainsi, il suffit d'appliquer une nouvelle mesure d'association à chaque $N$-gram positionnel et de "faire tourner" le GenLocalMaxs pour extraire un nouvel ensemble d'associations textuelles candidates. Cette caractéristique est un atout fondamental en faveur du GenLocalMaxs qui permet la comparaison directe de mesures d'association aussi différentes que le coefficient Dice [45], le coefficient d'association [27], la Probabilité Conditionnelle Symétrique [30], le test $\Phi^2$ [29], le coefficient de vraisemblance LogLike [28] et bien entendu l'Expectative Mutuelle. Nous illustrerons ces résultats dans la troisième partie de ce rapport.

**Associations Textuelles Obtenues par Composition**

Les méthodes de sélection qui préconisent l'utilisation de valeurs seuil ne permettent pas une analyse fine des variations des mesures d'association. En effet, une seule valeur seuil est calculée pour l'ensemble des séquences d'UTs postulant ainsi l'existence d'une valeur "universelle" qui permet de définir l'ensemble des associations textuelles. Or, il est clair que cette approche n'est pas en mesure de rendre compte de l'ensemble des subtilités mises en évidence par les associations textuelles. En particulier, la définition de valeurs seuil de fréquence[17] rend difficile l'extraction d'associations textuelles obtenues par composition. Une association textuelle obtenue par composition peut être définie comme une association textuelle qui contient en son sein une ou plusieurs autres associations textuelles. Par exemple, dans le cadre des associations lexicales, la séquence *Le Président*

[17] La méthodologie dicte une fréquence minimale en-dessous de laquelle un $N$-gram ne peut pas être considéré comme pertinent.

*de la République Jacques Chirac* peut être interprétée comme la concaténation des deux unités lexicales complexes suivantes : *Le Président de la République* et *Jacques Chirac*. Celle-ci est alors définie comme une association textuelle obtenue par composition.

D'après cette définition, le lecteur conviendra sans difficulté que les associations textuelles obtenues par composition sont moins fréquentes que les associations textuelles qui la composent. En effet, dans le cas contraire, il n'existerait aucune base permettant de dissocier les constituants de la séquence et par conséquent la suite d'UTs formerait une et une seule association textuelle. En nous basant sur ce constat, il est clair que la définition de valeurs seuil de fréquence n'est pas une solution idéale pour l'extraction d'associations textuelles candidates. En effet, il est très probable qu'une grande proportion d'associations textuelles ne soit pas extraite du fait de la définition *ad hoc* d'une quelconque valeur de fréquence limite[18]. Cette affirmation est d'autant plus vraie que l'on verra dans la suite de ce rapport que la plupart des associations lexicales complexes n'apparaissent que deux fois dans les énoncés.

Contrairement à ce que proposent les autres méthodologies d'extraction, l'algorithme GenLocalMaxs permet l'identification d'associations textuelles obtenues par composition sans définir de valeurs seuil. En effet, comme il extrait les associations textuelles candidates à partir de l'analyse des contextes immédiats des $N$-grams positionnels, le GenLocalMaxs permet l'identification d'associations entre constituants qui sont eux-mêmes repérés pertinents. Ainsi, le processus de sélection se répète pour tous les $N$-grams positionnels jusqu'à ce qu'il n'existe plus de maximum local. Nous illustrons cette situation à partir de l'exemple de la figure 5.1

Dans cet exemple, l'unité lexicale complexe *Ministre de l'Intérieur Jean-Pierre Chevènement* serait élue à partir des deux associations *Ministre de l'Intérieur* et *Jean-Pierre Chevènement* puisqu'elle correspond à un maximum local. Nous donnons une explication simplifiée de ce résultat. Il est clair que le degré de cohésion de la séquence *Ministre de l'Intérieur* est supérieur à la mesure d'association de n'importe quel tetragram

[18] On notera que ce phénomène est également vrai pour le cas des valeurs seuil de mesure d'association puisqu'elles se basent sur des calculs de fréquence.

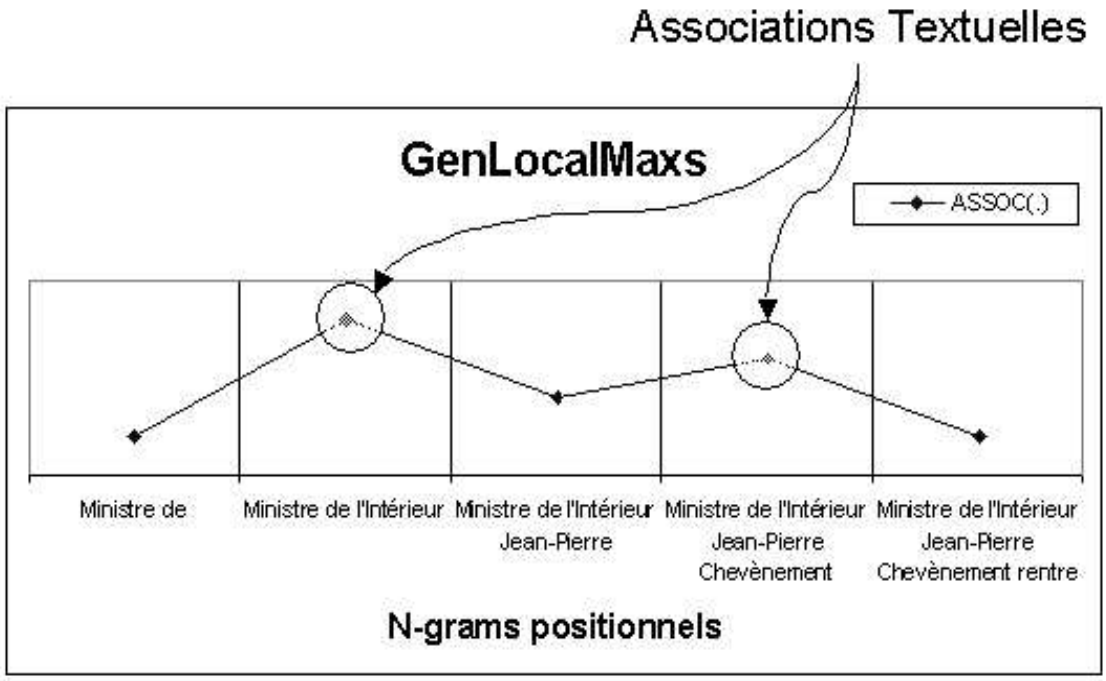


FIG. 5.1 – Exemple du GenLocalMaxs

positionnel[19] qui contient le 3-gram positionnel considéré. En effet, il n'existe aucune unité lexicale qui puisse renforcer la cohésion de l'ensemble des trois constituants. En particulier, dans le cadre de l'Union Européenne, il existe plusieurs ministres de l'intérieur et par conséquent plusieurs prénoms possibles qui peuvent se juxtaposer à l'association textuelle *Ministre de l'Intérieur*. Ainsi, la séquence *Ministre de l'Intérieur Jean-Pierre* démontre un degré de cohésion plus faible que celui du 3-gram positionnel [*0 Ministre 1 de 2 l'Intérieur*]. Parallèlement, la suppression de l'un des constituants du 3-gram positionnel se répercute négativement sur la mesure d'association de la séquence. En effet, le lecteur conviendra facilement que la séquence *Ministre de l'Intérieur* représente une suite d'UTs pertinente alors qu'aucun de ses sous-groupes de rang 2 ne peut être jugé digne d'intérêt. La mesure d'association du 3-gram positionnel [*0 Ministre 1 de 2 l'Intérieur*] est par conséquent un maximum local.

Cependant, la recherche de maxima locaux ne s'arrête pas à la découverte du premier maximum local pour une suite d'UTs donnée. Le GenLocalMaxs continue sa recherche pour des séquences plus longues. Ainsi, l'introduction du nom *Chevènement* à la suite de la séquence *Ministre de l'Intérieur Jean-Pierre* vient renforcer les liens qui associent tous

[19]Un tetragram positionnel est un 4-gram positionnel.

les constituants en relation. En effet, la probabilité de voir apparaître *Chevènement* après *Ministre de l'Intérieur Jean-Pierre* est très grande voire même totale[20]. Le pentagram [*0 Ministre 1 de 2 l'Intérieur 3 Jean-Pierre 4 Chevènement*] démontre ainsi un degré de cohésion plus fort que celui de *Ministre de l'Intérieur Jean-Pierre* et par extension de tous ses sous-groupes de rang 4. Parallèlement, il est peu vraisemblable qu'une unité lexicale vienne renforcer la cohésion globale du pentagram. En effet, dans un contexte immédiat maximum de taille 5[21], il est difficile de prévoir l'occurrence d'une UT spécifique. On stipulera par conséquent que le pentagram [*0 Ministre 1 de 2 l'Intérieur 3 Jean-Pierre 4 Chevènement*] met en évidence une mesure d'association plus forte que celle de tous ses sur-groupes de rang 6. La séquence *Ministre de l'Intérieur Jean-Pierre Chevènement* est donc identifiée par le GenLocalMaxs comme étant une association textuelle obtenue par composition.

## 5.4 Conclusion

L'algorithme GenLocalMaxs propose donc une solution robuste et flexible au problème de la sélection des associations textuelles. D'une part, comme il ne dépend pas de valeurs seuil définies de façon *ad hoc* par expérimentation, il propose une plateforme unique pour l'évaluation de différentes mesures d'association. D'autre part, il permet l'identification directe d'associations textuelles obtenues par composition grâce à l'analyse locale des variations des mesures d'association.

Finalement, combiné à l'Expectative Mutuelle, le GenLocalMaxs met en avant une solution innovatrice dans le cadre de l'extraction d'associations textuelles à partir de corpora. En effet, puisque l'Expectative Mutuelle n'est pas sensible aux particules fréquentes des textes — contrairement à un grand nombre de mesures statistiques proposées dans la littérature, son association avec le GenLocalMaxs permet d'extraire des unités complexes pertinentes sans recourir à la définition de listes de mots vides[22] et sans dépendre de la taille, la langue, le genre ou le domaine des énoncés considérés.

[20] Il est difficile d'attendre un autre nom que celui de *Chevènement* dans le contexte des textes de l'Union Européenne datés de 1996 à 1999.

[21] Plusieurs travaux en linguistique définissent un environnement de 5 UTs comme l'environnement maximum pour la recherche d'associations lexicales.

[22] L'ensemble des mots vides correspond généralement aux particules fonctionnelles du langage.

Avant d'exposer les résultats obtenus à partir de cette nouvelle combinaison, nous introduisons, en guise de préparation à l'évaluation, la normalisation de cinq mesures d'association couramment utilisées dans le cadre de l'extraction de relations textuelles. Ces mesures sont l'information mutuelle [27], le coefficient Dice [45], la probabilité conditionnelle symétrique [30], le test $\Phi^2$ [29] et le coefficient de vraisemblance LogLike [28]. En effet, puisqu'elles ne sont définies que pour les modèles digram ou bien pour les modèles $N$-gram classiques comme c'est le cas pour la probabilité conditionnelle symétrique, ces mesures ne permettent pas d'évaluer les forces d'attraction qui lient entre eux les constituants d'un $N$-gram positionnel pour tout $N \geq 2$. Par conséquent, une phase préliminaire de normalisation est nécessaire. Nous l'introduisons dans le prochain chapitre ■

# Chapitre 6

# Normalisation des Mesures Binaires

Un nombre important de mesures d'association ont été définies dans le but de mesurer le degré de cohésion existant entre différentes UTs. Cependant, la plupart d'entre elles ne sont définies que pour les associations binaires et doivent recourir aux méthodes d'amorçage pour l'acquisition d'associations de taille supérieure à deux UTs. Dans le but de contrecarrer cette méthodologie, plusieurs auteurs ont proposé la définition de mesures d'association $N$-aires. Malheureusement, ces mesures se sont révélées peu discriminantes et la communauté scientifique[1] s'est rattachée aux mesures binaires théoriquement mieux fondées et plus performantes[2]. A l'instar de ce courant idéologique, nous pensons que la définition de mesures d'association génériques — c'est-à-dire $\forall N, N \geq 2$ — est une notion à approfondir plutôt qu'à exclure. Dans ce cadre, nous proposons de construire un lien entre les deux approches binaire et $N$-aire. Ainsi, nous prétendons définir une méthodologie "universelle" pour le calcul des forces d'attraction entre toutes les UTs d'un $N$-gram positionnel générique. Dans ce sens, nous introduisons deux normalisations qui nous permettent de définir pour le cas d'un $N$-gram positionnel générique, cinq mesures d'association binaires couramment utilisées dans les applications du traitement automatique des langues : le coefficient d'association [27], le coefficient Dice [45], la probabilité conditionnelle symétrique[3] [30], le test $\Phi^2$ [29] et le coefficient de vraisemblance LogLike [28].

[1] En grande partie anglo-saxonne.

[2] Cette affirmation reste encore à prouver. En effet, à notre connaissance, aucune évaluation n'a été faite dans ce sens. Nous essaierons dans ce rapport de proposer un début de réponse.

[3] Nous considérons la version binaire de cette mesure.

## 6.1 Evénement Moyen Unique

L'analyse des différentes mesures d'association binaires nous a permis de définir deux groupes distincts : d'un côté, les mesures statistiques basées sur le test d'hypothèse et de l'autre, les mesures de la théorie des probabilités et de la théorie de l'information. En particulier, nous verrons que pour chacun des deux groupes, différentes normalisations devront être proposées. Dans un premier temps, nous abordons la normalisation du second groupe de mesures par le biais de la définition de l'événement moyen unique. L'événement moyen unique propose une solution originale au problème de la division des $N$-grams positionnels en sous-groupes d'UTs et permet ainsi la définition de mesures d'association $N$-aires à partir de mesures définies pour deux UTs. Afin d'illustrer nos propos, nous analysons cette situation de normalisation à partir d'un exemple.

### 6.1.1 Exemple

Considérons le coefficient d'association proposé par K. Church et P. Hanks [27]. Il est défini pour un digram positionnel d'UTs noté $[p_{11}u_1p_{12}u_2]$ dans l'équation 6.1 où $P([p_{11}u_1p_{12}u_2])$ correspond à la probabilité d'occurrence conjointe des deux UTs $u_1$ et $u_2$ selon les positions $p_{11}$ et $p_{12}$. Parallèlement, $P([p_{21}u_1])$ et $P([p_{12}u_2])$ correspondent aux probabilités marginales d'occurrence des formes $u_1$ et $u_2$.

$$I([p_{11}u_1p_{12}u_2]) \equiv \log_2 \frac{P([p_{11}u_1p_{12}u_2])}{P([p_{21}u_1]).P([p_{12}u_2])} \tag{6.1}$$

Conformément à sa définition, cette formule a pour objectif de calculer les forces d'attraction qui existent entre les formes $u_1$ et $u_2$ à partir de l'appréciation de la quantité d'informations que chacune des unités contient par rapport à l'autre.

| Groupe 1 | Groupe 2 |
|---|---|
| $[p_{21}u_1]$ | $[p_{22}u_2p_{23}u_3]$ |
| $[p_{12}u_2]$ | $[p_{11}u_1p_{13}u_3]$ |
| $[p_{13}u_3]$ | $[p_{11}u_1p_{12}u_2]$ |

TAB. 6.1 – Division d'un 3-gram en 2 sous-groupes complémentaires

A partir de cette analyse, la formule du coefficient d'association permet de mesurer la

quantité d'informations liant deux groupes de données. Dans ces conditions, supposons que l'on veuille définir le coefficient d'association pour le trigram $[p_{11}u_1p_{12}u_2p_{13}u_3]$. Il faudrait définir deux groupes d'UTs entre lesquelles il serait souhaitable de mesurer les attirances. Cependant, il existe exactement trois paires de sous-groupes complémentaires possibles comme nous l'illustrons dans le tableau 6.1. De ce fait, la question suivante s'impose : comment peut-on évaluer trois cohésions différentes en une seule mesure ? En effet, à chacune des lignes du tableau 6.1 correspond un coefficient d'association possible comme nous le montrent les trois équations suivantes.

$$I([p_{11}u_1p_{12}u_2p_{13}u_3]) \equiv \log_2 \frac{P([p_{11}u_1p_{12}u_2p_{13}u_3])}{P([p_{21}u_1]).P([p_{22}u_2p_{23}u_3])}$$

$$I([p_{11}u_1p_{12}u_2p_{13}u_3]) \equiv \log_2 \frac{P([p_{11}u_1p_{12}u_2p_{13}u_3])}{P([p_{12}u_2]).P([p_{11}u_1p_{13}u_3])}$$

$$I([p_{11}u_1p_{12}u_2p_{13}u_3]) \equiv \log_2 \frac{P([p_{11}u_1p_{12}u_2p_{13}u_3])}{P([p_{13}u_3]).P([p_{11}u_1p_{12}u_2])}$$

Ainsi, pour un même trigram positionnel, nous pouvons définir trois mesures d'association différentes. Face à cette situation, l'objectif de la normalisation est de définir une et une seule mesure capable de déterminer le degré de cohésion implicite à un $N$-gram positionnel quelconque. Dans ce cadre, deux solutions sont possibles. La première consiste à choisir, selon une heuristique définie, deux sous-groupes complémentaires particuliers parmi toutes les combinaisons possibles. Seulement, dans ce cas, le jugement que l'on porte sur la notion de degré d'attraction est biaisé par cette intervention extérieure. En effet, seule une partie des forces en présence est mesurée. L'autre solution préconise au contraire la prise en compte de toutes les combinaisons possibles de deux sous-groupes complémentaires d'UTs. Ainsi, toutes les attirances entre UTs à l'intérieur d'un $N$-gram positionnel générique seront jugées. C'est cette dernière méthode que l'on adoptera grâce à l'introduction de la notion d'événement moyen unique.

A ce stade de notre explication, il est important que le lecteur ait toujours à l'esprit le fait que notre objectif est de définir une et une seule formule qui prenne en compte le calcul du degré de cohésion d'un $N$-gram positionnel générique. Dans ces conditions, l'analyse des trois coefficients d'association montre que l'élément variable introduit par la

généralisation n'est autre que le point de séparation entre deux sous-groupes d'UTs. En effet, dans le premier cas — première ligne du tableau 6.1 — la séparation entre les deux sous-groupes est réalisée entre $u_1$ et le reste du $N$-gram positonnel. Dans le deuxième cas, la rupture s'effectue entre $u_2$ et le reste du $N$-gram de même que pour $u_3$. L'idée de base de l'événement moyen unique est par conséquent d'évaluer le **point moyen de séparation** entre deux sous-groupes complémentaires d'UTs. Or, cette séparation d'un $N$-gram positionnel en deux sous-groupes complémentaires est uniquement mise en évidence par le dénominateur du coefficient d'association. Dans le cadre de notre exemple, l'événement moyen unique — EMU — du trigram $[p_{11}u_1p_{12}u_2p_{13}u_3]$ serait la moyenne arithmétique des trois dénominateurs supposés par les trois coefficients d'association précédemment définis.

$$EMU([p_{11}u_1p_{12}u_2p_{13}u_3]) = \frac{1}{3} \times \begin{pmatrix} P([p_{21}u_1]).P([p_{22}u_2p_{23}u_3])+ \\ P([p_{12}u_2]).P([p_{11}u_1p_{13}u_3])+ \\ P([p_{13}u_3]).P([p_{11}u_1p_{12}u_2]) \end{pmatrix} \quad (6.2)$$

En effet, d'un coefficient d'association à l'autre, le numérateur reste inchangé et seul le dénominateur varie démontrant ainsi la variabilité du point de séparation. Par conséquent, l'événement moyen unique représente le dénominateur moyen du coefficient d'association. Ainsi, le coefficient d'association du trigram pourrait être formulé par l'équation 6.3.

$$I([p_{11}u_1p_{12}u_2p_{13}u_3]) \equiv \log_2 \frac{P([p_{11}u_1p_{12}u_2p_{13}u_3])}{EMU([p_{11}u_1p_{12}u_2p_{13}u_3])} \quad (6.3)$$

Avant de passer à la formalisation de l'événement moyen unique, nous mettons en évidence une autre préoccupation qu'il est nécessaire de prendre en compte lors de son calcul quand le $N$-gram positionnel considéré dépasse trois UTs. Là encore, nous utiliserons un exemple pour illustrer nos propos. Considérons le cas des modèles tetragram positionnels. Il existe deux possibilités de séparer un tetragram positionnel en deux sous-groupes complémentaires : deux sous-groupes de rang 2 ou bien un sous-groupe de rang 1 et un sous-groupe de rang 3. Dans ces conditions, nous devrons porter une attention toute particulière au calcul de toutes les combinaisons possibles de sous-groupes complémentaires. Supposons le tetragram suivant [*0 Président 1 de 2 la 3 République*]. Il est possible de le diviser en deux sous-groupes complémentaires de rang 2 comme le montre le tableau 6.2[4]. Cependant, il

[4]Nous rappelons que la propriété de changement de position de référence nous permet de formuler les

est également possible de séparer le tetragram en deux sous-groupes complémentaires, l'un de rang trois et l'autre de rang 1 comme nous l'illustrons dans le tableau 6.3.

| Groupe 1 | Groupe 2 |
|---|---|
| [*-2 Président -1 de*] | [*0 la 1 République*] |
| [*-1 Président 1 la*] | [*0 de 2 République*] |
| [*-1 Président 2 République*] | [*0 de 1 la*] |

TAB. 6.2 – Division d'un 4-gram en 2 sous-groupes de rang 2

Ainsi, afin de définir une normalisation complète, toutes les combinaisons de sous-groupes complémentaires qu'il est possible de calculer à partir d'un quelconque $N$-gram positionnel devront être contemplées dans le calcul de l'événement moyen unique. Ainsi, nous introduisons la formalisation du concept d'événement moyen unique qui sert à généraliser les mesures d'association binaires de la théorie des probabilités et de la théorie de l'information pour un $N$-gram positionnel générique.

| Groupe 1 | Groupe 2 |
|---|---|
| [*-1 Président*] | [*0 de 1 la 2 République*] |
| [*1 de*] | [*0 Président 2 la 3 République*] |
| [*2 la*] | [*0 Président 1 de 3 République*] |
| [*3 République*] | [*0 Président 1 de 2 la*] |

TAB. 6.3 – Division d'un 4-gram en 2 sous-groupes de rang 1 et 3

### 6.1.2 Formalisation

Dans cette section, nous nous proposons de formaliser la notion d'événement moyen unique pour un ensemble de mesures d'association binaires de la théorie des probabilités et de la théorie de l'information. Celles-ci sont le coefficient d'association [27], le coefficient Dice [45] et la probabilité conditionnelle symétrique [30]. Dans ce cadre, nous rappelons leurs définitions "fréquentistes" afin de présenter leurs similitudes. En particulier, nous notons $Dice([p_{11}u_1p_{12}u_2])$ le coefficient Dice pour deux UTs $u_1$ et $u_2$, $n$ le nombre d'UTs

digrams du groupe 1 en fonction de l'UT pivot ayant sa position équivalente à zéro. Par exemple, le digram [*-2 Président -1 de*] peut s'exprimer de la forme [*0 Président 1 de*].

de l'énoncé considéré et $k$ la fonction de fréquence d'occurrence d'un $N$-gram positionnel quelconque.

$$I([p_{11}u_1p_{12}u_2]) = \log_2 \frac{n.k([p_{11}u_1p_{12}u_2])}{k([p_{21}u_1]).k([p_{12}u_2])} \tag{6.4}$$

$$Dice([p_{11}u_1p_{12}u_2]) = \frac{2.k([p_{11}u_1p_{12}u_2])}{k([p_{21}u_1]) + k([p_{12}u_2])} \tag{6.5}$$

$$SCP([p_{11}u_1p_{12}u_2]) = \frac{k([p_{11}u_1p_{12}u_2])^2}{k([p_{21}u_1]).k([p_{12}u_2])} \tag{6.6}$$

A la lumière de ces trois formules, il est clair que chacune des mesures d'association démontre un comportement similaire. En effet, elles définissent toutes, dans leur dénominateur, la division d'un digram en deux sous-groupes complémentaires. Par conséquent, leur généralisation impliquera la division d'un $N$-gram positionnel générique en deux sous-groupes complémentaires d'UTs et donc la définition d'un événement moyen unique. Ainsi, les trois mesures d'association se verront attribuer au dénominateur le même événement moyen unique — au changement d'opérateur près.

### Notion d'Evénement

Dans ce paragraphe, nous introduisons une notion de vocabulaire qui facilitera notre argumentation : l'événement[5]. Nous considérons un événement comme étant un dénominateur particulier de l'une des trois mesures d'association pour une combinaison donnée de deux sous-groupes complémentaires d'un $N$-gram positionnel. Ainsi, dans le cadre des modèles digram positionnels, deux événements distincts peuvent être définis selon que l'on analyse le coefficient Dice — opérateur = addition — ou le coefficient d'association et la probabilité conditionnelle symétrique — opérateur = multiplication. Ces deux événements sont formulés de la forme suivante.

$$k([p_{21}u_1]) + k([p_{12}u_2])$$

$$k([p_{21}u_1]).k([p_{12}u_2])$$

[5]Nous tenons à noter que ce terme n'est en rien identique au concept d'événement défini dans le cadre de la théorie des probabilités.

Afin de généraliser cette notion, nous introduisons l'opérateur $\otimes$ qui représente soit l'addition soit la multiplication. Ainsi, nous pouvons définir un événement comme une opération $\otimes$ entre deux sous-groupes complémentaires d'un $N$-gram positionnel générique. Par conséquent, le seul événement à considérer pour un digram positionnel générique serait le suivant.

$$k([p_{21}u_1]) \otimes k([p_{12}u_2])$$

Comme nous l'avons précédemment formulé, l'événement moyen unique d'un $N$-gram positionnel peut être considéré comme étant la moyenne de tous ses événements. Dans ces conditions, il est nécessaire de calculer le nombre d'événements qu'implique un $N$-gram positionnel générique ainsi que la somme de tous ceux-ci. Nous nous attacherons à définir le nombre d'événements impliqués par un $N$-gram positionnel dans le paragraphe suivant.

**Nombre d'Evénements**

L'idée de base de l'événement moyen unique est d'évaluer la valeur moyenne des différents dénominateurs mis en évidence par la division d'un $N$-gram positionnel quelconque. Dans ce cadre, nous définissons la notion de point de séparation moyen — $PSM$ — qui peut être vu comme une frontière symbolique qui divise un $N$-gram positionnel en deux sous-groupes complémentaires. Ainsi, pour un $N$-gram positionnel donné, le $PSM$ peut prendre un ensemble de valeurs comprises entre 1 et $E(N/2)$ comme le montre le tableau 6.4 où $E$ est la fonction qui retourne la partie entière de son argument.

| $PSM$ | Nb. d'UTs $1^{er}$ sous-groupe | Nb. d'UTs $2^{eme}$ sous-groupe |
|---|---|---|
| 1 | 1 | $N-1$ |
| 2 | 2 | $N-2$ |
| ... | ... | ... |
| $E(N/2)$ | $E(N/2)$ | $N-E(N/2)$ |

TAB. 6.4 – Division d'un $N$-gram positionnel

Dans tous les cas, le tableau précédent ne met en évidence qu'une seule partie des événements possibles. En effet, pour chaque valeur du $PSM$, il existe plusieurs combinaisons de sous-groupes complémentaires. Par exemple, dans le cas spécifique où le $PSM$ vaut 1, il existe $N$ couples de sous-groupes complémentaires. De fait, il est facile de déterminer qu'il existe $N$ sous-groupes d'une UT et parallèlement $N$ sous-groupes de $N-1$ UTs. Cette situation est illustrée dans le tableau suivant 6.5.

| | $1^{er}$ sous-groupe | $2^{eme}$ sous-groupe |
|---|---|---|
| 1 | $[p_{21}u_1]$ | $[p_{22}u_2...p_{2i}u_i...p_{2N}u_N]$ |
| 2 | $[p_{12}u_2]$ | $[p_{11}u_1p_{13}u_3...p_{1i}u_i...p_{1N}u_N]$ |
| ... | ... | ... |
| $N$ | $[p_{1N}u_N]$ | $[p_{11}u_1p_{12}u_2...p_{1i}u_i...p_{1(N-1)}u_{(N-1)}]$ |

TAB. 6.5 – Sous-groupes complémentaires pour $PSM$=1

Ainsi, pour le seul cas du $PSM$=1, il existe $N$ événements possibles comme le montre le tableau suivant 6.6.

| Nb. | Evénement |
|---|---|
| 1 | $k([p_{21}u_1]) \otimes k([p_{22}u_2...p_{2i}u_i...p_{2N}u_N])$ |
| 2 | $k([p_{12}u_2]) \otimes k([p_{11}u_1p_{13}u_3...p_{1i}u_i...p_{1N}u_N])$ |
| ... | ... |
| $N$ | $k([p_{1N}u_N]) \otimes k\left(\left[p_{11}u_1p_{12}u_2...p_{1i}u_i...p_{1(N-1)}u_{(N-1)}\right]\right)$ |

TAB. 6.6 – Evénements pour $PSM$=1

Dans le but de calculer le nombre total d'événements d'un $N$-gram positionnel, nous devons donc distinguer deux sous-cas : le cas où $N$ est un nombre pair[6] et le cas où $N$ est un nombre impair[7].

Ainsi, si le $N$-gram positionnel contient un nombre pair d'UTs, le nombre total d'événements doit être calculé en deux étapes. Premièrement, nous calculons le nombre de

[6]Le $N$-gram positionnel contient un nombre pair d'UTs.

[7]Le $N$-gram positionnel contient un nombre impair d'UTs.

| Nb. | Evénement |
|---|---|
| 1 | $k([p_{31}u_1p_{32}u_2]) \otimes k([p_{33}u_3p_{34}u_4])$ |
| 2 | $k([p_{21}u_1p_{23}u_3]) \otimes k([p_{22}u_2p_{24}u_4])$ |
| 3 | $k([p_{21}u_1p_{24}u_4]) \otimes k([p_{22}u_2p_{23}u_3])$ |

TAB. 6.7 – Tetragram positionnel et $PSM$=2

combinaisons de $PSM$ UTs parmi $N$ pour tout $PSM = 1..E(N/2) - 1$ c'est-à-dire pour toutes les valeurs du $PSM$ à l'exception de la dernière. En effet, dans le cas spécifique du dernier point de séparation — $PSM = E(N/2)$, le $N$-gram positionnel est divisé en sous-groupes complémentaires de même taille. Dans ces conditions, le nombre d'événements doit être réduit de moitié. Ainsi, au nombre de combinaisons déjà calculées, nous devons ajouter la moitié des combinaisons de $E(N/2)$ UTs parmi $N$ UTs. Par exemple, pour le cas d'un tetragram positionnel, seulement trois événements peuvent être calculés si $PSM$=2 comme l'illustre le tableau 6.7. D'autre part, si le $N$-gram positionnel considéré contient un nombre impair d'UTs, le nombre total d'événements est le nombre de combinaisons de $PSM$ UTs parmi $N$ pour tout $PSM = 1..E(N/2)$. Ainsi, la fonction qui calcule le nombre total d'événements d'un $N$-gram positionnel est notée $nb\_ev(.)$ et définie dans l'équation suivante.

$$nb\_ev([p_{11}u_1...p_{1N}u_N]) =$$

$$\begin{cases} pair(N), & \begin{cases} N = 2, 1 \\ N > 2, \begin{cases} \sum_{PSM=1}^{E(N/2)-1} \mathrm{C}_N^{PSM} + \\ \frac{1}{2}\mathrm{C}_N^{E(N/2)} \end{cases} \end{cases} \\ impair(N), & \sum_{PSM=1}^{E(N/2)} \mathrm{C}_N^{PSM} \end{cases} \tag{6.7}$$

Nous devons maintenant formuler la somme proprement dite de tous les événements impliqués par un $N$-gram positionnel afin de formaliser la définition de l'événement moyen unique. C'est ce que nous nous proposons d'étudier dans le paragraphe suivant.

**Somme d'Evénements**

Une fois déterminé le nombre d'événements impliqués dans un $N$-gram positionnel, il est nécessaire de calculer leur somme. Dans ce contexte, nous distinguons deux calculs principaux. Premièrement, nous définissons une fonction "outil" notée $ev\_spec(.,.)$ qui nous permet de calculer la somme des événements possibles pour une valeur donnée du $PSM$. Deuxièmement, à partir de la définition de cette fonction, nous calculons la somme de tous les événements pour toutes les valeurs du $PSM$ grâce à la fonction $som\_ev(.)$.

$$ev\_spec\left(PSM, [p_{11}u_1 ... p_{1i}u_i ... p_{1N}u_N]\right) =$$

$$\sum_{i1=2}^{j} \sum_{i2=i1+1}^{j+1} \cdots \sum_{iPSM=i(PSM-1)+1}^{N} \left( \begin{array}{l} k\left([p_{i1i1}u_{i1}p_{i1i2}u_{i2} ... p_{i1iPSM}u_{iPSM}]\right) \\ \otimes \\ k\left([p_{11}u_1 ... p_{\hat{1}i1}u_{\hat{i1}} ... p_{\hat{1}ik}u_{\hat{ik}} ... p_{1N}u_N]\right) \end{array} \right)$$

où

$$j = N - PSM + 1 \tag{6.8}$$

Nous définissons donc dans un premier temps la fonction $ev\_spec(.,.)$. Elle est formulée dans l'équation précédente 6.8 pour une valeur du $PSM$ et un $N$-gram positionnel générique donnés. Nous rappelons que l'accent circonflexe est une notation empruntée à l'Algèbre qui suppose l'omission du terme marqué par l'accent dans une suite indexée de 1 à $N$. A partir de la fonction $ev\_spec(.,.)$ qui permet de définir la somme de tous les événements possibles pour une valeur donnée du $PSM$, nous devons maintenant prendre en compte les variations du point de séparation moyen. Dans ces conditions, nous distinguons le cas où $N$ est pair et le cas où $N$ est impair. Ainsi, on définit la fonction $som\_ev(.)$ pour un $N$-gram positionnel de la forme suivante dans l'équation 6.9.

$$som\_ev\left([p_{11}u_1...p_{1i}u_i...p_{1N}u_N]\right) = \tag{6.9}$$

$$\begin{cases} N = 2, ev\_spec(1, [p_{11}u_1...p_{1i}u_i...p_{1N}u_N]) \\ pair(N) \wedge (N > 2), \begin{cases} \sum_{PSM=1}^{E(N/2)-1} \\ \left( \begin{array}{l} ev\_spec(PSM, [p_{11}u_1...p_{1i}u_i...p_{1N}u_N]) + \\ ev\_spec(N - PSM, [p_{11}u_1...p_{1i}u_i...p_{1N}u_N]) \end{array} \right) \\ + ev\_spec(E\left(N/2\right), [p_{11}u_1...p_{1i}u_i...p_{1N}u_N]) \end{cases} \\ impair(N), \sum_{PSM=1}^{E(N/2)} \left( \begin{array}{l} ev\_spec(PSM, [p_{11}u_1...p_{1i}u_i...p_{1N}u_N]) + \\ ev\_spec(N - PSM, [p_{11}u_1...p_{1i}u_i...p_{1N}u_N]) \end{array} \right) \end{cases}$$

Finalement, nous formalisons, dans le prochain paragraphe, la notion d'événement moyen unique à partir de la définition des fonctions $nb\_ev(.)$ et $som\_ev(.)$.

**Evénement Moyen Unique**

L'événement moyen unique — EMU — n'est autre que la moyenne arithmétique de tous les événements qu'il est possible de construire à partir d'un $N$-gram positionnel donné. Par conséquent, il peut être facilement calculé à partir des deux fonctions définies antérieurement : $nb\_ev(.)$ et $som\_ev(.)$. En effet, l'EMU n'est autre que le quotient entre la somme des événements et le nombre d'événements considérés. Ainsi, l'EMU est défini dans l'équation suivante comme étant la valeur d'un événement moyen.

$$EMU([p_{11}u_1...p_{1i}u_i...p_{1N}u_N]) = \frac{som\_ev([p_{11}u_1...p_{1i}u_i...p_{1N}u_N])}{nb\_ev([p_{11}u_1...p_{1i}u_i...p_{1N}u_N])} \tag{6.10}$$

Grâce à cette définition, nous pouvons donc proposer une formule générique pour les mesures d'association binaires que sont le coefficient d'association, le coefficient Dice et la probabilité conditionnelle symétrique.

### 6.1.3 Mesures d'Association Binaires Normalisées

Afin de formuler les mesures d'association binaires de la théorie des probabilités et de la théorie de l'information pour le cas des modèles $N$-gram positionnels, nous proposons d'introduire la notion d'événement moyen unique dans la formule définie pour deux UTs. Ainsi, le coefficient d'association, le coefficient Dice et la probabilité conditionnelle symétrique peuvent être définis génériquement à partir de l'événement moyen unique de la forme suivante. Nous rappelons que l'opérateur $\otimes$ devra être remplacé par l'addition dans le cas du coefficient Dice et par la multiplication dans le cas du coefficient d'association et de la probabilité conditionnelle symétrique.

$$I([p_{11}u_1...p_{1i}u_i...p_{1N}u_N]) = \log_2 \frac{n.k([p_{11}u_1...p_{1i}u_i...p_{1N}u_N])}{EMU([p_{11}u_1...p_{1i}u_i...p_{1N}u_N])} \quad (6.11)$$

$$Dice([p_{11}u_1...p_{1i}u_i...p_{1N}u_N]) = \frac{2.k([p_{11}u_1...p_{1i}u_i...p_{1N}u_N])}{EMU([p_{11}u_1...p_{1i}u_i...p_{1N}u_N])} \quad (6.12)$$

$$SCP([p_{11}u_1...p_{1i}u_i...p_{1N}u_N]) = \frac{k([p_{11}u_1...p_{1i}u_i...p_{1N}u_N])^2}{EMU([p_{11}u_1...p_{1i}u_i...p_{1N}u_N])} \quad (6.13)$$

Grâce à cette normalisation, il est maintenant possible de déterminer le degré de cohésion existant entre $N$ UTs à partir de mesures d'association binaires bien connues. Malheureusement, la normalisation proposée pour ces trois mesures ne peut pas être appliquée aux mesures d'association binaires basées sur le test d'hypothèse. Nous proposons donc une nouvelle normalisation pour ce type de mesures.

## 6.2 Evénements Moyens Gauche et Droit

Dans le cadre des mesures d'association basées sur l'analyse des tableaux de contingence, nous proposons une nouvelle normalisation mettant en évidence le calcul de deux événements moyens : l'un gauche et l'autre droit. Comme nous l'avons vu précédemment, les tableaux de contingence se distinguent par la définition d'un critère de ligne — critère gauche — et d'un critère de colonne — critère droit — selon lesquels les individus à analyser sont classés. Ainsi, afin de généraliser les mesures d'association statistiques telles que le coefficient de vraisemblance Loglike [28] et le coefficient $\Phi^2$ [29], nous verrons qu'il est nécessaire de mesurer un événement moyen pour chacun des critères considérés. Les deux

événements seront nommés événement moyen gauche et événement moyen droit selon le critère considéré. Dans un premier temps, nous rappelons les principales caractéristiques des tableaux de contingence afin de familiariser le lecteur avec cette notion et d'illustrer les problèmes rencontrés lors de la généralisation.

### 6.2.1 Tableau de Contingence

Un tableau de contingence est une classification multiple. Ainsi, supposons que nous avons à classer $n$ individus suivant deux critères $A$ et $B$ : $A$ distinguant $r$ classes possibles pour un individu — $A_1$, $A_2$, ..., $A_r$ — et $B$ distinguant $s$ classes possibles pour un individu — $B_1$, $B_2$, ..., $B_s$. Si l'on considère $n_{ij}$ le nombre d'individus appartenant à $A_i$ et $B_j$, le tableau de contingence peut être défini comme un tableau de dimensions $r \times s$ tel que $\sum_{i=1}^{r} \sum_{j=1}^{s} n_{ij} = n$.

| | $B_1$ | $B_2$ | ... | $B_s$ |
|---|---|---|---|---|
| $A_1$ | $n_{11}$ | $n_{12}$ | ... | $n_{1s}$ |
| $A_2$ | $n_{21}$ | $n_{22}$ | ... | $n_{2s}$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| $A_r$ | $n_{r1}$ | $n_{r2}$ | ... | $n_{rs}$ |

TAB. 6.8 – Tableau de contingence générique

Dans le contexte des cooccurrences d'UTs, les mesures d'association qui ont été proposées par les chercheurs, sont définies pour des tableaux de contingence de dimensions $2 \times 2$. En effet, l'idée mise en évidence est de classer tous les digrams d'UTs selon deux critères, ces derniers distinguant deux classes possibles pour chaque individu — digram. Ces critères correspondent à deux UTs données selon lesquelles les digrams positionnels doivent être classés. Ainsi, on peut définir autant de tableaux de contingence qu'il existe de digrams positionnels pour une distribution donnée — c'est-à-dire approximativement la taille de l'énoncé. Un tableau de contingence pour les modèles 2-gram positionnels peut facilement être défini — voir Tableau 6.9 — pour deux UTs $u_1$ et $u_2$, et leurs positions correspondantes $p_{11}$ et $p_{12}$. On considèrera que l'UT $u_1$ est le critère gauche du tableau de contingence et $u_2$ son critère droit.

A partir de ce tableau de contingence, il est possible de classer tous les digrams d'UTs

| | $[p_{12}u_2]$ | $[p_{12}\neg u_2]$ | Total |
|---|---|---|---|
| $[p_{11}u_1]$ | $k([p_{11}u_1p_{12}u_2])$ | $k([p_{11}u_1p_{12}\neg u_2])$ | $k([p_{11}u_1])$ |
| $[p_{11}\neg u_1]$ | $k([p_{11}\neg u_1p_{12}u_2])$ | $k([p_{11}\neg u_1p_{12}\neg u_2])$ | $k([p_{11}\neg u_1])$ |
| Total | $k([p_{12}u_2])$ | $k([p_{12}\neg u_2])$ | $n$ |

TAB. 6.9 – Tableau de contingence $2 \times 2$

d'un énoncé selon les deux critères $u_1$ et $u_2$ pour un ensemble de positions $p_{11}$ et $p_{12}$ données — une distribution donnée[8]. Ainsi, l'étude de l'hypothèse d'indépendance peut être évaluée sans difficulté à partir des différentes méthodologies proposées par la théorie statistique comme le test du $\chi^2$ ou le rapport de vraisemblance $\lambda$.

Cependant, que se passe-t-il si l'on veut tester l'hypothèse nulle d'indépendance selon plus de deux critères ? Cette question doit être posée si l'on prétend caractériser le degré d'attraction entre $N$ UTs. Par exemple, comment peut-on analyser la cooccurrence de trois UTs — trigram positionnel — à partir d'un tableau de contingence ? Deux solutions sont possibles : l'une théorique et l'autre empirique. La première solution, théorique, est définie par un tableau de contingence à trois entrées. Parallèlement à ce qui a été développé pour le cas de deux critères, il est possible de classer une population selon trois critères $A_i(i = 1, 2, ..., s_1)$, $B_j(j = 1, 2, ..., s_2)$ et $C_k(k = 1, 2, ..., s_3)$ définissant un tableau de contingence à trois entrées de dimensions $s_1 \times s_2 \times s_3$. Le nombre d'éléments dans chacune des cases est alors noté $n_{ijk}$ tel que $\sum_{i=1}^{s_1} \sum_{j=1}^{s_2} \sum_{k=1}^{s_3} n_{ijk} = n$. Dans ce cadre, il est souvent commode de se représenter les cases comme des cubes contenus dans un parallélépipède de largeur $s_1$, de longueur $s_2$ et de hauteur $s_3$.

$$H_0 \quad : \quad p_{ijk} = p_i.p_j.p_k$$

Ainsi, à partir de ce tableau de contingence, il est possible de tester quatre hypothèses. On peut tester si les trois critères sont mutuellement indépendants, auquel cas l'hypothèse nulle $H_0$ peut être formulée comme précédemment où $p$ est une fonction de probabilité. On peut également tester si l'un des critères est indépendant des deux autres. Ainsi, pour

[8] On remarquera qu'il existe $n = T - F$ — $T$ étant la taille de l'énoncé et $F$ l'environnement immédiat considéré — digrams positionnels pour une distribution donnée.

tester si la classification $B$ est indépendante de $A$ et $C$, l'hypothèse $H_0$ serait définie de la forme suivante.

$$H_0 \quad : \quad p_{ijk} = p_{ik}.p_j$$

Selon cette définition, l'extension aux modèles 3-gram positionnels est directe. Afin de calculer le degré d'attraction entre 3 UTs, il suffirait de classer l'ensemble des trigrams positionnels — construits pour trois positions données — par rapport à trois critères qui ne seraient autres que trois UTs données[9]. Cependant, un problème se pose pour le calcul du degré d'attraction : quelle hypothèse choisir pour le test d'indépendance ? En effet, chacune des quatre hypothèses devrait logiquement être testée. Pour un trigram positionnel donné $[p_{11}u_1p_{12}u_2p_{13}u_3]$, il serait intéressant de mesurer la validité du test d'indépendance entre les trois UTs $u_1$, $u_2$ et $u_3$, mais aussi entre $u_1$ et le couple $u_2$, $u_3$. Pareillement, on devrait tester l'hypothèse nulle entre $u_2$ (resp. $u_3$) et le couple $u_1$, $u_3$ (resp. $u_1$, $u_2$). Malheureusement, la théorie statistique ne propose aucune solution pour le calcul commun de l'ensemble des hypothèses possibles. Ainsi, nous proposons une solution empirique qui nous permet de résoudre ce problème.

Plutôt que de définir des tableaux de contingence de dimensions complexes, nous proposons d'adapter la situation des modèles $N$-gram positionnels aux tableaux de contingence de dimensions $2 \times 2$. En effet, la structure des tableaux de contingence de dimensions $2 \times 2$ suppose la définition de deux critères et par conséquent la division de chaque $N$-gram positionnel en deux sous-groupes complémentaires. Par exemple, dans le cas des trigrams positionnels, il serait possible de définir trois tableaux de contingence conformément aux six critères qu'il est possible de définir. Ainsi, le critère gauche de chaque tableau serait représenté par l'une des trois UTs et le critère droit par le couple d'UTs complémentaires. Nous illustrons cette situation dans les trois tableaux précédents où, pour simplifier la présentation des données, nous proposons les notations suivantes.

Afin de définir un seul tableau de contingence — tableau de contingence moyen — regroupant l'ensemble des trois tableaux précédents, nous proposons l'introduction des notions

[9] Le processus peut être directement copié de ce que nous avons démontré précédemment pour le cas des modèles digram positionnels.

| 1 | $[p_{12}u_2p_{13}u_3]$ | $[p_{12}\neg u_2p_{13}\neg u_3]$ | Total |
|---|---|---|---|
| $[p_{11}u_1]$ | $k\,(W_1)$ | $k\,(W_2)$ | $k\,([p_{11}u_1])$ |
| $[p_{11}\neg u_1]$ | $k\,(W_3)$ | $k\,(W_4)$ | $k\,([p_{11}\neg u_1])$ |
| Total | $k\,([p_{12}u_2p_{13}u_3])$ | $k\,([p_{12}\neg u_2p_{13}\neg u_3])$ | $n$ |

| 2 | $[p_{11}u_1p_{13}u_3]$ | $[p_{11}\neg u_1p_{13}\neg u_3]$ | Total |
|---|---|---|---|
| $[p_{12}u_2]$ | $k\,(W_1)$ | $k\,(W_5)$ | $k\,([p_{12}u_2])$ |
| $[p_{12}\neg u_2]$ | $k\,(W_6)$ | $k\,(W_4)$ | $k\,([p_{12}\neg u_2])$ |
| Total | $k\,([p_{11}u_1p_{13}u_3])$ | $k\,([p_{11}\neg u_1p_{13}\neg u_3])$ | $n$ |

| 3 | $[p_{11}u_1p_{12}u_2]$ | $[p_{11}\neg u_1p_{12}\neg u_2]$ | Total |
|---|---|---|---|
| $[p_{13}u_3]$ | $k\,(W_1)$ | $k\,(W_7)$ | $k\,([p_{13}u_3])$ |
| $[p_{13}\neg u_3]$ | $k\,(W_8)$ | $k\,(W_4)$ | $k\,([p_{13}\neg u_3])$ |
| Total | $k\,([p_{11}u_1p_{12}u_2])$ | $k\,([p_{11}\neg u_1p_{12}\neg u_2])$ | $n$ |

TAB. 6.10 – Tableaux 1, 2 et 3

$$W_1 \equiv [p_{11}u_1p_{12}u_2p_{13}u_3] \quad W_2 \equiv [p_{11}u_1p_{12}\neg u_2p_{13}\neg u_3]$$
$$W_3 \equiv [p_{11}\neg u_1p_{12}u_2p_{13}u_3] \quad W_4 \equiv [p_{11}\neg u_1p_{12}\neg u_2p_{13}\neg u_3]$$
$$W_5 \equiv [p_{11}\neg u_1p_{12}u_2p_{13}\neg u_3] \quad W_6 \equiv [p_{11}u_1p_{12}\neg u_2p_{13}u_3]$$
$$W_7 \equiv [p_{11}\neg u_1p_{12}\neg u_2p_{13}u_3] \quad W_8 \equiv [p_{11}u_1p_{12}u_2p_{13}\neg u_3]$$

d'événements moyens gauche et droit — $EMG$ et $EMD$. Nous en donnons une intuition en continuant avec notre exemple. Dans un premier temps, nous définissons la structure du tableau de contingence moyen. Dans ce cadre, chacune des cases du tableau de contingence moyen doit correspondre à la moyenne arithmétique des cases des trois tableaux de contingence. En guise de notation, la case $ct_{ij}$ représentera la case de la ligne $i$ et de la colonne $j$ du tableau de contingence $t$. Dans ces conditions, nous écrirons les équations suivantes pour la définition du tableau de contingence moyen dont les cases seront notées $cm_{ij}$ pour chaque ligne $i$ et colonne $j$.

$$
\begin{aligned}
cm_{11} &= \frac{1}{3}\left(c1_{11} + c2_{11} + c3_{11}\right) \\
&= \frac{1}{3}\begin{pmatrix} k([p_{11}u_1 p_{12}u_2 p_{13}u_3])+ \\ k([p_{11}u_1 p_{12}u_2 p_{13}u_3])+ \\ k([p_{11}u_1 p_{12}u_2 p_{13}u_3]) \end{pmatrix} \\
&= k([p_{11}u_1 p_{12}u_2 p_{13}u_3]) \qquad (6.14)
\end{aligned}
$$

$$
\begin{aligned}
cm_{12} &= \frac{1}{3}\left(c1_{12} + c2_{12} + c3_{12}\right) \\
&= \frac{1}{3}\begin{pmatrix} k([p_{11}u_1 p_{12}\neg u_2 p_{13}\neg u_3])+ \\ k([p_{11}\neg u_1 p_{12}u_2 p_{13}\neg u_3])+ \\ k([p_{11}\neg u_1 p_{12}\neg u_2 p_{13}u_3]) \end{pmatrix} \\
&= \frac{1}{3}\begin{pmatrix} (k([p_{11}u_1]) - k([p_{11}u_1 p_{12}u_2 p_{13}u_3])) + \\ (k([p_{12}u_2]) - k([p_{11}u_1 p_{12}u_2 p_{13}u_3])) + \\ (k([p_{13}u_3]) - k([p_{11}u_1 p_{12}u_2 p_{13}u_3])) \end{pmatrix} \\
&= \begin{array}{l} \frac{1}{3}\left(k([p_{11}u_1]) + k([p_{12}u_2]) + k([p_{13}u_3])\right) - \\ k([p_{11}u_1 p_{12}u_2 p_{13}u_3]) \end{array} \qquad (6.15)
\end{aligned}
$$

$$
\begin{aligned}
cm_{21} &= \frac{1}{3}\left(c1_{21} + c2_{21} + c3_{21}\right) \\
&= \frac{1}{3}\begin{pmatrix} k([p_{11}\neg u_1 p_{12}u_2 p_{13}u_3])+ \\ k([p_{11}u_1 p_{12}\neg u_2 p_{13}u_3])+ \\ k([p_{11}u_1 p_{12}u_2 p_{13}\neg u_3]) \end{pmatrix} \\
&= \frac{1}{3}\begin{pmatrix} (k([p_{12}u_2 p_{13}u_3]) - k([p_{11}u_1 p_{12}u_2 p_{13}u_3])) + \\ (k([p_{11}u_1 p_{13}u_3]) - k([p_{11}u_1 p_{12}u_2 p_{13}u_3])) + \\ (k([p_{11}u_1 p_{12}u_2]) - k([p_{11}u_1 p_{12}u_2 p_{13}u_3])) \end{pmatrix}
\end{aligned}
$$

$$= \frac{1}{3}\begin{pmatrix} k([p_{11}u_1p_{12}u_2])+ \\ k([p_{11}u_1p_{13}u_3])+ \\ k([p_{12}u_2p_{13}u_3]) \end{pmatrix} - k([p_{11}u_1p_{12}u_2p_{13}u_3]) \quad (6.16)$$

$$cm_{22} = \frac{1}{3}(c1_{22} + c2_{22} + c3_{22})$$

$$= \frac{1}{3}\begin{pmatrix} k([p_{11}\neg u_1p_{12}\neg u_2p_{13}\neg u_3])+ \\ k([p_{11}\neg u_1p_{12}\neg u_2p_{13}\neg u_3])+ \\ k([p_{11}\neg u_1p_{12}\neg u_2p_{13}\neg u_3]) \end{pmatrix}$$

$$= \frac{1}{3}\begin{pmatrix} n - k([p_{11}u_1p_{12}u_2p_{13}u_3])+ \\ n - k([p_{11}u_1p_{12}u_2p_{13}u_3])+ \\ n - k([p_{11}u_1p_{12}u_2p_{13}u_3]) \end{pmatrix}$$

$$= n - k([p_{11}u_1p_{12}u_2p_{13}u_3]) \quad (6.17)$$

L'analyse des quatre équations montre qu'il est nécessaire d'introduire la moyenne des occurrences des unigrams et des digrams positionnels qui forment les critères des différents tableaux de contingence. En effet, pour le calcul des cases, seules ces moyennes sont nécessaires et démontrent la division d'un trigram positionnel en deux sous-groupes complémentaires, division qu'il est nécessaire d'évaluer. Dans ce cadre, la moyenne des occurrences des unigrams représentera l'événement moyen gauche et la moyenne des occurrences des digrams, l'événement moyen droit. En effet, le critère gauche regroupe l'ensemble de tous les unigrams contenus dans un $N$-gram positionnel et le critère droit l'ensemble des digrams contenus dans ce même $N$-gram positionnel. Le lecteur comprendra facilement la notion d'événement moyen — gauche et droit — à partir des équations suivantes dans le cas des modèles trigram positionnels.

$$EMG([p_{11}u_1p_{12}u_2p_{13}u_3]) = \frac{1}{3}\begin{pmatrix} k([p_{11}u_1])+ \\ k([p_{12}u_2])+ \\ k([p_{13}u_3]) \end{pmatrix} \quad (6.18)$$

$$EMD([p_{11}u_1p_{12}u_2p_{13}u_3]) = \frac{1}{3}\begin{pmatrix} k([p_{11}u_1p_{12}u_2])+ \\ k([p_{11}u_1p_{13}u_3])+ \\ k([p_{12}u_2p_{13}u_3]) \end{pmatrix} \quad (6.19)$$

La normalisation proposée met donc en évidence la division de tous les trigrams positionnels en deux critères de classification. En effet, chaque trigram positionnel est à l'origine d'un tableau de contingence moyen qui implique la définition de deux événements moyens. Dans ces conditions, il est possible de formuler les tests statistiques basés sur l'étude des tableaux de contingence de dimensions $2 \times 2$ pour trois UTs. Considérons le test $\Phi^2$ défini dans l'équation 6.20 pour deux UTs et deux positions.

$$\Phi^2\left([p_{11}u_1p_{12}u_2]\right) = \frac{(n\times k([p_{11}u_1p_{12}u_2]) - k([p_{11}u_1])\times k([p_{12}u_2]))^2}{k([p_{11}u_1])\times(n-k([p_{11}u_1]))\times k([p_{12}u_2])\times(n-k([p_{12}u_2]))} \tag{6.20}$$

Ce test serait défini de la façon suivante pour un trigram positionnel quelconque $[p_{11}u_1p_{12}u_2p_{13}u_3]$ que l'on note $W$ pour simplifier la lecture.

$$\Phi^2\left(W\right) = \frac{(n\times k(W) - EMG(W)\times EMD(W))^2}{EMG(W)\times(n-EMG(W))\times EMD(W)\times(n-EMD(W))} \tag{6.21}$$

Une fois encore, nous proposons l'étude des modèles tetragram positionnels afin d'évaluer tous les impératifs de la division d'un $N$-gram positionnel en deux critères moyens. En effet, il existe deux possibilités de séparer un tetragram positionnel en deux critères moyens droit et gauche : deux sous-groupes de rang 2 complémentaires ou bien un sous-groupe de rang 1 et un sous-groupe de rang trois complémentaires. Considérons le tetragram positionnel suivant $[p_{11}\ u_1\ p_{12}\ u_2\ p_{13}\ u_3\ p_{14}\ u_4]$. Il est possible de définir deux types de tableaux de contingence pour chacun des cas exposés. Nous illustrons cette situation dans les tableaux suivants 6.11 où, dans le premier cas, la partie gauche du tableau de contingence considéré est représentée par le digram positionnel $[p_{11}\ u_1\ p_{12}\ u_2]$ et où, dans le deuxième cas, le critère gauche est défini par l'UT $u_1$. Nous formulons les notations suivantes pour simplifier la présentation des résultats.

$$\begin{aligned}
W_1 &\equiv [p_{11}u_1p_{12}u_2p_{13}u_3p_{14}u_4] \\
W_2 &\equiv [p_{11}u_1p_{12}\neg u_2p_{13}\neg u_3p_{14}\neg u_4] \\
W_3 &\equiv [p_{11}\neg u_1p_{12}u_2p_{13}u_3p_{14}u_4] \\
W_4 &\equiv [p_{11}\neg u_1p_{12}\neg u_2p_{13}\neg u_3p_{14}\neg u_4] \\
W_5 &\equiv [p_{11}u_1p_{12}u_2p_{13}\neg u_3p_{14}\neg u_4] \\
W_6 &\equiv [p_{11}\neg u_1p_{12}\neg u_2p_{13}u_3p_{14}u_4]
\end{aligned}$$

| 1 | $[p_{13}u_3p_{14}u_4]$ | $[p_{13}\neg u_3p_{14}\neg u_4]$ | Total |
|---|---|---|---|
| $[p_{11}u_1p_{12}u_2]$ | $k(W_1)$ | $k(W_5)$ | $k([p_{11}u_1])$ |
| $[p_{11}\neg u_1p_{12}\neg u_2]$ | $k(W_6)$ | $k(W_4)$ | $k([p_{11}\neg u_1])$ |
| Total | $k([p_{13}u_3p_{14}u_4])$ | $k([p_{13}\neg u_3p_{14}\neg u_4])$ | $n$ |

| 2 | $[p_{12}u_2p_{13}u_3p_{14}u_4]$ | $[p_{12}\neg u_2p_{13}\neg u_3p_{14}\neg u_4]$ | Total |
|---|---|---|---|
| $[p_{11}u_1]$ | $k(W_1)$ | $k(W_2)$ | $k([p_{11}u_1])$ |
| $[p_{11}\neg u_1]$ | $k(W_3)$ | $k(W_4)$ | $k([p_{11}\neg u_1])$ |
| Total | $k([p_{12}u_2p_{13}u_3p_{14}u_4])$ | $k([p_{12}\neg u_2p_{13}\neg u_3p_{14}\neg u_4])$ | $n$ |

TAB. 6.11 – Deux tableaux de contingence possibles

Dans le premier cas pour lequel le critère gauche est défini par un digram positionnel, on remarquera qu'il est nécessaire de définir une règle selon laquelle un digram positionnel, sous-groupe d'un $N$-gram, est un critère gauche ou droit. En effet, comme un tetragram peut être divisé en deux digrams complémentaires, il est indispensable de définir lequel des deux sera critère gauche et lequel sera critère droit. Dans ce cadre, nous émettons l'hypothèse que tous les digrams positionnels contenant l'UT pivot $u_1$ sont critères gauche et que ceux qui ne contiennent pas $u_1$ sont critères droit. Cette intervention extérieure sur le calcul des attirances est bien entendu discutable mais il nous fallait décider de la catégorie de chacun des sous-groupes de façon systématique. Dans ce sens, il nous a paru plus correct de définir l'UT pivot comme élément décisif de ce choix puisque l'UT $u_1$ apparaît nécessairement à gauche de toutes les autres UTs dans le texte. Cette règle devra donc s'étendre au cas générique des modèles $N$-gram positionnels c'est-à-dire $\forall N, N \geq 2$. Dans le cadre de la généralisation de cette technique de normalisation, il sera donc nécessaire de proposer une méthode qui divise un $N$-gram positionnel générique en deux critères : l'un gauche et l'autre droit. Ainsi, on pourra évaluer les événements moyens gauche et droit d'un $N$-gram positionnel quelconque et réaliser les tests statistiques basés sur les tableaux de contingence de dimensions $2 \times 2$ pour tout groupe de $N$ UTs[10].

### 6.2.2 Généralisation

Afin de diviser un $N$-gram positionnel en deux critères, l'un gauche et l'autre droit, il est nécessaire de réintroduire la notion de point de séparation moyen — $PSM$. Dans

[10]Il est important de noter que la technique de normalisation que nous proposons ne permet pas d'évaluer le test d'indépendance entre chacun des critères considérés — $p_{i1,i2,\ldots,iN} = p_{i1}.p_{i2}...p_{iN}$.

ce cadre, le $PSM$ est une frontière qui nous permet de diviser un $N$-gram positionnel en deux sous-groupes complémentaires. Ainsi, de la même façon que nous avons utilisé le $PSM$ pour le cas des mesures d'association binaires de la théorie des probabilités et de l'information, nous nous servons de cette frontière symbolique pour définir les critères gauche et droit d'un $N$-gram positionnel générique. Dans ce cadre, nous définissons le critère gauche d'un $N$-gram positionnel comme étant l'ensemble de tous ses sous-groupes de rang 1 à $E(N/2)$ et son critère droit comme étant l'ensemble de ses sous-groupes de rang $N - E(N/2)$ à $N - 1$. Nous illustrons cette situation dans le tableau suivant.

| $PSM$ | Nb. d'UTs Critère Gauche | Nb. d'UTs Critère Droit |
|---|---|---|
| 1 | 1 | $N-1$ |
| 2 | 2 | $N-2$ |
| ... | ... | ... |
| $E(N/2)$ | $E(N/2)$ | $N-E(N/2)$ |

TAB. 6.12 – Critères Gauche et Droit

Nous pouvons analyser cette situation à partir de la figure 6.1 où le $PSM$ se déplace de gauche à droite du $N$-gram positionnel définissant ainsi les sous-groupes d'UTs appartenant au critère gauche et au critère droit.

$$\underbrace{p_{11}u_1...p_{1i}u_i}_{\text{critère gauche}} \Big|^{PSM} \underbrace{p_{1(i+1)}u_{i+1}...p_{1N}u_N}_{\text{critère droit}}$$

FIG. 6.1 – Variabilité du point de séparation moyen

Dans ce cadre, chacun des deux critères sera l'ensemble des sous-groupes d'UTs mentionnés dans chacune des deux colonnes du tableau 6.12 pour tout $N$-gram positionnel. Ainsi, à partir de la définition des critères gauche et droit, nous abordons le calcul des événements moyens gauche et droit d'un $N$-gram positionnel.

**Evénement Moyen Gauche**

Avant de proposer la formalisation des deux événements moyens, nous redéfinissons la notion d'événement. En effet, dans le cas spécifique de cette nouvelle normalisation, la notion d'événement est réduite à la simple fréquence d'un sous-groupe d'un $N$-gram positionnel. Dans ce cadre, l'événement moyen gauche d'un $N$-gram positionnel peut être défini comme étant la moyenne arithmétique des événements impliqués par son critère gauche. On le notera $EMG([p_{11}u_1...p_{1i}u_i...p_{1N}u_N])$. Dans ces conditions, nous devons donc calculer le nombre d'événements impliqués dans le critère gauche d'un $N$-gram positionnel ainsi que la somme de tous ces événements. Or, le nombre d'événements impliqués par le critère gauche d'un $N$-gram positionnel peut être calculé à partir de la fonction $nb\_ev(.)$ définie précédemment. En effet, comme le critère gauche d'un $N$-gram positionnel représente l'ensemble de ses sous-groupes de rang 1 à $E(N/2)$, il définit par opposition tous les sous-groupes complémentaires qui forment le critère droit. Ainsi, le nombre d'événements impliqués par le critère gauche est égal au nombre d'événements impliqués par le critère droit et par conséquent peut être évalué par la fonction $nb\_ev(.)$.

Une fois déterminé le nombre d'événements impliqués dans un $N$-gram positionnel générique, il est nécessaire de calculer leur somme. Dans ce cadre, nous définissons trois fonctions principales. Premièrement, nous définissons deux fonctions "outil" notées $ev\_spec\_g(.,.)$ et $ev\_spec\_gd(.,.)$ qui nous permettent de calculer la somme des événements du critère gauche d'un $N$-gram positionnel pour une valeur donnée du $PSM$. En particulier, la fonction $ev\_spec\_g(.,.)$ impose la présence de l'UT pivot $u_1$. Deuxièmement, nous introduisons une troisième fonction notée $som\_ev\_g(.)$ qui calcule la somme de tous les événements impliqués par le critère gauche d'un $N$-gram positionnel, ceci pour toutes les valeurs du $PSM$. Nous définissons donc les deux fonctions $ev\_spec\_g(.,.)$ et $ev\_spec\_gd(.,.)$ pour une valeur donnée du $PSM$ et un $N$-gram positionnel quelconque. Elles sont formulées dans les deux équations suivantes.

$$ev\_spec\_g\left(PSM, [p_{11}u_1...p_{1i}u_i...p_{1N}u_N]\right) = \tag{6.22}$$

$$\sum_{i2=2}^{j+1} \sum_{i3=i2+1}^{j+2} ... \sum_{iPSM=i(PSM-1)+1}^{N} k\left([p_{11}u_1 p_{1i2}u_{i2}...p_{1iPSM}u_{iPSM}]\right)$$

où $j = N - PSM + 1$

$$ev\_spec\_gd\left(PSM, [p_{11}u_1...p_{1i}u_i...p_{1N}u_N]\right) = \tag{6.23}$$

$$\sum_{i1=1}^{j} \sum_{i2=i1+1}^{j+1} ... \sum_{iPSM=i(PSM-1)+1}^{N} k\left([p_{i1i1}u_{i1}p_{i1i2}u_{i2}...p_{i1iPSM}u_{iPSM}]\right)$$

où $j = N - PSM + 1$

Dans le but de calculer la somme de tous les événements du critère gauche, nous devons maintenant prendre en compte les variations du point de séparation moyen. Dans ces conditions, nous distinguons le cas dans lequel $N$ est pair et le cas dans lequel $N$ est impair. Ainsi, on définira la fonction $som\_ev\_g(.)$ pour un $N$-gram positionnel de la forme suivante.

$$som\_ev\_g\left([p_{11}u_1...p_{1i}u_i...p_{1N}u_N]\right) =$$

$$\begin{cases} N = 2, ev\_spec\_g(1, [p_{11}u_1...p_{1i}u_i...p_{1N}u_N]) \\ \\ pair(N) \wedge (N > 2), \begin{cases} \sum_{PSM=1}^{E(N/2)-1} \\ ev\_spec\_gd(PSM, [p_{11}u_1...p_{1i}u_i...p_{1N}u_N]) \\ \\ +ev\_spec\_g(E\left(N/2\right), [p_{11}u_1...p_{1i}u_i...p_{1N}u_N]) \end{cases} \\ \\ impair(N), \sum_{PSM=1}^{E(N/2)} ev\_spec\_gd(PSM, [p_{11}u_1...p_{1i}u_i...p_{1N}u_N]) \end{cases} \tag{6.24}$$

A partir des résultats définis antérieurement, l'événement moyen gauche peut être facilement calculé. En effet, l'EMG n'est autre que le quotient entre la somme — donnée par la fonction $som\_ev\_g(.)$ — des événements compris dans le critère gauche et le nombre de tous les événements considérés — formulé par la fonction $nb\_ev(.)$. Ainsi, l'EMG est défini dans l'équation suivante pour un $N$-gram positionnel générique.

$$EMG([p_{11}u_1...p_{1i}u_i...p_{1N}u_N]) = \frac{som\_ev\_g([p_{11}u_1...p_{1i}u_i...p_{1N}u_N])}{nb\_ev([p_{11}u_1...p_{1i}u_i...p_{1N}u_N])} \tag{6.25}$$

De la même façon que nous avons défini l'EMG, nous présentons la définition de l'événement moyen droit — EMD — dans le prochain paragraphe.

**Evénement Moyen Droit**

L'événement moyen droit d'un $N$-gram positionnel peut être défini comme étant la moyenne arithmétique des événements impliqués par son critère droit. On le notera $EMD([p_{11}u_1...p_{1i}u_i...p_{1N}u_N])$. Dans ces conditions, nous devons calculer le nombre d'événements impliqués par le critère droit d'un $N$-gram positionnel ainsi que la somme de tous ces événements. Or, comme nous l'avons déjà déterminé, le nombre d'événements mis en évidence dans le critère droit d'un $N$-gram positionnel est donné par la fonction $nb\_ev(.)$. Ainsi, nous détaillons plus particulièrement le calcul de la somme des événements à considérer dans le cas du critère droit d'un $N$-gram positionnel. Dans ce cadre, parallèlement à ce que nous avons démontré pour l'événement moyen gauche, nous définissons deux nouvelles fonctions. Premièrement, nous proposons une nouvelle fonction "outil" notée $ev\_spec\_d(.,.)$ qui nous permet de calculer la somme des événements du critère droit d'un $N$-gram positionnel pour une valeur donnée du $PSM$. En particulier, la fonction $ev\_spec\_d(.,.)$ impose l'absence de l'UT pivot $u_1$. Deuxièmement, à partir de cette fonction et de la fonction $ev\_spec\_gd(.,.)$ précédemment définie, nous calculons dans la fonction $som\_ev\_d(.)$ la somme de tous les événements impliqués par le critère droit d'un $N$-gram positionnel, ceci pour toutes les valeurs du $PSM$. Dans un premier temps, nous définissons la fonction "outil" $ev\_spec\_d(.,.)$ pour une valeur donnée du $PSM$ et un $N$-gram positionnel quelconque. Elle est formulée dans l'équation suivante.

$$ev\_spec\_d\left(PSM, [p_{11}u_1...p_{1i}u_i...p_{1N}u_N]\right) = \tag{6.26}$$

$$\sum_{i1=2}^{j} \sum_{i3=i2+1}^{j+1} ... \sum_{iPSM=i(PSM-1)+1}^{N} k\left([p_{i1i1}u_{i1}p_{i1i2}u_{i2}...p_{i1iPSM}u_{iPSM}]\right)$$

où $j = N - PSM + 1$

Comme nous l'avons fait précédemment, nous prenons maintenant en compte les variations du point de séparation moyen. Dans ces conditions, nous distinguons les deux cas où $N$ est pair et impair. Ainsi, nous définissons la fonction $som\_ev\_d(.)$ pour un $N$-gram positionnel dans l'équation suivante.

$$som\_ev\_d\left([p_{11}u_1...p_{1i}u_i...p_{1N}u_N]\right) =$$

$$\begin{cases} N=2, ev\_spec\_d(N-E(N/2),[p_{11}u_1...p_{1i}u_i...p_{1N}u_N]) \\ \\ pair(N)\wedge \\ \\ (N>2) \end{cases} , \begin{cases} \sum_{PSM=N-E(N/2)+1}^{N-1} \\ \\ (ev\_spec\_gd(PSM,[p_{11}u_1...p_{1i}u_i...p_{1N}u_N])) \\ \\ +ev\_spec\_d(N-E(N/2),[p_{11}u_1...p_{1i}u_i...p_{1N}u_N]) \end{cases} \quad (6.27)$$

$$impair(N), \sum_{PSM=N-E(N/2)}^{E(N/2)} ev\_spec\_gd(PSM,[p_{11}u_1...p_{1i}u_i...p_{1N}u_N])$$

Finalement, à partir de ces résultats, l'événement moyen droit d'un $N$-gram positionnel peut être évalué. En effet, l'EMD peut être défini comme étant le quotient entre la somme des événements et le nombre d'événements stipulés par le critère droit d'un $N$-gram positionnel. Ainsi, l'EMD est défini dans l'équation suivante pour un $N$-gram positionnel générique.

$$EMD([p_{11}u_1...p_{1i}u_i...p_{1N}u_N]) = \frac{som\_ev\_d([p_{11}u_1...p_{1i}u_i...p_{1N}u_N])}{nb\_ev([p_{11}u_1...p_{1i}u_i...p_{1N}u_N])} \quad (6.28)$$

Grâce aux deux définitions d'événements moyens gauche et droit, nous sommes désormais en mesure de définir génériquement — $\forall N, N \geq 2$ — les deux mesures d'association binaires proposées par T. Dunning [28] et W. Gale [29] que sont le coefficient de vraisemblance Loglike et le test $\Phi^2$.

### 6.2.3 Mesures d'Association Binaires Normalisées

Afin de formuler les mesures d'association binaires basées sur l'analyse des tableaux de contingence pour le cas des modèles $N$-gram positionnels, nous proposons d'introduire les notions d'événement moyen droit et gauche dans les formules définies pour deux UTs. Ainsi, le coefficient de vraisemblance Loglike [28] et le test $\Phi^2$ [29] peuvent être définis génériquement à partir de l'EMG et de l'EMD. En effet, il suffit de remplacer les valeurs des critères gauche et droit par leurs valeurs moyennes définies par les événements moyens

gauche et droit. Dans ce cadre, le coefficient de vraisemblance Loglike est défini de la forme suivante pour un $N$-gram positionnel quelconque.

$$Loglike\left([p_{11}u_1...p_{1i}...u_ip_{1N}u_N]\right) = -2\log\lambda =$$

$$\begin{aligned} &2 \times \left(\log\theta_1^{s1}\left(1-\theta_1\right)^{n1-s1} + \log\theta_2^{s2}\left(1-\theta_2\right)^{n2-s2}\right. \\ &\quad \left. - \log\theta^{s1}\left(1-\theta\right)^{n1-s1} - \log\theta^{s2}\left(1-\theta\right)^{n2-s2}\right) \end{aligned} \tag{6.29}$$

où

$$\begin{aligned} s1 &= k\left([p_{11}u_1...p_{1i}u_i...p_{1N}u_N]\right) \\ s2 &= \left(EMD\left([p_{11}u_1...p_{1i}u_i...p_{1N}u_N]\right) - k\left([p_{11}u_1...p_{1i}u_i...p_{1N}u_N]\right)\right) \\ n1 &= EMG\left([p_{11}u_1...p_{1i}u_i...p_{1N}u_N]\right) \\ n2 &= n - EMG\left([p_{11}u_1...p_{1i}u_i...p_{1N}u_N]\right) \\ \theta_1 &= \tfrac{s1}{n1} \\ \theta_2 &= \tfrac{s2}{n2} \\ \theta &= \tfrac{EMD([p_{11}u_1...p_{1i}u_i...p_{1N}u_N])}{n} \end{aligned}$$

Dans les mêmes conditions, le test $\Phi^2$ peut être formulé de la forme suivante.

$$\Phi^2\left([p_{11}u_1...p_{1i}u_i...p_{1N}u_N]\right) = \frac{\left(n \times k\left([p_{11}u_1...p_{1i}u_i...p_{1N}u_N]\right) - \left(\begin{array}{l} EMG\left([p_{11}u_1...p_{1i}u_i...p_{1N}u_N]\right) \times \\ EMD\left([p_{11}u_1...p_{1i}u_i...p_{1N}u_N]\right) \end{array}\right)\right)^2}{\left(\begin{array}{l} EMG\left([p_{11}u_1...p_{1i}u_i...p_{1N}u_N]\right) \times \left(\begin{array}{l} n- \\ EMG\left([p_{11}u_1...p_{1i}u_i...p_{1N}u_N]\right) \end{array}\right) \times \\ EMD\left([p_{11}u_1...p_{1i}u_i...p_{1N}u_N]\right) \times \left(\begin{array}{l} n- \\ EMD\left([p_{11}u_1...p_{1i}u_i...p_{1N}u_N]\right) \end{array}\right) \end{array}\right)} \tag{6.30}$$

## 6.3 Conclusion

Dans ce chapitre, nous avons proposé une méthodologie ayant pour objectif de définir un ensemble de mesures d'association binaires de façon générique pour le cas où un $N$-gram positionnel contient plus de deux UTs. Dans ce cadre, nous avons mis en évidence deux normalisations qui s'appliquent conformément à la catégorie de chaque mesure d'association. Ainsi, la définition de l'événement moyen unique permet de généraliser le coefficient d'association [27], le coefficient Dice [45] et la probabilité conditionnelle symétrique [30]. Parallèlement, dans le cadre des mesures d'association statistiques basées sur les tests d'hypothèses [28] [29], nous avons introduit deux événements moyens — gauche et droit — qui proposent une normalisation par critère. Cette étude s'inscrit dans l'idée "d'universalité" que nous souhaitons atteindre. En effet, l'évaluation des différentes mesures d'association est rendue difficile par leurs définitions non généralisées. Dans le but de contrecarrer cette tendance, nous avons proposé la définition de mesures d'association normalisées qui puissent être évaluées de forme systématique à partir de l'application de l'algorithme GenLocalMaxs. En effet, comme nous l'avons vu dans le chapitre précédent, le GenLocalMaxs propose une plateforme d'évaluation indépendante des mesures d'association utilisées. Dans ce cadre, dans la prochaine partie de notre rapport, nous étudions les résultats des différentes mesures d'association normalisées combinées avec le GenLocalMaxs sur un ensemble d'énoncés en Français et Portugais ■

## Troisième partie

# Evaluation par Comparaison

**"L'Evaluation joue un rôle crucial dans le cadre du traitement automatique des langues naturelles tant au niveau des développeurs que des utilisateurs"**

---

*H. Thompson [87]*

# Chapitre 7

# Evaluation

L'évaluation est aujourd'hui un domaine de recherche à part entière dans le cadre spécifique du TALN. La réalisation de conférences internationales comme *LREC (Language Resources & Evaluation Conference)*, *TREC (Text Retrieval Conference)* ou *MUC (Message Understanding Conference)* en est la preuve flagrante. Traditionnellement, l'évaluation était le domaine réservé de l'Intelligence Artificielle. Cependant, la nécessité d'offrir des solutions de plus en plus fiables sur le marché des nouvelles technologies et l'urgence de comparer le comportement des prototypes de recherche se sont avérées être les éléments déclencheurs de ce qui est aujourd'hui une activité reconnue comme essentielle par tous les acteurs du TALN — tant linguistes qu'informaticiens. Ainsi, des études spécifiques ont été massivement proposées dans le cadre de la standardisation et de la validation. L'extraction d'associations lexicales n'échappe pas à cette règle. Cependant, le flou qui règne autour de la définition des phénomènes de figement impose une certaine attention. En effet, s'il est "simple" de déterminer avec précision si deux phrases sont des équivalents de traduction, décider si une séquence de mots est une association lexicale ou non dépend intrinsèquement de l'application considérée. Par exemple, pour un terminologiste, la locution complexe *en matière de* ou le déterminant *un tas de* ne devraient pas être retenus comme pertinents. Par contre, pour un lexicographe, celles-ci seraient certainement sélectionnées. De la même façon, du point de vue de la recherche documentaire, seuls les noms composés sont considérés détenteurs de sens propre. Mais qu'en est-il des verbes composés tels que *mettre au point* ou *entrer en vigueur* ? N'ont-ils pas eux aussi un sens propre ? A l'extrême, les associations lexicales comme *il y a* ou *ne — pas* paraissent bien relever des phénomènes de figement, mais ne sont que très rarement considérées comme telles dans la plupart des applications. Or, leur identification est fondamentale pour un bon nombre de travaux, notamment en traduction automatique. Les résultats du processus d'extraction dépendent donc des appli-

cations considérées. Toutefois, cette dépendance implique un certain nombre de restrictions que nous voulons éviter dans le cadre de notre travail purement exploratoire. En effet, nous sommes plus intéressés par les résultats de l'extraction que par leur pertinence par rapport à une application spécifique. Dans ce cadre, nous proposons une étude des méthodes d'évaluation existantes afin de déterminer celle que nous retiendrons pour la validation de nos résultats.

## 7.1 Terminologie de l'Evaluation

Parmi l'ensemble des stratégies d'évaluation, nous en retiendrons trois qui poursuivent des objectifs différents : l'évaluation par adéquation, l'évaluation par diagnostic et l'évaluation par performance.

### 7.1.1 Evaluation par Adéquation

Au même moment qu'un prototype de recherche sort du laboratoire et s'installe sur le marché, la question de l'adéquation entre l'offre et la demande doit se poser. Ceci revient à savoir si une application développée pour un domaine spécifique va à l'encontre des exigences de l'utilisateur et si d'autres systèmes s'approchent encore plus de ses besoins. Cette approche est souvent mise en évidence comme le paradigme du rapport du consommateur[1]. En effet, l'objectif n'est pas de déterminer le meilleur système mais de proposer une étude comparative qui permette à l'utilisateur de faire son choix en connaissance de cause. Dans le cadre du TALN, cette approche n'a traditionnellement pas connu un fort engouement. Il est même étonnant de noter qu'il faut attendre l'année 1992 pour voir publier le premier rapport sur les critères d'évaluation en traduction automatique sous l'égide de l'association Japonaise pour le développement de l'industrie électronique [88]. Dans ce cadre, la commission d'évaluation s'est concentrée sur trois aspects fondamentaux :

- Evaluation des facteurs économiques (analyse de marché),
- Evaluation technique par les utilisateurs (conformité à l'utilisation),
- Evaluation technique par les programmeurs (conformité à la spécification).

Cette approche a toutefois connu un essor important dans le cadre de l'Union Européenne. En particulier, on remarquera les efforts du groupe d'évaluation EAGLES qui définit une méthodologie d'évaluation dans laquelle les besoins des utilisateurs sont systématiquement

[1]Nous traduisons ainsi le terme anglo-saxon *Consumer Report*.

pris en compte. Dans le cadre spécifique de l'extraction de noms composés, on soulignera également l'étude réalisée par D. Bourigault et B. Habert [89]. Cependant, certains points sont discutables. Comme le soulignent J. Galliers et K. Sparck Jones [90], ce ne sont pas les systèmes qui sont évalués mais plutôt des environnements[2] c'est-à-dire des systèmes englobés dans un contexte d'utilisation particulier. Dans le même contexte, mais de manière différente, les séries ISO 9000 sur la qualité des logiciels corroborent ce point de vue :

> "*L'importance de chaque caractéristique de qualité varie suivant la classe du logiciel. Par exemple, la fiabilité est plus importante pour un logiciel critique, l'efficacité pour un logiciel en temps réel et l'utilisation pour un logiciel d'interaction avec l'utilisateur.*" [91]

Dans ce sens, l'évaluation par adéquation impose la définition des besoins des utilisateurs. Mais, il existe autant de points de vue que d'utilisateurs ! Dès lors, il n'est pas souhaitable de proposer une étude basée sur les exigences d'individualités. Dans le meilleur des cas, une étude approfondie des besoins, réalisée à partir d'un nombre suffisamment grand d'individus, devrait permettre d'identifier des classes d'utilisateurs potentiels et de construire des profils d'utilisation pour chacune de ces classes. Ainsi, chaque profil pourrait être utilisé pour déterminer les attributs d'intérêt de chaque produit pour une classe d'utilisateur. Dans le cadre de l'extraction d'associations lexicales, cette situation peut être représentée par le biais du graphique suivant 7.1.

En effet, il est clair que les classes d'utilisateurs intéressées par les systèmes d'extraction d'associations lexicales sont nombreuses. Les terminologistes autant que les traducteurs en passant par les ingénieurs du langage sont préoccupés par la qualité des résultats de leurs applications. Dans ce cadre, l'identification des unités lexicales complexes joue un rôle fondamental dû aux phénomènes d'opacité que celles-ci mettent en évidence lors des phases de compréhension et de production du langage. Ainsi, vérifier l'adéquation d'un extracteur reviendrait à le comparer aux extracteurs existants pour chacune des classes d'utilisateurs. Ceci est évidemment impossible. D'une part, l'analyse de toutes les classes d'utilisateurs suppose un travail dantesque qui ne peut être abordé dans le cadre de notre étude. D'autre part, les différents points de vue à l'intérieur des différentes classes d'utilisateurs rendent difficile la rédaction d'un "cahier des charges" cohérent pour un contexte

[2] Traduction de *setups*.

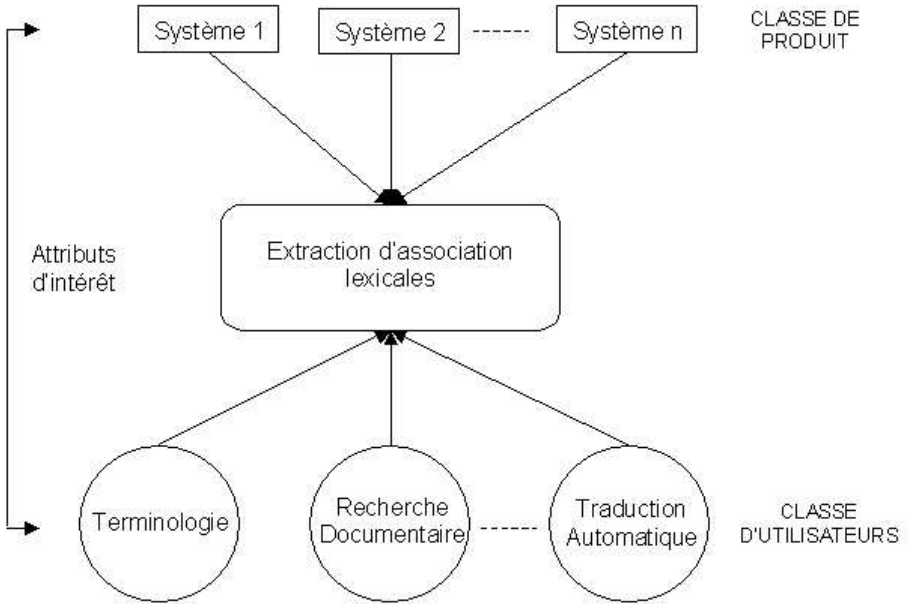


FIG. 7.1 – Evaluation par adéquation

donné. Par exemple, la construction de terminologie est un domaine en constante évolution qui suppose un certain nombre d'approches et de théories différentes. Ainsi, certains auteurs défendent l'introduction des verbes composés alors que d'autres la condamnent — ceci dépendant fortement des domaines considérés. L'évaluation par adéquation impose donc que le domaine considéré soit particulièrement stable. Or, cette condition est loin d'être satisfaite dans le domaine du TALN. En effet, les courants théoriques et empiriques n'ont pas encore atteint tel objectif, loin s'en faut. La méthode d'évaluation par adéquation ne présente donc pas une option satisfaisante dans le cadre de notre étude exclusivement exploratoire. Par conséquent, nous analysons une autre approche d'évaluation : l'évaluation par diagnostic.

### 7.1.2 Evaluation par Diagnostic

L'évaluation par diagnostic suppose la définition d'une suite de tests[3] suffisamment représentative des données d'entrée du système pour que celui-ci puisse être jugé sur la base de résultats connus *a priori*. Cette technique est généralement utilisée par les développeurs dans leur quête incessante de meilleurs résultats, mais peut également être proposée aux utilisateurs du produit. Dans le cadre du TALN, cette approche est largement diffusée dans les domaines de la traduction automatique [92] [93] et de la construction de

[3]Le terme anglais utilisé dans ce cadre est *test suite*.

grammaires explicites [94] [19] où le taux de couverture est particulièrement important. Dans ce cadre, une suite de tests est constituée d'un ensemble d'exemples dont le but est d'énumérer les phénomènes linguistiques élémentaires du domaine considéré ainsi que leurs combinaisons les plus probables. Ainsi, une suite de tests est généralement structurée en plusieurs dimensions définies par les phénomènes linguistiques élémentaires considérés, et peut également contenir des contre-exemples repérés comme tels dans la suite[4]. Une évaluation par diagnostic peut donc être résumée par la figure suivante.

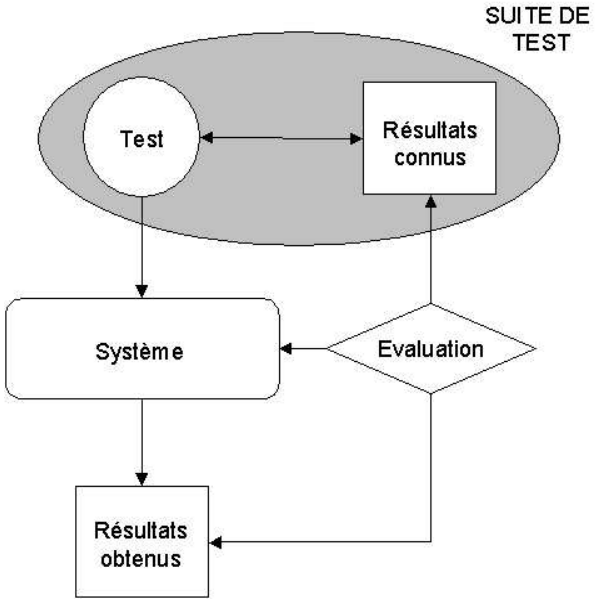


FIG. 7.2 – Evaluation par diagnostic

Les suites de tests sont particulièrement utiles aux développeurs et aux responsables de maintenance. En effet, elles permettent de s'assurer que les changements résultant de l'évaluation ont bien les résultats escomptés et pas d'autres. Cette validation s'opère le plus souvent par l'application de mesures d'analyse des suites de données. Dans le cadre spécifique de l'extraction d'associations lexicales, peu de travaux ont été proposés dans ce sens. On soulignera tout de même ceux de I. Blank [95] qui propose une évaluation à partir du système INTEX développé par M. Silberztein [14]. Bien que cette approche ne s'identifie pas exactement à une évaluation par diagnostic, elle en est néanmoins suffisamment proche pour être considérée comme telle.

[4]Cette technique est directement importée des méthodes d'apprentissage automatique.

Cependant, certaines remarques doivent être considérées. La première prend en compte le fait que les suites de tests peuvent ne pas refléter la distribution des phénomènes linguistiques présents dans tous les domaines d'application. En effet, une suite de tests n'est valide que dans le contexte d'une expérience et ne prend pas forcément en compte toutes les spécificités d'un domaine. Ainsi, dans le cadre de notre étude, est-on capable de définir une suite de tests complète c'est-à-dire un ensemble d'exemples englobant tous les phénomènes linguistiques des associations lexicales ? Ceci est peu probable. D'une part, il n'existe pas une définition claire des phénomènes de figement. D'autre part, sera-t-on jamais sûr de la validité d'un échantillon face à la diversité du matériel textuel ? La deuxième remarque attaque le fait que la création d'une suite de tests dépend forcément de l'utilisateur considéré, voire du domaine dans lequel il se trouve. En effet, le flou théorique mis en évidence dans le cadre de la définition des associations lexicales est une raison suffisante pour qu'il soit impossible d'atteindre un concensus sur la construction d'une suite de tests "normalisée". Contrairement à l'étiquetage morpho-syntaxique où les divergences sur la construction d'une suite de tests sont peu nombreuses, la définition d'une suite de tests dans le domaine de l'extraction d'associations lexicales divergerait obligatoirement selon le domaine d'activité de l'utilisateur : terminologiste, traducteur ou lexicographe. Cette situation n'est pas supportable. En effet, définir un domaine d'application pour lequel le système serait évalué, conduirait forcément à omettre certaines caractéristiques qui pourraient être bénéfiques pour un certain nombre d'autres applications. Cette méthodologie va donc contre notre volonté de définir certains principes universels afin de ne pas dépendre d'un domaine, d'une langue ou de ressources pré-existantes. Dans ce sens, nous devons introduire la dernière méthode d'évaluation : l'évaluation par performance.

### 7.1.3 Evaluation par Performance

Traditionnellement, l'évaluation par performance a été la méthode la plus utilisée par les chercheurs du TALN. Celle-ci, ayant connu une forte popularité dans le domaine de la recherche documentaire [95], s'est rapidement imposée comme étant la démarche à suivre. Dans ce contexte, trois niveaux de spécificité peuvent être distingués.

- Critère : Quel est le critère que l'on cherche à évaluer ? Précision ? Rapidité ? Couverture ? Taux d'erreur ? etc.
- Mesure : Pour chaque critère considéré, quelle est la mesure à retenir ? Pourcentage des succès par rapport aux échecs ? Temps de traitement en secondes ? Pourcentage

de succès obtenus par rapport aux succès réels ? Pourcentage d'échecs ? etc.

– Méthode : Comment détermine-t-on la valeur de la mesure considérée ? Utilise-t-on un banc d'essai ? Fait-on appel à l'analyse humaine ? etc.

Par exemple, dans le cadre de la recherche documentaire, le critère de précision est le plus souvent utilisé. On recherche alors à déterminer si les documents qui ont été retenus par le processus de fouille forment un ensemble cohérent correspondant aux besoins exprimés dans la requête initiale. Pour se faire, on calcule le pourcentage des documents extraits, réellement pertinents. Il ne reste plus alors qu'à déterminer la meilleure méthode pour réaliser ces calculs [96]. L'approche préconisée est de définir un ensemble de requêtes de tests pour lesquelles on évaluera les résultats de l'extraction. On remarquera que cette méthode n'est applicable qu'à la condition *sine qua non* que l'on sâche préalablement quel est l'ensemble des textes pertinents correspondant aux requêtes de tests. Cette situation peut être exprimée graphiquement par la figure suivante.

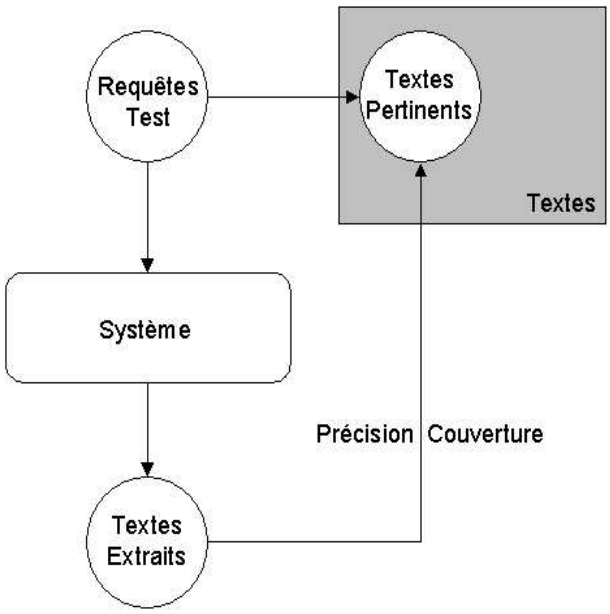


FIG. 7.3 – Evaluation par performance

Parallèlement, le taux de rappel — couverture — s'est imposé comme le second critère nécessaire pour l'évaluation des systèmes de recherche documentaire. Selon la méthode du critère de précision, le taux de rappel correspond au pourcentage de documents pertinents extraits par rapport à l'ensemble des documents préalablement identifiés comme

pertinents. Aujourd'hui, ces deux critères de précision et de rappel sont largement utilisés par un bon nombre de recherches du TALN. Cependant, l'utilisation des critères de performance cache un certain nombre d'inconvénients qu'il est nécessaire d'énumérer. Premièrement, l'évaluation par performance considère le système en analyse comme une "boîte noire". Ainsi, les caractéristiques internes du système ne sont pas analysées en tant que telles. Les données en entrée et les résultats obtenus sont les seuls retenus pour l'évaluation. Or, ceci est dangereux. Par exemple, comparer deux systèmes à partir des critères de précision et de rappel ne peut se limiter à la seule lecture des différentes valeurs obtenues. En effet, il n'est pas pensable de comparer un extracteur d'associations lexicales construit spécifiquement pour le Français — contenant un certain nombre d'heuristiques dépendantes de la langue — avec un extracteur multilingue développé à partir de la simple analyse des formes graphiques. Dans le même sens, il est difficile de comparer un système utilisant une technologie innovatrice avec un autre produit développé selon des méthodes connues. Deuxièmement, la plupart des évaluations par performance dépendent de l'application considérée. Suivant le domaine choisi pour l'évaluation, les résultats de précision et de couverture seront forcément différents d'un domaine d'application à l'autre. Troisièmement, les méthodes utilisées pour calculer les mesures de précision et de couverture préconisent la définition préalable du caractère de pertinence qui comme nous l'avons vu précédemment pose d'énormes problèmes dans le cadre de l'extraction d'associations lexicales. Face à toutes ces constatations, nous proposons, dans la prochaine section, une démarche originale pour l'évaluation des résultats de notre extracteur : l'évaluation par comparaison.

## 7.2 Evaluation par Comparaison

Si un bon nombre d'applications du TALN sont testées grâce aux méthodes d'évaluation "classiques", la tâche spécifique de l'extraction d'associations lexicales est l'exception qui confirme la règle. En effet, de nombreuses contraintes intrinsèques à ce domaine ne permettent pas de bénéficier d'un ensemble de techniques développées de façon régulière et constante. Dans un premier temps, nous présentons les différentes méthodes qui ont été utilisées dans la littérature afin de pouvoir fonder notre proposition. Dans ce domaine, plusieurs méthodes ont été proposées différant souvent selon l'objectif prétendu. Dans ce contexte, F. Smadja [31] propose une évaluation par performance pour laquelle il définit son domaine d'application comme étant celui de la lexicographie.

Ainsi, F. Smadja prétend analyser les résultats de son système XTRACT par le biais d'un expert humain, en l'occurrence le lexicographe J. Triggs du service de recherche de *Bell Communications*. L'évaluation des résultats s'est réalisée sur un échantillon de 4000 associations candidates extraites de forme aléatoire d'un ensemble de 15000 résultats de l'application de XTRACT sur un corpus de 10 millions de mots du quotidien *The Jerusalem Post*. J. Triggs s'est vu confier la tâche délicate de classer toutes ces séquences suivant trois critères : YY, Y et N. Ainsi, YY représente une association pertinente de grande qualité, Y une séquence pertinente de plus faible valeur et N une suite qui n'est en aucun cas une association pertinente. Dans ces conditions, 20% des associations extraites ont été classées YY, 20% Y et 60% N. A ce stade, plusieurs remarques se doivent d'être exprimées. Premièrement, le choix du domaine d'application est tendancieux. Pourquoi privilégier un domaine particulier au détriment d'un autre ? Pourquoi la lexicographie et non la terminologie ou la traduction automatique ? Deuxièmement, l'analyse s'est réalisée uniquement sur un échantillon d'un ensemble plus vaste d'éléments. Or, quel est le degré de certitude que l'on peut avoir sur la représentativité de cet ensemble ? En effet, aucune mention n'est faite à la théorie de l'échantillonnage et encore moins au taux d'erreur introduit par cette méthodologie. Troisièmement, la taille et le domaine du corpus de tests demandent réflexion. En effet, F. Smadja le dit lui-même dans son étude : "*les résultats dépendent de la taille du corpus ... et du contenu du corpus*". Quatrièmement, F. Smadja introduit la notion vague de qualité des associations lexicales. Mais, qu'est-ce qu'une séquence de qualité ? Finalement, on pourra regretter l'absence d'évaluation croisée entre plusieurs experts. En effet, un seul point de vue est pris en compte, celui de J. Triggs. Toutes ces questions sont importantes lorsque l'on prétend évaluer avec précision un système du TALN. Néanmoins, nous devons louer les efforts prodigués par F. Smadja dans le but de réaliser une évaluation indépendante c'est-à-dire extérieure à celle du propre développeur du système. De même, les résultats mis en évidence indiquent clairement la bonne foi de l'auteur qui ne prétend pas sur-évaluer les résultats de son système. Les remarques que nous proposons sont purement indicatives et ne retirent en rien le mérite du travail de F. Smadja. En effet, cette section a pour but principal d'alerter le lecteur sur les problèmes intrinsèques à l'évaluation des résultats dans le domaine de l'extraction des associations lexicales.

Suivant la même politique, J. Justeson et S. Katz [22] proposent une évaluation par

performance de leur algorithme de reconnaissance de termes complexes. Leur analyse se base sur l'étude de trois corpora traitant de domaines différents : chromatographie liquide, classification statistique et sémantique lexicale. Dans ce cadre, les textes sont préalablement traités avant d'être soumis à l'algorithme d'extraction. Ainsi, les tableaux, les figures et les formules sont éliminés du matériel textuel. Cette remarque est importante. En effet, comparer plusieurs systèmes implique la vérification du matériel effectivement traité ainsi que les techniques utilisées. Nous reviendrons plus loin sur cette question. Dans le cadre de leur évaluation proprement dite, J. Justeson et S. Katz notent qu'il n'est pas difficile de repérer les séquences qui ne sont pas des termes. Ainsi, de nombreux termes candidats peuvent être facilement éliminés. Cependant, lorsqu'il s'agit de définir si un groupe nominal est un terme technique ou non, la question n'est pas aussi simple. Ces auteurs utilisent même le terme de subjectivité de l'analyse des résultats et affirment qu'ils "*... n'ont pas été capables de définir une méthode qui permette de mesurer directement la qualité terminologique des groupes nominaux*". Ils proposent malgré tout une mesure de rappel basée sur l'existence d'un dictionnaire terminologique et une mesure de précision calculée à partir de jugements subjectifs. D'une part, le taux de rappel a été évalué grâce au dictionnaire de Physique et de Mathématique *Lapedes* [97] à partir duquel tous les termes présents dans le texte ont été repérés. Dans ce cadre, le taux de rappel atteint 85% mais la faible couverture du dictionnaire impose une vision très critique des ces résultats. D'autre part, le taux de précision a été calculé pour les trois textes à partir d'une définition subjective et très souple de la notion de terme. Dans ces conditions, des résultats très disparates ont été obtenus pour l'ensemble des textes. Les taux de précision vont de 76% pour le texte sur la chromatographie liquide à 94% pour le cas de l'énoncé sur la classification statistique en passant par 87% pour le texte sur la sémantique lexicale. A ce stade, plusieurs remarques méritent d'être énoncées. Premièrement, J. Justeson et S. Katz mettent en évidence l'importance du domaine des textes pour l'évaluation des performances d'un système. En effet, différents domaines impliquent différents résultats d'extraction. Deuxièmement, ils proposent une alternative possible au problème du calcul du taux de rappel en recourant à l'existence de ressources linguistiques préalablement compilées. Seulement, il est important de noter que de telles ressources ne sont pas disponibles dans tous les domaines et encore moins dans toutes les langues. De plus, la véritable utilité de telles données reste encore à prouver. Les auteurs en sont d'ailleurs bien conscients lorsqu'ils pointent le faible nombre de termes réellement reconnus dans

les textes. Certains éléments de l'évaluation doivent cependant faire l'objet d'un certain nombre de critiques. D'une part, les résultats sont analysés par les propres auteurs du système ce qui est loin de garantir une vision indépendante de l'évaluation. D'ailleurs, J. Justeson et S. Katz affirment sans complexe que la définition de la notion de terme complexe est très souple et que leur jugement sur les résultats est somme toute subjectif. On regrettera donc que ceux-ci n'aient pas validé leurs résultats à partir d'une analyse croisée externe. D'autre part, les deux auteurs ne mentionnent pas la taille des corpora utilisés réduisant ainsi le degré de confiance des résultats comme le souligne F. Smadja. Finalement, une question évidente se pose. Peut-on comparer les taux de précision de Justeson/Katz avec ceux obtenus par F. Smadja ? La réponse est bien évidemment Non. En effet, l'objectif de chacun des systèmes est clairement distinct. Alors que F. Smadja propose une étude purement exploratoire, J. Justeson et S. Katz s'attaquent à la tâche spécifique de l'extraction de terminologies qui peut être considérée comme une sous-tâche de l'extraction d'associations lexicales. D'autre part, les méthodologies employées ainsi que les traitements réalisés sur les textes sont fondamentalement différents. Alors que la méthodologie de F. Smadja est indépendante de la langue, celle proposée par J. Justeson et S. Katz dépend intrinsèquement de la langue Anglaise.

Dans le même ordre d'idée, S. Shimohata [32] propose une catégorisation des associations lexicales afin de produire le taux de précision de son algorithme. Ainsi, il définit dans un premier temps quatre types d'éléments potentiellement identifiables : les phrases complètes CS, les unités grammaticales GU, les unités sémantiques non grammaticales MU et les fragments fonctionnels F. Son évaluation s'est réalisée sur un corpus en langue anglaise de 1 311 522 mots dans le domaine de l'informatique. L'une des particularités de l'analyse de S. Shimohata réside dans la présentation de différents résultats suivant les paramètres utilisés lors de l'extraction, notamment les différentes valeurs seuil d'entropie. Ainsi, S. Shimohata met en évidence une caractéristique importante de l'évaluation des systèmes d'extraction à valeur limite. En effet, les résultats dépendent intrinsèquement de l'utilisation de valeurs seuil plus ou moins permissives pour que le taux de précision et le taux de rappel soient les plus satisfaisants possibles. Dans ce cadre, trois valeurs seuil d'entropie — 2, 1,5 et 1 — donnant respectivement lieu à l'extraction de 650, 1950 et 6774 séquences de formes graphiques sont proposées. Dans le premier cas, 25% des unités appartiennent au groupe des CS, 46% à la classe des GU, 17% à celle des MU et 12% au

groupe des F, ce qui permet à S. Shimohata de mettre en évidence un taux de précision de 88%. Malheureusement, les résultats obtenus pour les deux autres valeurs seuil ne sont pas accessibles. Néanmoins, il est clair, selon notre expérience, que ceux-ci sont notablement plus faibles du fait du degré de liberté qui est introduit. Dans un deuxième temps, dans le but d'évaluer un sous-ensemble d'unités potentiellement pertinentes par le biais d'un expert humain, S. Shimohata propose un raffinement par valeur seuil des unités précédemment extraites. Ainsi, il définit un ensemble unique de trois valeurs seuil — une de fréquence, une d'entropie et une de quotient — qui sont censées maximiser le taux de précision de l'extraction. A partir des 650 séquences retenues préalablement, seulement 269 sont repérées par la deuxième étape d'extraction. Dans ce cadre, les résultats de l'expertise humaine mettent en évidence 180 associations lexicales représentant un taux de précision de 67%. Comme nous l'avons remarqué précédemment, S. Shimohata souligne clairement la dépendance aux valeurs seuil du processus d'extraction.

Cependant, il ne propose pas les chiffres de précision obtenus pour ces différentes expériences. En effet, il est clair que ceux-ci seraient foncièrement plus faibles. Mais il convient de s'arrêter sur ce point. Quelle est la légitimité des valeurs seuil utilisées ? Celles-ci sont optimisées pour ce corpus en particulier mais ne sont certainement pas applicables à de nouveaux corpora. Ainsi, il serait nécessaire de réévaluer les valeurs limites choisies pour quelconque autre énoncé. Ceci complique notablement la tâche d'évaluation d'un système d'extraction. Parallèlement, S. Shimohata ne définit pas avec précision chacune des catégories proposées. Notamment, la catégorie MU mériterait une explication. Encore une fois, il n'est pas possible de comparer ce système avec les deux précédents à partir des seuls taux de performance. En effet, il faudrait considérer en toute équité le risque de proposer une nouvelle méthode par rapport à une méthode dont les preuves ont déjà été faites. Particulièrement au taux de rappel, S. Shimohata affirme raisonnablement que celui-ci ne peut être valablement calculé à partir des données dont il dispose. En effet, il n'existe pas de corpus de référence où toutes les associations lexicales ont été repérées et la difficulté de définir ce que sont exactement les associations lexicales empêche la normalisation des corpora utilisés. Dans ce sens, de nombreux efforts restent à mener afin de proposer des ressources indispensables à cette tâche spécifique.

Selon cette approche, l'une des évaluations les plus complètes est proposée par K. Frantzi

et S. Ananiadou [98] dans le cadre spécifique de la construction de terminologies. Dans un premier temps, elles proposent une analyse croisée des résultats grâce à l'intervention extérieure de deux spécialistes. Dans ce cadre, un sous-ensemble des associations lexicales candidates extraites à partir d'un corpus médical de 860 000 mots est proposé à un terminologiste et à un expert du domaine d'application afin d'être analysé. K. Frantzi et S. Ananiadou suggèrent ainsi d'éviter la subjectivité de l'analyse en proposant plusieurs points de vue sur les données. Dans un deuxième temps, elles proposent de déterminer les différents taux de précision et de rappel constatés lors de l'analyse de plusieurs scenari d'extraction. Ainsi, suivant les valeurs seuil de *C-value* utilisées, les résultats diffèrent radicalement. Nous rappelons les résultats présentés par K. Frantzi et S. Ananiadou dans leur article [98] dans le tableau 7.1.

| *C-value* | Précision | Rappel |
|---|---|---|
| 2057 - 150 | 82% | 30% |
| 2057 - 13 | 65% | 82% |
| 2057 - 7 | 58% | 85% |
| 2057 - 4 | 45% | 93% |

TAB. 7.1 – Précision et Rappel

La distribution entre taux de précision et taux de rappel est révélatrice de la nécessité de prendre en compte les différents paramètres d'un système afin de pouvoir l'évaluer avec rigueur. En effet, l'ensemble des systèmes à valeurs seuil démontrent le même comportement. Le gain en précision se fait au détriment de la couverture du système et vice versa. Que doit-on alors évaluer ? La précision ou le rappel ? Les deux ? Difficile de répondre à cette question. Peut-être, ni l'une ni l'autre. Dans ce sens, K. Frantzi et S. Ananiadou montrent leurs préoccupations relativement à leur évaluation. En effet, elles affirment que seules des estimations sur les taux de performance peuvent être produites. Entre autres, le taux de rappel n'est calculé qu'à partir d'une faible proportion du texte initial — 1000 mots seulement. D'autre part, le taux de précision n'est valide que pour ce texte en particulier et varie selon les paramètres utilisés. Les auteurs font également remarquer les erreurs provenant du pré-traitement linguistique de l'énoncé. En effet, quel est le pourcentage d'erreur induit par l'étiquetage morpho-syntaxique dans le processus

d'extraction ? L'évaluation d'un système ne peut donc en aucun cas se limiter au simple calcul des performances d'une "boîte noire" ; elle doit prendre en compte un certain nombre de paramètres internes au système comme la technologie utilisée, son champ d'application ou encore sa portabilité. Dans ces conditions, il ne nous semble pas opportun d'évaluer un système d'extraction d'associations textuelles à partir de la seule mesure de ses performances.

A l'autre extrême, un certain nombre d'auteurs se sont limités à lister leurs résultats sans proposer de comparaison rigoureuse avec d'autres systèmes ou d'autres méthodologies ni de validation des résultats obtenus. Parmi ceux-ci, on soulignera D. Bourigault [15] qui donne quelques exemples de termes extraits par LEXTER dans le cadre de la constitution d'un glossaire de documentation technique et de la construction d'un système de recherche documentaire chez EDF. Malheureusement, aucune allusion sur la qualité des résultats n'est exposée. Parallèlement, K. Church et P. Hanks [27] dévoilent une étude sur le coefficient d'association à partir de l'analyse de certains cas particuliers c'est-à-dire à partir de l'étude des termes cooccurrents avec *doctor*, *set*, *save* et *save ... from*. Les deux auteurs proposent également une liste de huit associations lexicales calculées à partir d'un corpus de 44 millions de mots ! Là encore, aucune alternative n'est proposée à la valeur seuil définie de façon *ad hoc*, ce qui pourrait certainement représenter une analyse comparative intéressante. De plus, aucune comparaison n'est effectuée à partir de mesures d'association similaires. Finalement, on remarquera l'analyse multilingue de C. Enguehard [23] dans laquelle elle propose deux listes de termes complexes extraits à partir de deux corpora différents, l'un en Français de 120 000 mots et l'autre en Anglais de 22 000 mots. En particulier, le premier sur le nucléaire donne lieu à l'extraction de 3000 nouveaux concepts alors que le deuxième contenant un ensemble d'articles scientifiques met en évidence l'extraction de 200 termes complexes. Bien que C. Enguehard ne propose pas de point de référence par rapport à certaines méthodologies parallèles et ne tente pas de classer l'ensemble des phénomènes linguistiques extraits, elle a le mérite d'illustrer la flexibilité de son système ANA qui peut être adapté à d'autres langues sans trop de problèmes. La position de ces auteurs est toutefois fortement critiquable. En effet, rien ne permet réellement d'évaluer les méthodes proposées.

Afin de faire face à ces critiques, un certain nombre d'auteurs, notamment dans les

domaines syntaxico-numérique et numérique, proposent de comparer leurs résultats d'extraction à partir de l'utilisation de différentes mesures d'association. Dans ce cadre, l'idée fondamentale est d'étudier le comportement de chaque mesure en considération et, dans le meilleur des cas, d'établir laquelle serait une bonne mesure pour l'extraction d'associations lexicales. Ainsi, on est loin des méthodes d'évaluation présentées précédemment qui considèrent les systèmes d'extraction comme des "boîtes noires". En effet, suivant cette approche, l'outil logiciel se distingue par sa flexibilité permettant de tester diverses mesures d'association en son sein. L'évaluation est ainsi réalisée directement sur la technologie employée. Suivant ces constatations, nous ferons référence à cette approche comme étant la méthode d'évaluation par comparaison. T. Dunning [28] est le premier à proposer une étude comparative entre deux mesures d'association. Dans ce cadre, il compare le coefficient de vraisemblance et le test $\Phi^2$ appliqués sur le même corpus de 31 777 mots de l'Union de Banque Suisse. T. Dunning se donne ainsi comme objectif principal d'illustrer l'inadéquation du test $\Phi^2$ — proposé préalablement par W. Gale [29] — lorsque la taille du corpus testé est faible. Les résultats obtenus montrent que les termes extraits par le coefficient de vraisemblance sont notablement plus naturels que ceux mis en évidence par le test du $\Phi^2$ qui ... *sur-évalue dramatiquement les associations lexicales rares.* T. Dunning se limite ainsi à réhausser les différences de comportement entre les deux mesures d'association. En effet, son but n'est pas de viser une application particulière mais plutôt de définir clairement les différences qui peuvent exister entre différentes mesures. Ainsi, il ne porte aucun jugement de valeur sur l'adéquation à une application ou à un domaine donnés des associations extraites. Son objectif est purement exploratoire. Cette vision de l'évaluation est particulièrement séduisante dans un domaine où les éléments recherchés sont difficiles à définir de façon claire. Deux remarques s'imposent tout de même. D'une part, l'absence d'une évaluation sur plusieurs textes est regrettable. D'autre part, T. Dunning ne propose pas une méthode complète d'extraction d'associations lexicales laissant sans réponse le problème des valeurs seuil.

Suivant la même approche, on soulignera également les travaux réalisés par R. Feldman *et al.* [34] pour la construction de taxonomies hiérarchiques. Dans ce cadre, les termes complexes sont extraits à partir de filtres linguistiques et de l'utilisation de méthodes d'amorçage. Ainsi, deux unités textuelles sont considérées cooccurrentes si elles mettent en évidence une structure syntaxique répertoriée et si leur degré de cohésion, mesuré

à l'aide d'un modèle mathématique donné, dépasse une certaine valeur seuil. Afin d'obtenir un taux de rappel le plus expressif possible, R. Feldman *et al.* proposent donc d'utiliser quatre mesures d'association différentes : la fréquence d'occurrence, le test $\Phi^2$, le coefficient d'association et le coefficient de vraisemblance. Ainsi, les quatre mesures sont successivement implémentées dans le système d'extraction donnant lieu à quatre ensembles distincts d'associations lexicales qui sont alors utilisés comme un seul tout dans la phase de filtrage[5]. Suivant cette méthodologie, R. Feldman *et al.* révèlent dans leur article [34] une expérience réalisée à partir d'un corpus extrait d'une collection de l'agence *Reuters* d'environ 44 millions de mots. En particulier, ils mettent en évidence le recours à deux valeurs seuil : l'une pour la fréquence d'occurrence et l'autre pour les mesures d'association utilisées. Malheureusement, aucune information n'est donnée sur les valeurs définies. En effet, les auteurs se limitent simplement à mentionner qu'une valeur seuil unique a été déterminée pour chacune des mesures d'association utilisées sans pour autant préciser sa valeur. Dans ces conditions, l'évaluation ne peut en aucun cas être jugée concluante. La même remarque peut être formulée par rapport à l'utilisation d'un texte unique de test. Cependant, ces travaux s'inscrivent dans une optique bien particulière. L'intérêt principal est en effet de proposer un taux de rappel le plus satisfaisant possible pour la tâche spécifique de l'extraction d'associations lexicales. Dans ce cadre, R. Feldman *et al.* soulignent un point important de la méthode d'évaluation par comparaison. Chaque mesure d'association réhausse en effet une caractéristique particulière des associations lexicales. Ainsi, on suppose qu'il n'existerait pas une seule mesure — et donc une seule méthodologie — pour l'extraction d'associations lexicales mais plutôt un ensemble d'elles. Dans le cadre spécifique de l'évaluation, cette remarque est fondamentale. Ainsi, plutôt que de tenter de découvrir ce qui serait la "meilleure" mesure d'association, une étude détaillée du comportement de chacune serait bien plus utile au développement de nouvelles techniques d'extraction. C'est en tout cas notre intuition.

Dans le même ordre d'idée, C. Zhai [99] propose une évaluation par comparaison de quatre mesures d'association : le coefficient d'association MI, la mesure de cooccurrence mot-segment PWC, la mesure d'association de mots WA et celle de similitude de contexte CS[6]. L'étude proposée dans son article "*Exploiting Context to Identify Lexical*

[5]Nous ne nous attarderons pas sur cette notion. Le lecteur intéressé trouvera les informations nécessaires dans leur article [34].

[6]De plus amples informations sur ces mesures pourront être trouvées dans l'article [99] de l'auteur.

*Atoms — a Statistical View of Linguistic context*"[99] à partir d'un texte en anglais de 20 Méga-octets tiré du quotidien *Association Press Newswire*, est particulièrement intéressante. Contrairement à R. Feldman *et al.*, C. Zhai propose une analyse qualitative des résultats obtenus à partir des quatre mesures énoncées précédemment.

Ainsi, il expose la liste des 10 "meilleures" associations lexicales extraites ainsi que les 5 plus "mauvaises" pour chacune des mesures testées. Dans ce cadre, il réalise une rapide analyse qualitative qui lui suggère que les quatre mesures sont de bonnes heuristiques mathématiques pour l'extraction d'associations lexicales ! Afin d'approfondir son évaluation, il fait appel à la validation "manuelle" d'un échantillon de 400 associations extraites pour chacune des mesures considérées. Dans ce cadre, il offre une classification intéressante des résultats de précision[7]. Ainsi, il propose successivement les taux de précision de chacune des mesures pour les 10 premières paires extraites, puis pour les 20 premières et ainsi de suite jusqu'aux 400 premières paires. Les résultats montrent que toutes les mesures ont pratiquement le même comportement à l'exception de la similitude de contexte qui semblerait donner des résultats moins pertinents. Dans le cadre de cette étude, on remarquera également que C. Zhai a évalué différentes tailles de contexte — environnement immédiat — pour vérifier l'indépendance des résultats par rapport à l'ensemble des paramètres du système. Néanmoins, peu de références concrètes sont faites à ce sujet mais C. Zhai affirme que le processus d'extraction ne varie que très peu suivant les fluctuations de l'environnement immédiat.

Dans le cadre spécifique de la construction de terminologies, nous pourrions également citer les travaux de B. Daille [21] qui propose une analyse croisée externe de différents résultats obtenus à partir de son extracteur ACABIT pour lequel la fréquence, le test $\Phi^2$ et le coefficient de vraisemblance sont successivement testés. Cependant, nous n'introduirions pas d'éléments nouveaux qui pourraient guider notre proposition d'évaluation. Ainsi, nous annonçons finalement les différentes étapes de l'évaluation de notre extracteur. D'une part, l'objectif de notre travail est purement exploratoire. Dans ce cadre, nous ne voulons en aucun cas évaluer notre système à partir d'une application donnée. Cette position peut paraître radicale, mais comme nous l'avons déjà mentionné, dans l'absolu,

[7] Nous ne reviendrons pas sur les remarques déjà énoncées sur les problèmes intrinsèques de la définition du taux de précision. C. Zhai le souligne lui-même : "*Il est difficile de définir un critère satisfaisant ... pour juger si une séquence est une association lexicale ou non*".

la validation des résultats de l'extracteur serait nécessairement biaisée par une analyse subjective des résultats. Face à cette constatation et aux solutions qui ont été proposées dans la littérature, proposer une simple liste des associations extraites serait foncièrement réducteur. Par conséquent, nous aborderons une approche d'évaluation par comparaison dans laquelle nous analyserons successivement les résultats obtenus à partir de six mesures d'association normalisées : le coefficient d'association [27], la probabilité conditionnelle symétrique [30], le coefficient Dice, le test $\Phi^2$ [29], le coefficient de vraisemblance LogLike [28] et bien entendu l'Expectative Mutuelle. Il est clair que d'autres mesures d'association pourraient être considérées[8], mais le lecteur comprendra facilement qu'il n'est pas possible d'évaluer toutes les mesures d'association existantes. Nous avons donc dû centrer notre recherche sur un sous-ensemble de mesures couramment utilisées dans les applications du TALN[9].

L'un des problèmes principaux mis en évidence par les approches précédentes dans le cadre de l'évaluation par comparaison réside dans la capacité des systèmes à proposer une plateforme commune à toutes les mesures d'association. En effet, les valeurs seuil doivent être spécifiquement déterminées pour chaque cas de figure c'est-à-dire chaque fois que la langue, le domaine, la longueur ou le genre de l'énoncé changent. Ainsi, les résultats de l'extraction sont forcément biaisés par la définition plus ou moins fine des valeurs seuil. D'autre part, la comparaison entre mesures d'association est rendue difficile par l'application de cette méthodologie puisque pour chaque mesure il est nécessaire de préciser une valeur seuil particulière. Or, s'il est difficile de définir les valeurs seuil pour une même mesure suivant les conditions d'expérience, il est bien plus compliqué de calculer les valeurs seuil optimales pour différentes mesures d'association. Dans ce cas, l'utilisateur doit être familier à l'ensemble des mesures statistiques qu'il veut comparer. Suivant ces constatations, l'algorithme GenLocalMaxs propose une solution originale au problème de l'évaluation par comparaison. En effet, le GenLocalMaxs constitue une plateforme unique capable de recevoir sans limitation ni ajustement n'importe quel type de mesure d'association normalisée. Ainsi, il suffit d'appliquer une nouvelle mesure d'association à chaque $N$-gram positionnel et d'utiliser le GenLocalMaxs pour extraire un

[8] En particulier, notre évaluation peut pêcher sur ce point.

[9] A titre d'information, différents tests ont été réalisés à partir des mesures de Cramer [82] et du coefficient de Pearson [82]. Mais les résultats obtenus étant très proches de ceux du test $\Phi^2$, leur présence dans cette évaluation ne se justifie pas.

nouvel ensemble d'associations lexicales. Cette caractéristique est un atout fondamental en faveur du GenLocalMaxs dans le cadre de la validation des résultats d'extraction.

Finalement, afin d'illustrer les résultats d'extraction obtenus à partir de chaque mesure d'association, nous proposerons dans un premier temps une étude qualitative des associations lexicales extraites par le GenLocalMaxs. Dans ce cadre, nous essaierons de regrouper les séquences candidates suivant certaines catégories que l'on déterminera sur la base de différentes études qui se sont chargées de définir un certain nombre de classes génériques d'associations lexicales[10]. Notre objectif sera alors d'informer le lecteur sur les différents comportements de chacune des mesures afin de lui permettre un choix avisé selon l'application envisagée. Parallèlement, nous évaluerons le processus d'extraction à partir de l'analyse de textes de différentes langues — notamment le Français et le Portugais. Le but de cette opération est évidemment de démontrer la flexibilité du système et de vérifier son indépendance aux caractéristiques du matériel textuel proposé en entrée. Dans un deuxième temps, nous présenterons une analyse quantitative des résultats. En particulier, nous nous intéresserons aux pourcentages d'associations lexicales extraites — continues ou non, à la longueur moyenne des séquences repérées, à leur fréquence moyenne etc. Cette analyse aura pour but de renforcer l'analyse qualitative afin d'illustrer encore plus les différences mises en évidence par chacune des mesures d'association. Finalement, dans le même contexte, nous évaluerons les influences du changement de taille de l'environnement immédiat ainsi que la longueur des textes en entrée sur les résultats du processus d'extraction.

## 7.3 Conclusion

Bien entendu, des critiques sûrement judicieuses peuvent être exprimées sur notre vision de l'évaluation par comparaison. Notre propre analyse montre d'ailleurs quelques lacunes dont l'absence regrettable d'une validation externe des résultats ou plutôt d'une classification externe des séquences candidates. Malheureusement, il ne nous a pas été possible d'orienter notre analyse dans ce sens pour diverses raisons — notamment de temps et de logistique. Cependant, nous sommes conscients qu'une analyse externe croisée est bien souvent nécessaire à la "validation" rigoureuse et impartiale des résultats. On remarquera cependant qu'une telle analyse a été menée à bien par la linguiste Spela Vintar de l'Uni-

[10] En particulier, on soulignera les travaux de B. Daille [21] et de G. Gross [51].

versité de Ljubljana en Slovénie dans le cadre d'un corpus bilingue de petite taille Anglais-Slovène [100]. De même, Jaan-Heiki Kaalep de l'Université de Tartu en Estonie a conduit un certain nombre d'études sur les verbes composés de l'Estonien [101]. Parallèlement, une étude inter-domaine a également été proposée dans [102] montrant que les résultats d'extraction dépendent effectivement du domaine d'application. Là encore, nous n'aborderons pas cet aspect de l'évaluation dans ce rapport pour des raisons évidentes d'espace. Dans tous les cas, il nous semble préférable et surtout plus utile de proposer une analyse exhaustive des résultats plutôt que de tenter résumer un ensemble riche de phénomènes linguistiques sous la forme de taux de performance souvent sur-évalués. Néanmoins, cette affirmation ne nous empêchera pas de proposer certains chiffres sur les performances des différents scenari testés. Cependant, ces chiffres serviront uniquement de guide au lecteur et ne pourront en aucun cas être utilisés comme mesures de comparaison entre systèmes ■

# Chapitre 8

# Analyse Qualitative des Résultats

L'objectif principal de ce chapitre est triple. Dans un premier temps, nous prétendons tester la validité de nos hypothèses initiales sur la rigidité des associations lexicales. Ainsi, nous voulons vérifier la justesse de nos théories à partir de la réalité du matériel textuel. En fait, nous désirons montrer que notre architecture est réellement capable d'identifier un certain nombre de phénomènes linguistiques relevant des expressions figées. Dans ce sens, nous nous baserons sur deux études fondamentales proposées par G. Gross [51] et B. Daille [21]. Ainsi, G. Gross propose une classification des phénomènes de figement suivant sept catégories : les noms et déterminants composés, et les locutions verbales, adjectivales, adverbiales, prépositives et conjonctives. Parallèlement, B. Daille complète cette analyse par l'introduction de trois nouvelles notions : la surcomposition, la modification et la coordination. Dans un deuxième temps, nous voulons illustrer le bien fondé de la nécessité de définir une nouvelle mesure d'association, en l'occurrence l'Expectative Mutuelle. Pour se faire, nous utiliserons les résultats d'extraction obtenus à partir des mesures d'association suivantes : le coefficient d'association [27], le coefficient Dice [45], la Probabilité Conditionnelle Symétrique [30], le test $\Phi^2$ [29] et le coefficient de vraisemblance LogLike [28]. Ainsi, nous rapporterons leurs caractéristiques propres ainsi que leurs lacunes principales. Finalement, dans un troisième temps, nous mettons en évidence la flexibilité de notre système en l'appliquant au Français et au Portugais sans pour autant le modifier en quoi que ce soit. Du fait de son caractère purement statistique et de l'analyse exclusive de textes non traités, cette architecture peut être facilement appliquée à tout type de matériel textuel et *a fortiori* à n'importe quelle langue[1].

[1]Nous noterons que différentes expériences ont été réalisées à partir de corpora en Italien [103], Slovène [100], Estonien [101] et Anglais [104] [105] outre le Français et le Portugais [106] [107].

## 8.1 Analyse par Catégories

Dans cette première section, nous proposons de vérifier que l'ensemble des phénomènes linguistiques mis en évidence par G. Gross et B. Daille, dans leurs études respectives, sont réellement présents dans les résultats d'extraction. Nous voulons ainsi tester la validité de nos hypothèses initiales sur la non flexibilité des associations lexicales et sur le principe d'intégrité du matériel textuel. Pour se faire, nous avons donc extrait d'une base de textes de la Commission Européenne un corpus parallèle Français-Portugais d'exactement 200 000 formes graphiques[2]. A partir de celui-ci, nous avons appliqué le GenLocalMaxs associé à l'Expectative Mutuelle pour un environnement immédiat de trois[3] unités — i.e. trois mots à droite et trois mots à gauche de l'unité pivot. Par conséquent, un ensemble désordonné de $N$-grams positionnels — $\forall N, N = 2..6$ — que nous nous proposons d'analyser rigoureusement a été extrait.

### 8.1.1 Classification de G. Gross

G. Gross [51] propose une étude exhaustive du phénomène de figement pour la langue Française. Dans celle-ci, il suggère sept catégories suivant lesquelles les associations lexicales peuvent être classées : les noms et déterminants composés, ainsi que les locutions verbales, adjectivales, adverbiales, prépositives et conjonctives. Notre première analyse revient donc à vérifier que l'ensemble de ces phénomènes linguistiques sont couverts par notre extracteur tant pour le Français que pour le Portugais.

#### Noms Composés

Le nom composé a traditionnellement été repéré par l'évocation dans l'esprit d'une image unique et non de l'ensemble des images distinctes répondant à chacun des mots constituants. Dans ce sens, il n'est pas étonnant de constater que la terminologie ait été l'un des instigateurs principaux de l'étude des noms composés. Ainsi, comme le souligne G. Gross, "*si l'on fait l'inventaire du vocabulaire des langues de spécialités, on se rend compte que les noms composés s'y taille la part du lion*". Les noms composés ont la même distribution syntaxique que les noms simples et fonctionnent donc dans

[2]La base de textes a été acquise auprès de l'Association Européenne pour les Ressources Linguistiques — http ://www.icp.inpg.fr/ELRA/home.html — et est cataloguée sous la référence : W0023. Les corpora extraits de celle-ci ont pour nom *den30419.html* pour le Français et *dpt30419.html* pour le Portugais.

[3]Nous justifierons ce choix dans la partie suivante de ce rapport.

la phrase comme ces derniers. Cependant, d'un point de vue interne, ce sont des suites qui n'ont pas la liberté de fonctionnement des groupes nominaux ordinaires. Ainsi, on opposera le groupe nominal ordinaire *livre vert* comprenant un substantif accompagné d'un adjectif qualificatif — *le livre qui est vert* — au nom composé *Livre Vert* — *Rapport de la Commission Européenne* — suite inanalysable du point de vue syntaxique et dont le sens est des plus opaques. Les noms composés ont donc cette particularité d'allier l'unité à la pluralité. Afin de repérer ces associations lexicales de type nominal, G. Gross propose une série de règles qu'elles doivent respecter : absence de libre actualisation des éléments composants, non-prédication et structures internes atypiques, entre autres. Comme nous l'espérions, de nombreux noms composés ont été repérés par notre système, pour la plupart spécifiques du domaine de l'Union Européenne. Nous en présentons une liste non exhaustive tant pour le Français que pour le Portugais dans les tableaux 8.1 et 8.2.

Il est intéressant de noter qu'une grande proportion des noms composés extraits sont communs aux deux langues. Ceci était cependant prévisible du fait de la proximité du Français et du Portugais. Malgré tout, quelques différences subsistent. Par exemple, le concept *renvoi en commission* ne forme pas une association lexicale de type nominal en Portugais. Cette notion est ainsi rapportée par l'usage d'une forme complexe telle que *seja enviado de novo para a Commissão* (*soit renvoyé de nouveau à la Commission*). Il en est de même pour les *vifs remerciements* qui sont tantôt traduits par *agradecer muito especialmente* (*remercier très spécialement*) ou bien par *gostaria de agradecer* (*j'aimerais remercier*). Parallèlement, le caractère atomique des associations lexicales est clairement mis en évidence par les résultats de l'extraction. En effet, en Français, le concept *Extrême-Orient* ne forme qu'une seule unité graphique alors que pour le Portugais celui-ci est réalisé par l'agglomérat des deux formes *Extremo* et *Oriente*. En particulier, il serait intéressant d'analyser l'évolution "historique" de la réalisation lexicale de cette notion. L'hypothèse la plus probable est que la force qui lie ces deux mots soit devenue telle que leur réunion par un trait d'union est apparue inévitable. Le même phénomène s'applique aux *Etats Membres* qui sont traduits en Portugais par le nom composé *Estados-membros* qui forme une et une seule forme graphique.

| ME | Fréq. | Noms composés |
|---|---|---|
| 0.000395161 | 140 | Etats membres |
| 0.000134914 | 84 | Parlement européen |
| 0.000108456 | 53 | formation professionnelle |
| 7.80531e-05 | 21 | réforme de la PAC |
| 4.48148e-05 | 11 | Nations unies |
| 1.8375e-05 | 7 | renvoi en commission |
| 1.875e-05 | 5 | voie à suivre |
| 2.27273e-05 | 5 | commission des libertés publiques |
| 9e-06 | 3 | chantier naval |
| 9e-06 | 3 | conseil d'administration |
| 8.18182e-06 | 3 | chrétiens démocrates |
| 1.35e-05 | 3 | Sir Jack Stewart-Clark |
| 6.66667e-06 | 2 | boucs émissaires |
| 5e-06 | 2 | vifs remerciements |
| 5e-06 | 2 | purification ethnique |
| 8.57143e-06 | 2 | Fonds social européen |
| 1e-05 | 2 | dioxyde de carbone |
| 8e-06 | 2 | commission économique et monétaire |
| 1e-05 | 2 | Année européenne des personnes âgées |
| 1e-05 | 2 | Politique de neutralité d'Etats membres potentiels |

TAB. 8.1 – Noms Composés du Français

**Déterminants Composés**

La deuxième catégorie proposée par G. Gross est celle des déterminants composés. La détermination comprend un ensemble de moyens morphologiques dont le rôle est d'actualiser les substantifs. Dans ce sens, la détermination est un élément qui actualise non seulement les noms mais également la phrase toute entière. La vision de G. Gross sur les déterminants composés est somme toute révolutionnaire dans le sens où il met en avant un certain nombre de phénomènes qui ne sont pas retenus comme pertinents dans la linguistique classique. Par exemple, il défend que la détermination composée d'un substantif peut être constituée d'un prédéterminant et d'un modifieur. Dans ce cas, le modi-

| ME | Fréq. | Noms composés |
|---|---|---|
| 0.000148591 | 102 | Parlamento Europeu |
| 0.000186936 | 73 | formação profissional |
| 0.000150955 | 37 | reforma da PAC |
| 0.000103408 | 32 | fundos estruturais |
| 0.000144964 | 29 | Comissão dos Assuntos Económicos e Monetários |
| 5.5563e-05 | 17 | construção naval |
| 4.49888e-05 | 9 | Nações Unidas |
| 1.95869e-05 | 8 | Subcomissão das Pescas |
| 2.99925e-05 | 6 | Comissão das Liberdades Públicas |
| 1.9995e-05 | 4 | Comissão dos Transportes e do Turismo |
| 1.38427e-05 | 3 | o caminho a seguir |
| 1.4059e-05 | 3 | mercado dos produtos da pesca |
| 1.49963e-05 | 3 | Sir Jack Stewart-Clark |
| 1.49963e-05 | 3 | bodes expiatórios |
| 8.17977e-06 | 3 | turbulências monetárias |
| 7.998e-06 | 2 | Extremo Oriente |
| 7.49813e-06 | 2 | Fundo Social Europeu |
| 9.9975e-06 | 2 | Conselho de Segurança da ONU |
| 7.49813e-06 | 2 | Conferência do Rio |
| 3.15711e-06 | 2 | carne de caça |

TAB. 8.2 – Noms Composés du Portugais

fieur peut être un adjectif, un complément de nom ou bien une proposition relative et le prédéterminant un article défini ou indéfini. Ainsi, dans les phrases suivantes, extraites de [51], les déterminants composés sont repérés en caractère gras et la séquence "* ?" identifie les phrases incorrectes.

*Luc a* **un large** *front*

** ? Luc a un front*

*Luc m'a répondu d'***une** *manière* **impolie**

** ?Luc m'a répondu d'une manière*

Afin de classer l'ensemble des déterminants composés, G. Gross propose ainsi plusieurs catégories : les combinaisons de plusieurs déterminants simples telles que *tous les*, *ces trois*, les associations prédéterminants et modifieurs ainsi que les déterminants nominaux tels que *un ensemble de*, *un tas de*. Nous présentons, dans les tableaux 8.3 et 8.4, une liste de déterminants composés qui ont été repérés par notre extracteur.

| ME | Fréq. | Déterminants composés |
|---|---|---|
| 2.52246e-05 | 118 | les ____ qui |
| 1.89204e-05 | 86 | toutes les |
| 3.80934e-06 | 55 | le ____ ____ que |
| 7.38931e-06 | 22 | ces deux |
| 2.81789e-06 | 21 | une autre |
| 2.16156e-06 | 17 | un nouveau |
| 1.6e-05 | 4 | l'un ou l'autre |
| 9.47368e-06 | 3 | un niveau de ____ qui |
| 6.42857e-06 | 2 | une partie ____ ____ de |
| 8.88889e-06 | 2 | une grande partie de |

TAB. 8.3 – Déterminants Composés du Français

La première nouveauté est l'apparition d'unités complexes non contiguës. En effet, comme le montrent les exemples précédents, un déterminant composé n'est pas forcément une séquence continue de formes graphiques. Ainsi, dans le deuxième exemple proposé par G. Gross, l'unité lexicale complexe devrait être représentée de la forme suivante : *une* ____ *impolie*. Les déterminants composés sont souvent difficiles à repérer et, au premier abord, leur structure paraît dénuée de sens. Ainsi, afin d'éclairer le lecteur sur la validité des résultats présentés, nous utiliserons un concordanceur qui permettra de replacer l'unité complexe dans son contexte. L'un des patrons récurrents les plus importants est sans aucun doute l'exemple type de la relation entre prédéterminant et modifieur. Ainsi, tant pour le Portugais que pour le Français, les unités complexes *les* ____ *qui* et *o* ____ ____ *que* démontrent la forte relation qui lie entre eux article défini et pronom relatif. Dans ce cadre, G. Gross défend que le prédéterminant — en l'occurrence l'article défini — et le modifieur — la proposition relative commencée par le pronom relatif — constituent une et une seule unité. Ainsi, dans les deux exemples suivants proposés dans [51], deux

| ME | Fréq. | Déterminants composés |
|---|---|---|
| 0.000132472 | 158 | todos os |
| 1.1595e-05 | 108 | o ____ ____ que |
| 4.56703e-06 | 29 | uma maior |
| 2.74468e-06 | 20 | um novo |
| 5.67974e-05 | 14 | um certo número de |
| 2.08281e-05 | 5 | uma ____ desse tipo |
| 5.9985e-06 | 2 | uma ou outra |
| 7.93919e-06 | 3 | uma ____ bem clara |
| 1.12472e-05 | 3 | uma tão grande |
| 5.71286e-06 | 2 | um elevado grau de |
| 8.88667e-06 | 2 | uma grande parte da |

TAB. 8.4 – Déterminants Composés du Portugais

déterminants composés sont mis en évidence en caractères gras.

*Donne-moi* **le** *stylo* **que je viens de t'offrir**

*Ramasse* **le** *stylo* **qui est par terre**

Si G. Gross s'intéresse plus particulièrement à l'ensemble de la proposition relative comme étant le modifieur par excellence, notre extracteur a régulièrement repéré la liaison entre l'article défini et le pronom relatif. Il semble donc que le déterminant composé puisse être réduit à cette simple attraction. Cette conclusion est somme toute logique. En effet, la première évidence de la relation entre le prédéterminant et le modifieur est l'association entre ses éléments initiaux.

Un certain nombre de déterminants composés ont été extraits. Parmi ceux-ci, certains sont clairement repérés comme tels et par conséquent ne nécessitent pas d'explication. D'autres cependant sont plus opaques, et l'utilisation d'un concordanceur s'est avérée indispensable à leur identification. C'est le cas en particulier des unités non contiguës. Par exemple, l'association *une partie* ____ ____ *de* n'est pas des plus claires. Cependant, son sens apparaît nettement dans un contexte plus large. Les résultats du concordanceur sont présentés dans le tableau 8.5.

| | | | | | |
|---|---|---|---|---|---|
| constituent | *une* | *partie* | non négligeable | *de* | la population |
| seulement | *une* | *partie* | très vulnérable | *de* | notre société |

TAB. 8.5 – Concordanceur pour [*0 une 1 partie 4 de*]

Cette situation se répète dans le cas du déterminant composé du Portugais *uma* ____ *bem clara*. En effet, le retrait du deuxième composant du $N$-gram impliquerait une phrase sans queue ni tête. Là encore, l'utilisation du concordanceur est indispensable à l'analyse linguistique — Tableau 8.6.

| | | | | | |
|---|---|---|---|---|---|
| tipicamente , | *uma* | amostra | *bem* | *clara* | da atitude |
| se dar | *uma* | definição | *bem* | *clara* | dos vocábulos |
| é disto | *uma* | prova | *bem* | *clara* | . Quando |

TAB. 8.6 – Concordanceur pour [*0 uma 2 bem 3 clara*]

Le couple *bem clara* est indissociable de l'article indéfini *uma* pour la bonne compréhension de chacune des phrases. Ainsi, la dernière phrase du tableau 8.6 deviendrait incompréhensible si on en éliminait son modifieur — en l'occurrence *bem clara*.

*é disto uma prova bem clara* (*est de ceci une preuve irréfutable*)

**?é disto uma prova* (**?est de ceci une preuve*)

Pareillement, certains déterminants composés contigus ne sont pas faciles à comprendre hors de leur contexte. Dans ce cadre, nous présentons un exemple pour le Français — Tableau 8.7 — et pour le Portugais — Tableau 8.8 — afin de familiariser le lecteur avec le type de phénomènes représentés par les déterminants composés.

| | | |
|---|---|---|
| par chance, | *l'un ou l'autre* | de ces faits |
| jour où | *l'un ou l'autre* | fonctionnaire |
| toujours bien | *l'un ou l'autre* | prétexte |
| visible de | *l'un ou l'autre* | système national |

TAB. 8.7 – Concordanceur pour *l'un ou l'autre*

Le nombre de déterminants composés extraits est largement inférieur au nombre de noms composés identifiés par notre architecture. Ceci était somme toute prévisible. Néanmoins,

| | | |
|---|---|---|
| Havendo | *uma tão grande* | distância |
| coloca | *uma tão grande* | ênfase na |
| formam | *uma tão grande* | parte do sector |

TAB. 8.8 – Concordanceur pour *uma tão grande*

il est intéressant de remarquer que ceux-ci ne représentent pas un phénomène marginal. En effet, un nombre non négligeable de déterminants composés a été extrait, en grande partie grâce à la puissance de représentation des modèles $N$-gram positionnels.

**Locutions Verbales**

Parmi les descriptions classiques, on peut citer celle de G. Gougenheim [108] qui propose un certain nombre de critères pour la reconnaissance des locutions verbales : absence d'article, verbe assez vide sémantiquement mettant en valeur le sens du nom, impossibilité d'une substitution sémantique portant sur le complément. Dans le but de définir schématiquement ce phénomène, G. Gross propose qu'*une suite verbe + compléments est une locution verbale si l'assemblage verbe-complément n'est pas compositionnel ou si les groupes nominaux sont figés (c'est-à-dire qu'on ne peut les modifier d'aucune manière : les déterminants sont fixes et les modifieurs interdits).*

De plus, G. Gross complète la liste proposée par G. Gougenheim et suggère un ensemble de paramètres de figement permettant d'identifier les locutions verbales : les compléments ne peuvent pas former de classes (*porter le chapeau, * ? porter le bonnet*), les compléments ne sont pas actualisés (*donner une gifle, * ? donner la gifle*), les transformations syntaxiques sont bloquées (*le bateau a pris l'eau, * ? l'eau a été prise par le bateau*).

Comme nous l'espérions, un nombre non négligeable de locutions verbales a été repéré. Nous en décrivons quelques unes dans les tableaux 8.9 et 8.10.

La première remarque importante qu'il convient de rehausser sur les résultats obtenus est le fait que de nombreuses locutions verbales ont été identifiées dans leur forme infinitive bien que le texte initial ne soit pas traité. Ceci renforce directement nos hypothèses initiales qui prônent que l'information contenue dans les textes est suffisante pour mettre

| ME | Fréq. | Locutions verbales |
|---|---|---|
| 0.000178304 | 58 | il y a |
| 2.97674e-05 | 16 | mettre en oeuvre |
| 1.6e-05 | 8 | faire face à |
| 1.38679e-05 | 7 | prendre une décision |
| 1.07143e-05 | 5 | prendre la parole |
| 8.27586e-06 | 4 | prendre en compte |
| 1.03846e-05 | 3 | mettre l'accent sur |
| 1.22727e-05 | 3 | mener à bien |
| 1.22727e-05 | 3 | faire pression sur |
| 1e-05 | 2 | perdre notre temps |
| 2.30769e-06 | 2 | entrer en vigueur |
| 3.15789e-06 | 2 | mettre à profit |
| 3.33333e-06 | 2 | apporter une solution |
| 8.57143e-06 | 2 | réduire à néant |
| 8.57143e-06 | 2 | passer sous silence |
| 6e-06 | 2 | traduire ____ ____ dans la pratique |
| 8.88889e-06 | 2 | prendre bonne note de |
| 1e-05 | 2 | rester en marge du développement |
| 8.57143e-06 | 2 | franchir le cap |
| 8.88889e-06 | 2 | se laver les mains |

TAB. 8.9 – Locutions verbales du Français

en évidence un grand nombre d'associations lexicales linguistiquement motivées[4]. La deuxième remarque qui s'impose est à mettre au crédit des modèles $N$-gram positionnels qui permettent l'extraction d'un nombre non négligeable de locutions verbales non contiguës tant pour le Français que pour le Portugais. Par exemple, la locution verbale *traduire* ____ ____ *dans la pratique* qui pourrait être représentée génériquement par la suite *traduire QUELQUE CHOSE dans la pratique*, n'aurait pu en aucun cas être extraite dans le cadre des modèles $N$-gram classiques — i.e. contigus. Ainsi, l'utilisation

[4]Il est évident, cependant, surtout pour le cas des locutions verbales, qu'une lemmatisation serait souhaitable comme nous le montrons dans [109]. En effet, les flexions verbales sont nombreuses pour le cas du Français et du Portugais et diminuent les décomptes de phénomènes "identiques".

| ME | Fréq. | Locutions verbales |
|---|---|---|
| 7.25819e-05 | 22 | chamar a atenção |
| 3.62813e-05 | 15 | ter em conta |
| 1.57104e-05 | 11 | pôr termo |
| 1.5996e-05 | 8 | tomar uma decisão |
| 8.5206e-06 | 5 | entrar em vigor |
| 8.56929e-06 | 4 | levar a sério |
| 8.88667e-06 | 4 | fazer o favor |
| 4.08989e-06 | 3 | tomar uma atitude |
| 4.35375e-06 | 3 | tomar em consideração |
| 7.49813e-06 | 3 | ter a certeza de |
| 2.49938e-06 | 2 | enfrentar ____ ____ problema |
| 6.665e-06 | 2 | colocar ____ ____ pergunta |
| 5.45318e-06 | 2 | lavar as mãos |
| 5.45318e-06 | 2 | levar à prática |
| 1.42821e-06 | 2 | dar a palavra |
| 5.9985e-06 | 2 | tomar posição sobre |
| 7.49813e-06 | 2 | tomar como exemplo |
| 5.9985e-06 | 2 | ler e escrever |
| 9.9975e-06 | 2 | tomar ____ iniciativa no sentido de |
| 9.9975e-06 | 2 | tirar conclusões para o futuro |

TAB. 8.10 – Locutions verbales du Portugais

des positions nous permet de vérifier que seulement deux mots séparent normalement les deux blocs de l'unité lexicale complexe. Ils sont illustrés dans le tableau 8.11 qui montre les résultats du concordanceur.

Comme nous l'avons déjà mentionné, ce phénomène n'est pas exclusif au Français mais s'étend également au Portugais. Dans ce cadre, nous illustrons l'expression verbale complexe *colocar* ____ ____ *pergunta* (*poser une question*) dans son contexte à partir du tableau 8.12. La première phrase du tableau 8.12 représente ainsi la variante *poser une question de plus* alors que la seconde met en évidence la suite *poser ma question*. Dans ces deux cas, la locution verbale se caractérise par la force qui unit les mots *colocar* (*poser*) et

| | | | | | |
|---|---|---|---|---|---|
| Et | *traduire* | ces | mesures | *dans la pratique* | à l'issue |
| manière de | *traduire* | ce | principe | *dans la pratique* | communautaire |

TAB. 8.11 – Concordanceur pour [*0 traduire 3 dans 4 la 5 pratique*]

*pergunta* (*question*) et non pas par les particules qui varient en son sein. On remarquera de plus le caractère clairement opaque du sens de la locution. En effet, dans la réalité, on ne "pose" pas une question dans le même sens que l'on pose un livre sur une table.

| | | | | | |
|---|---|---|---|---|---|
| tencionou | *colocar* | mais | uma | *pergunta* | à Comissão |
| gostaria de | *colocar* | a | minha | *pergunta* | ao contrário |

TAB. 8.12 – Concordanceur pour [*0 colocar 3 pergunta*]

Finalement, il est intéressant de noter que les fréquences des locutions verbales extraites sont particulièrement faibles. Un bon nombre d'entre elles n'apparaissent que deux fois dans le corpus mais sont néanmoins identifiées par notre extracteur. Ces résultats mettent en évidence les lacunes des méthodes d'extraction par valeurs seuil qui préconisent, comme dans [31], que le seuil de fréquence doit être particulièrement élevé pour permettre des taux de précision intéressants. Nous reviendrons sur ces résultats dans la prochaine partie de notre rapport.

**Locutions adjectivales**

Dans le cadre des locutions adjectivales, G. Gross propose une nouvelle définition de la notion d'adjectif qualificatif. Ainsi, il suggère une définition syntaxique de cette catégorie. Les adjectifs sont des formes (simples ou composées) qui correspondent aux deux critères suivants : elles figurent en position d'attribut à droite du verbe *être* et elles peuvent être nominalisées par le pronom invariable *le*. Par exemple, dans les deux phrases suivantes, *gentil* montre les caractéristiques d'un adjectif.

*Cet enfant a été gentil aujourd'hui.*

*Il le sera aussi demain.*

Cette définition vaut également pour les suites de nature polylexicale. Ainsi, la séquence *de bonne humeur* démontre le comportement d'un adjectif.

*Cet enfant a été de bonne humeur aujourd'hui.*
*Il le sera aussi demain.*

Dans cette optique, G. Gross [110] a proposé un recensement méthodique des locutions adjectivales qui nous a servi de base pour l'analyse de nos résultats d'extraction que nous illustrons dans les deux tableaux 8.13 et 8.14.

| ME | Fréq. | Locutions adjectivales |
|---|---|---|
| 2.32258e-05 | 12 | haute définition |
| 1.09051e-05 | 4 | à long terme |
| 3.75e-06 | 3 | d'extrême droite |
| 3.52941e-06 | 2 | en cours d'élaboration |
| 1.33333e-06 | 2 | dans l'attente de |
| 6.66667e-06 | 2 | dans l'espoir que |
| 2.85714e-06 | 2 | à votre disposition |
| 1.0e-05 | 2 | d'ordre technique |
| 3.15789e-06 | 2 | économiques et sociales |
| 2.6087e-06 | 2 | de ma circonscription |

TAB. 8.13 – Locutions adjectivales du Français

Il est évident que la typologie fournie par G. Gross s'est révélée particulièrement utile pour le repérage des locutions adjectivales du Français. Par contre, le passage au Portugais s'est avéré difficile. En effet, un nombre important de règles ne peut pas être traduit directement, rendant ainsi difficile l'identification des adjectifs composés du Portugais. De plus, il semble que la langue Portugaise n'utilise pas le concept de locution adjectivale dans les mêmes proportions que le Français. En effet, une étude exhaustive des données a été nécessaire pour repérer un ensemble fourni d'unités pertinentes.

### Locutions adverbiales

Parmi l'ensemble des locutions adverbiales, G. Gross distingue deux grandes catégories : les adverbes complexes qui sont figés et ceux qui ne le sont pas. Dans le

| ME | Fréq. | Locutions adjectivales |
|---|---|---|
| 4.54432e-05 | 10 | regionais e locais |
| 2.99925e-05 | 6 | em linha de conta |
| 1.13608e-05 | 5 | económico e social |
| 1.12472e-05 | 3 | de boa fé |
| 1.22697e-05 | 3 | fora de questão |
| 1.28539e-05 | 3 | ao abrigo do artigo |
| 1.9995e-06 | 2 | de natureza geral |
| 3.52853e-06 | 2 | económica e monetária |
| 3.74906e-06 | 2 | igual para todos |
| 6.665e-06 | 2 | de extrema- direita |

TAB. 8.14 – Locutions adjectivales du Portugais

premier cas, il met en évidence certains indices permettant de les repérer. Une locution adverbiale peut ainsi apparaître comme non compositionnelle quand l'un de ses éléments n'est pas reconnu comme un mot de la langue. Un bon exemple est l'adverbe complexe *grosso modo*. Le figement peut également provenir de l'emprunt métaphorique. Ainsi, *blanc comme neige* ou *fort comme un boeuf* sont des suites particulièrement figées. A l'opposé, certaines locutions adverbiales sont des constructions régulières et libres, même s'il existe pour chacune d'elles des restrictions lexicales. Dans ce cadre, nous retiendrons les catégories suivantes : les équivalents des adverbes en *-ment* comme *de façon* ou *de manière* et les structures productives telles que *PREP (DET) N*[5] qui constituent des moules de formation des adverbes complexes. Comme nous l'espérions, un grand nombre de locutions adverbiales a été extrait. Nous en illustrons quelques exemples dans les deux tableaux suivants 8.15 et 8.16.

La première remarque qu'il convient de rehausser tient au fait que la virgule joue un rôle important dans la structure des adverbes complexes. En particulier, les adverbes de marque de ponctuation du discours tels que *en effet*, *d'autre part* ou *du reste* sont régulièrement suivis d'une virgule. Cette caractéristique est largement mise en évidence par notre extracteur qui présuppose que la virgule serait partie intégrante de la locution

[5]Nous rappelons que PREP est l'étiquette morpho-syntaxique de la préposition, DET celle du déterminant et N celle du nom.

| ME | Fréq. | Locutions adverbiales |
|---|---|---|
| 9e-05 | 60 | En effet, |
| 0.000233103 | 52 | tout à fait |
| 1.97388e-05 | 23 | sans doute |
| 1.9542e-05 | 16 | jusqu'à présent |
| 4.5283e-06 | 12 | bien entendu |
| 5.30702e-06 | 11 | sans cesse |
| 1.98361e-05 | 11 | D'autre part, |
| 1.48171e-05 | 9 | à cet égard |
| 3.42857e-05 | 8 | en même temps |
| 2.94e-05 | 7 | en premier lieu |
| 9.24528e-06 | 7 | Tout d'abord, |
| 2.53448e-05 | 7 | à juste titre |
| 9.61538e-06 | 5 | Du reste, |
| 1.70455e-05 | 5 | en temps voulu |
| 1.125e-05 | 3 | sans nul doute |
| 1.22727e-05 | 3 | Pour ma part, |
| 1.5e-05 | 3 | de moins en moins |
| 1e-05 | 2 | autant que faire se peut |
| 1.875e-06 | 2 | de tout coeur |
| 1e-05 | 2 | à n'en point douter |

TAB. 8.15 – Locutions adverbiales du Français

adverbiale. On remarquera même que la virgule peut servir de délimiteur de la suite complexe comme dans *muito simplesmente* (*très simplement*) ou *no entanto* (*cependant*). Devant ce cadre, il est important de louer l'utilisation de toute l'information contenue dans les textes et de ne pas se laisser tenter par la solution facile qui consisterait à effacer du texte toutes les particules "soit disant" dénuées de sens ou d'intérêt.

La deuxième remarque digne d'être mise en valeur est encore une fois à mettre au crédit de l'utilisation des modèles $N$-gram positionnels qui permettent de rendre compte de phénomènes non continus. Ceci est d'autant plus intéressant que la locution adverbiale la plus fréquente du Portugais est justement discontinue : *"Em ___ lugar ,"*. En

| ME | Fréq. | Locutions adverbiales |
|---|---|---|
| 0.000443415 | 106 | Em ____ lugar , |
| 0.000244007 | 60 | , por exemplo , |
| 0.000215864 | 46 | , ou seja , |
| 3.37416e-05 | 27 | sem dúvida |
| 8.44789e-05 | 26 | cada vez mais |
| 0.00010478 | 26 | , no entanto , |
| 8.59309e-05 | 19 | ao mesmo tempo |
| 3.83904e-05 | 16 | antes de mais |
| 1.78527e-05 | 5 | em meu entender |
| 6.85543e-05 | 16 | a meu ver |
| 1.45418e-05 | 4 | em todo o caso |
| 1.49963e-05 | 4 | em larga medida |
| 1.34966e-05 | 3 | ponto por ponto |
| 3.74906e-06 | 3 | apesar de tudo |
| 8.99775e-06 | 3 | a seu tempo |
| 5.45318e-06 | 2 | tanto mais ____ quanto |
| 4.61423e-06 | 2 | muito em breve |
| 8.56929e-06 | 2 | no bom sentido |
| 5.71286e-06 | 2 | , muito simplesmente , |
| 9.9975e-06 | 2 | sem sombra de dúvida |

TAB. 8.16 – Locutions adverbiales du Portugais

effet, l'élément variable contribue à l'identification d'un nombre important de variantes du patron général. Ainsi, il est possible d'extraire des suites telles que "*Em primeiro lugar ,*" ("*En premier lieu ,*") ou encore "*Em segundo lugar ,*" ("*En deuxième lieu ,*"). Nous illustrons cette situation à partir des résultats du concordanceur dans le tableau 8.17.

Finalement, il semblerait[6] que les locutions adverbiales soient plus nombreuses en Français qu'en Portugais. En effet, alors que nous avons dû choisir nos exemples parmi un vaste ensemble de locutions potentielles pour le Français, cette situation ne s'est pas répétée pour le Portugais pour lequel il a été difficile d'identifier différentes catégories d'adverbes

[6]Cette hypothèse mériterait d'être confirmée.

| | | | | |
|---|---|---|---|---|
| Tecnologia . | *Em* | primeiro | *lugar ,* | verifico que |
| mercado . | *Em* | segundo | *lugar ,* | 10 a 15% |
| higiene . | *Em* | terceiro | *lugar ,* | depois de |
| países ? | *Em* | quarto | *lugar ,* | considera que |
| pesca . | *Em* | quinto | *lugar ,* | é preciso |

TAB. 8.17 – Exemples du Concordanceur pour [*0 Em 2 lugar 3 ,*]

complexes.

**Locutions prépositives et conjonctives**

G. Gross propose de traiter les locutions prépositives et les locutions conjonctives de la même façon. En effet, il suggère que *[...] leur fonctionnement est parallèle.* Cependant, pour la clarté de nos propos, nous structurerons notre analyse en deux parties bien distinctes : l'une pour les prépositions complexes, l'autre pour les conjonctions polylexicales.

**Locutions prépositives :** On attribue généralement aux prépositions la fonction d'introduire un complément — indirect — après un prédicat, que celui-ci soit verbal, nominal ou adjectival. G. Gross se réfère ainsi au rôle d'indicateur d'argument. Cependant, certaines prépositions peuvent avoir des emplois prédicatifs, c'est-à-dire avoir des arguments. C'est le cas des prépositions locatives ou temporelles telles que *à côté de* ou *au terme de* pour lesquelles le verbe joue généralement le rôle de support comme dans la phrase *la mairie est à côté de l'église* qui pourrait être représentée par le prédicat du premier ordre *à_côté(mairie,église).* Dans ce cadre, nous présentons un ensemble de locutions prépositives qui ont été repérées dans notre expérience — Tableaux 8.18 et 8.19.

Il importe de noter que le genre du corpus utilisé a particulièrement avantagé le processus d'acquisition. En effet, le caractère formel des textes de la Commission Européenne est propice à l'occurrence de prépositions complexes telles que *en matière de* ou *em nome da* (*au nom de la*). Ainsi, un bon nombre de locutions prépositives a été repéré dans les deux langues avec un nombre important d'occurrences.

Une remarque importante doit cependant être formulée sur la nature des locutions prépositives extraites pour le Portugais. En effet, dans *em nome da* et *em benefício dos*, les

| ME | Fréq. | Locutions prépositives |
|---|---|---|
| 0.000154341 | 80 | en matière de |
| 2.28863e-05 | 59 | au cours |
| 0.000157091 | 48 | en tant que |
| 1.875e-05 | 5 | sans pour autant |
| 8e-06 | 4 | par le biais de |
| 1.28e-05 | 4 | en matière de ____ et de |
| 1.125e-05 | 3 | Eu égard à |
| 1.15385e-06 | 2 | Au lieu de |
| 2.4e-06 | 2 | A cause de |
| 2.4e-06 | 2 | vu l'absence de |

TAB. 8.18 – Locutions prépositives du Français

prépositions *da* et *dos* correspondent respectivement aux deux contractions $PREP+ART$ suivantes : *de + a* et *de + os*. Il serait donc légitime de ne considérer comme véritables unités polylexicales que les suites *em nome de* et *em benefício de*. Cette situation se répète pour les séquences *com vista à* et *no tocante à* où *à* correspond à la contraction $PREP + ART$ *a + a*.

**Locutions conjonctives :** Les grammaires scolaires analysent la phrase complexe comme constituée d'une principale et d'une subordonnée reliées par une conjonction ou une locution conjonctive. Cependant, cette vision est fortement réductrice et ne rend pas compte de l'ensemble des comportements des conjonctions complexes. En effet, parallèlement aux prépositions, les conjonctions ont soit une fonction prédicative comme par exemple *à telle enseigne que* ou bien introduisent des arguments comme dans *le fait que*. Aussi, nous utiliserons les mêmes indices qui ont été définis pour les locutions prépositives pour repérer les conjonctions complexes. Les résultats sont illustrés dans les deux tableaux suivants 8.20 et 8.21.

Alors qu'il a été facile d'identifier un nombre non négligeable de locutions conjonctives pour le Français, cette tâche s'est révélée particulièrement difficile pour le Portugais. Il a même relevé du miracle pour présenter les dix exemples du tableau 8.21. Cette caractéristique tient du fait que les locutions conjonctives sont peu utilisées en Portugais

| ME | Fréq. | Locutions prépositives |
|---|---|---|
| 0.000353528 | 89 | em nome da |
| 0.000206845 | 60 | de acordo com |
| 0.000152504 | 60 | em matéria de |
| 0.000132084 | 52 | no sentido de |
| 0.000107732 | 25 | com vista à |
| 1.05856e-05 | 6 | por meio de |
| 2.322e-05 | 6 | sob a forma de |
| 4.99875e-06 | 3 | em benefício dos |
| 2.06845e-06 | 2 | numa lógica de |
| 5.45318e-06 | 2 | no tocante à |

TAB. 8.19 – Locutions prépositives du Portugais

courant mettant plus aisément en évidence les conjonctions simples.

Afin de résumer cette première analyse, il est important de souligner que tous les phénomènes linguistiques catalogués par G. Gross ont été repérés par notre architecture. Et ceci tant pour le Français que pour le Portugais. Ces résultats sont particulièrement réconfortants dans le sens où ils mettent en évidence la possibilité d'extraire un nombre important d'associations lexicales à partir de la seule information contenue dans les textes et ceci indépendamment de la langue considérée. Parallèlement, la puissance de représentation des modèles $N$-gram positionnels s'est révélée être un atout fondamental pour l'identification de suites discontinues qui comme nous l'avons montré ne forment pas un phénomène marginal. Seulement, réduire l'analyse des résultats d'extraction à une "simple" étude des catégories proposées par G. Gross ne serait en aucun cas faire justice à la diversité des phénomènes qui ont été identifiés. Parmi ceux-ci, on trouve les associations lexicales obtenues à partir d'opérations telles que la surcomposition, la modification et la coordination qui ont été mises en évidence par B. Daille [21]. Nous en proposons donc une étude approfondie dans le prochain paragraphe.

### 8.1.2 Classification de B. Daille

Dans la partie précédente, nous nous sommes attachés à rendre compte de l'ensemble des phénomènes linguistiques extraits par notre architecture sur la base de la classification

| ME | Fréq. | Locutions conjonctives |
|---|---|---|
| 1.39165e-05 | 70 | parce que |
| 2.36293e-05 | 92 | ainsi que |
| 1.22727e-05 | 9 | pour dire que |
| 2.04255e-05 | 8 | dans quelle mesure |
| 1.6875e-05 | 6 | à quel point |
| 2.16e-05 | 6 | chaque fois que |
| 6.66667e-06 | 2 | dû au fait que |
| 2.6087e-06 | 2 | à commencer par |
| 6.66667e-06 | 2 | en ce sens que |
| 7.27273e-06 | 2 | ce qui signifie que |

TAB. 8.20 – Locutions conjonctives du Français

de G. Gross. Cependant, cette dernière, bien que remarquable en tout point, n'aborde pas la structure interne des unités lexicales complexes. En fait, peu d'études se sont penchées sur ce problème. L'une d'entre elles a été proposée par B. Daille [21] dans le cadre de la construction automatique de terminologies. Dans ce contexte, B. Daille met en évidence l'existence de trois opérations qui permettent la réalisation de termes complexes de taille supérieure à trois unités lexicales. Celles-ci sont la surcomposition, la modification et la coordination. Ainsi, B. Daille prétend différencier les termes de base — qui contiennent au plus deux mots pleins — de toutes les autres associations selon leurs structures internes.

On regrettera cependant que cette étude ne s'étende pas aux associations lexicales autres que les constructions nominales complexes — i.e. les noms composés. En effet, B. Daille restreint ainsi le champ d'application de son étude. Nous verrons cependant que les trois opérations de construction qu'elle propose couvrent un ensemble important de phénomènes linguistiques. Notre objectif sera donc dans un premier temps de montrer que notre architecture est capable de repérer des noms composés obtenus par surcomposition, modification ou coordination. Dans un deuxième temps, nous tenterons d'étendre ces opérations à un ensemble plus large d'associations lexicales.

Avant de commencer notre étude, il est important de noter que les travaux de B. Daille définissent une solution théorique qui n'a pas fait l'objet d'une proposition computation-

| ME | Fréq. | Locutions conjonctives |
|---|---|---|
| 2.35524e-05 | 16 | assim como |
| 5.37643e-05 | 11 | na medida em que |
| 1.54247e-05 | 6 | do modo como |
| 2.666e-05 | 6 | ao passo que |
| 9.14057e-06 | 4 | de modo a que |
| 1.12472e-05 | 3 | à medida que |
| 8.56929e-06 | 3 | num momento em que |
| 9.08864e-06 | 2 | à semelhança do que |
| 3.3325e-06 | 2 | como de resto |
| 7.998e-06 | 2 | de cada vez que |

TAB. 8.21 – Locutions conjonctives du Portugais

nelle. Grâce aux caractéristiques propres du GenLocalMaxs, nous espérons répondre de forme élégante à cette lacune. En particulier, nous mettrons en évidence la versatilité de notre architecture autour du fait qu'elle ne recourt à aucun effort computationnel supplémentaire pour extraire des unités lexicales de plus de deux mots pleins, contrairement à ce qui a été proposé par un certain nombre d'auteurs [31] [32] et qui est encore la plus importante.

**Surcomposition**

Dans son étude, B. Daille définit un terme de base comme étant une construction motivée ayant l'une des structures syntaxiques suivantes où $N$ est l'étiquette morpho-syntaxique du nom, $ADJ$ celle de l'adjectif, $DET$ celle du déterminant et $PREP$ celle de la préposition :

- $N\ ADJ$
- $N$ *de* $(DET)\ N$
- $N$ *à* $(DET)\ N$
- $N\ PREP\ N$
- $N\ N$

Ainsi, deux types de surcomposition peuvent être identifiés : la juxtaposition et la substitution. Dans le premier cas — juxtaposition —, toute séquence construite à partir d'un terme de base et dont la structure est inchangée peut être considérée comme un

terme juxtaposé. Par exemple, la séquence morpho-syntaxique $N_1$ $PREP_1$ [$N_2$ $PREP_2$ $N_3$] où [$N_2$ $PREP_2$ $N_3$] repère le terme de base est un terme juxtaposé. Dans le deuxième cas — substitution —, si l'un des éléments d'un terme de base est substitué par un autre terme de base dont la tête est identique, la structure résultante doit être considérée comme un terme obtenu par substitution. Ainsi, si $N_1$ est substitué par le terme de base [$N_1$ $PREP_2$ $N_3$] dans la structure $N_1$ $PREP_1$ $N_2$, la séquence [$N_1$ $PREP_2$ $N_3$] $PREP_1$ $N_2$ est un terme obtenu par substitution.

Il est important de noter que B. Daille s'aide de l'information morpho-syntaxique pour déterminer la structure des termes obtenus par surcomposition. Or, nous ne disposons pas de telle information. De plus, nos résultats d'extraction mettent en évidence l'identification de phénomènes linguistiques qui dépassent le cadre des noms composés. Par conséquent, nous simplifierons la définition donnée par B. Daille tout en en gardant l'esprit. Ainsi, nous considérerons que toute association lexicale construite à partir d'une structure également repérée par l'extracteur sera une association lexicale surcomposée[7]. Nous illustrons cette situation dans les deux tableaux suivants 8.22 et 8.23.

Comme le montrent les résultats des deux tableaux 8.22 et 8.23, les phénomènes linguistiques repérés dépassent le cadre des noms composés et mettent en évidence un ensemble d'opérations plus large que celui proposé par B. Daille. Parallèlement, l'opération de substitution n'a pas été identifiée dans ce corpus alors que la juxtaposition est effectivement présente dans les résultats tant pour le Français que pour le Portugais.

L'un des résultats intéressants est à mettre au crédit des noms composés construits sur la base d'au moins deux associations lexicales préalablement identifiées par l'extracteur. C'est le cas par exemple de la séquence *Conselho de Segurança das Nações Unidas* (*Conseil de Sécurité des Nations Unies*) qui est le résultat de la juxtaposition des deux noms composés *Conselho de Segurança* (*Conseil de Sécurité*) et *Nações Unidas* (*Nations Unies*).

[7]Nous ne considèrerons pas ici les séquences obtenues par coordination pour lesquelles cette définition s'applique. Cette caractéristique sera présentée dans le prochain paragraphe.

| Surcomposée | Composée |
|---|---|
| Etats membres de la Communauté | Etats membres |
| la Communauté économique européenne | la Communauté |
| ni plus ni moins | ni ____ ni |
| ne ____ pas en mesure de | ne ____ pas |
| les gouvernements des Etats membres | Etats membres |
| la Commission en tant que collège | la Commission |
| je tiens à féliciter le rapporteur | féliciter le rapporteur |
| notre collègue Valverde López | Valverde López |
| les douze Etats membres | Etats membres |
| la directive sur l'hygiène alimentaire | l'hygiène alimentaire |
| la capacité concurrentielle des entreprises | la capacité concurrentielle |
| la ____ entre les partenaires sociaux | partenaires sociaux |
| sécurité dans les transports aériens | transports aériens |
| le secteur des transports aériens | transports aériens |
| partis de la droite démocratique | partis de ____ droite |
| la ____ du président en exercice | président en exercice |
| le secteur des fruits et légumes | fruits et légumes |
| groupe des verts au Parlement européen | Parlement européen |
| le ____ de la téléphonie vocale | téléphonie vocale |

TAB. 8.22 – Surcomposition du Français

Parallèlement, la surcomposition n'est pas une opération spécifique aux associations lexicales de type nominal. Par exemple, dans l'unité complexe *je tiens à féliciter le rapporteur* qui se comporte comme une phrase idiomatique du discours de la Commission Européenne, l'expression verbale *féliciter le rapporteur* joue le rôle de constituant de la juxtaposition. Le même phénomène est visible pour le Portugais dans l'unité complexe *política comunitária em matéria de ambiente* (*politique communautaire en matière d'environnement*) où la locution prépositive *em matéria de* (*en matière de*) est la seule association lexicale préalablement identifiée.

| Surcomposée | Composée |
|---|---|
| senhor deputado De Piccoli | De Piccoli<br>senhor deputado |
| acesso à formação profissional | formação profissional |
| em nome do Grupo Socialista | Grupo Socialista |
| Jornal Oficial das Comunidades Europeias | Jornal Oficial |
| Comissão da Agricultura do Parlamento | Comissão da Agricultura |
| a fixação ____ preços agrícolas | preços agrícolas |
| relatório anual do Tribunal de Contas | Tribunal de Contas |
| a importação de peixe fresco | peixe fresco |
| participação das PME nos contratos públicos | contratos públicos |
| um firme desiderato do Parlamento Europeu | Parlamento Europeu |
| Livro Branco sobre o mercado interno | Livro Branco<br>mercado interno |
| Conselho de Segurança das Nações Unidas | Nações Unidas<br>Conselho de Segurança |
| Conselho de Segurança da ONU | Conselho de Segurança |
| Grupo dos ____ no Parlamento Europeu | Parlamento Europeu |
| o mercado interno da Comunidade | mercado interno |
| extremismo de direita na Europa | extremismo de direita |
| Programa de Acção em ____ de Ambiente | Programa de Acção |
| uma política ____ de formação profissional | formação profissional |
| milhares de milhões de ecus | milhões de ecus |
| política comunitária em matéria de ambiente | em matéria de |

TAB. 8.23 – Surcomposition du Portugais

Finalement, l'utilisation des modèles $N$-gram positionnels met en évidence la réalisation de certains phénomènes linguistiques non contigus intéressants que nous détaillerons dans la suite de notre rapport. Par exemple, la locution adverbiale *ni plus ni moins* est construite sur la base du patron récurrent de la conjonction négative *ni* ____ *ni*. Parallèlement, certains noms composés juxtaposés peuvent être identifiés *a posteriori* à partir de l'identification de patrons non contigus, comme dans la séquence *la* ____ *du président en exercice*

où l'espace laissé libre peut être rempli par les substantifs *réponse* ou *déclaration*. De même pour le Portugais, la surcomposée *uma política ____ de formação profissional* (*une politique ____ de formation professionnelle*), rend compte des deux noms composés *uma política racional de formação profissional* (*une politique rationnelle de formation professionnelle*) et *uma política comum de formação profissional* (*une politique commune de formation professionnelle*). L'un des intérêts majeurs de cette caractérisation des associations lexicales est de pouvoir repérer des unités polylexicales qui n'apparaissent qu'une seule fois dans les textes. Ainsi, le grand problème des systèmes statistiques qui se doivent de prendre en compte uniquement les $N$-grams dont la fréquence est au moins égale à deux, est attaqué de façon pertinente. En effet, grâce aux modèles $N$-gram positionnels, il est possible d'identifier des patrons de fréquence élevée dont les réalisations complètes peuvent mettre en évidence des associations lexicales de fréquence unitaire. Même si cette hypothèse n'est pas toujours valide, elle laisse entrevoir une solution prometteuse.

**Modification**

Au-delà de la réalisation de noms composés à partir des opérations de juxtaposition et substitution, B. Daille met en évidence la construction d'associations lexicales de type nominal selon le principe de modification. Ainsi, un nouveau terme peut être construit à partir d'un terme de base par l'insertion d'un modificateur soit à l'intérieur soit après celui-ci. Ainsi, les adjectifs et les adverbes sont généralement des modificateurs qui sont susceptibles d'être insérés à l'intérieur des termes de base. Par exemple, on parlera indifféremment de *réseaux entièrement numériques* ou de *réseaux cablés numériques*. Dans ce cadre, l'élément modificateur suggère un certain degré de flexibilité de l'unité polylexicale que notre extracteur devrait être en mesure d'identifier grâce à la représentation en $N$-grams positionnels. En effet, à partir de l'exemple précédent, il est clair que l'unité lexicale complexe devrait être représentée par le patron *réseaux ____ numériques*. Nous illustrons certains de ces phénomènes dans les deux tableaux suivants 8.24 et 8.25.

Encore une fois, les résultats observés démontrent une large variété de phénomènes qui s'étend bien au-delà de la simple insertion d'adverbes ou d'adjectifs. En effet, de nombreux exemples montrent que l'élément modificateur peut être un nom et l'adjectif l'élément fixe de l'association lexicale comme dans la suite *prendre des ____ spécifiques* où les deux substantifs possibles sont *engagements* et *mesures*. Parallèlement, un bon

| Modification | Modificateur |
|---|---|
| la ____ guerre mondiale | Première<br>Deuxième |
| le ____ grand intérêt | plus<br>très |
| les ____ points de vue exprimés | divers<br>principaux |
| partis de ____ droite | la<br>l'extrême |
| prendre des ____ spécifiques | mesures<br>engagements |
| n'est ____ ____ possible de | certes pas<br>particulièrement plus |
| augmentation ____ des quotas | temporaire<br>éventuelle |
| l'article ____ de la directive | 4<br>5 |
| il est ____ important de | particulièrement<br>tellement |
| vieillissement ____ de la population active | progressif<br>relatif |

TAB. 8.24 – Modification du Français

nombre d'unités lexicales complexes obtenues par modification suggèrent l'introduction de différentes particules pour la réalisation du même phénomène linguistique. Ainsi, la séquence *associação* ____ *regiões marítimas* (*association* ____ *régions maritimes*) peut être complètement réalisée par l'introduction de deux prépositions possibles : *de* (*de*) ou *das* (*des*).

Dans le cadre des opérations de modification, B. Daille propose une seconde construction possible. Elle la nomme post-modification. Suivant ce nouvel opérateur, un terme de base peut être modifié par l'ajout d'un adjectif ou d'un adverbe à sa suite. Ainsi, une *station terrienne* peut être *brouilleuse* et donner naissance au terme complexe *station terrienne*

| Modification | Modificateur |
|---|---|
| o pacote ____ preços | de<br>dos |
| nos últimos ____ anos | dois<br>cinco<br>dez |
| as regiões ____ da Comunidade | costeiras<br>desfavorecidas<br>periféricas |
| no mais ____ prazo | breve<br>curto |
| execução das dotações ____ ____ investigação | destinadas à<br>para a |
| associação ____ regiões marítimas | de<br>das |
| política ____ de transportes | comunitária<br>comum |
| todo ____ conjunto de | o<br>um |
| adopção de ____ comunitárias | medidas<br>normas |
| gostaria ____ de referir | ainda<br>precisamente |

TAB. 8.25 – Modification du Portugais

*brouilleuse*. Les caractéristiques de l'algorithme de sélection GenLocalMaxs ne permettent malheureusement pas l'extraction de telles associations. En effet, un $N$-gram positionnel est élu s'il constitue un maximum local. Dans ce cas, pour un $N$-gram positionnel élu, aucun de ses sous-groupes de rang $N-1$ ni aucun de ses sur-groupes de rang $N+1$ ne pourra être retenu comme pertinent. Par contre, le GenLocalMaxs est capable de proposer des marqueurs c'est-à-dire des associations lexicales incomplètes qui supposent l'existence de plusieurs modificateurs. Par exemple, *un travail très* suggère l'occurrence des deux adjectifs *difficile* et *fouillé*. Nous exemplifions ce phénomène dans 8.27 et 8.26.

Comme les résultats le montrent, deux grandes catégories de marqueurs ont été identifiées. Dans le premier cas, le modificateur vient se greffer en avant de l'unité fixe élue. Par exemple, la séquence *de substances radioactives* peut être précédée des formes graphiques suivantes pour réaliser un nom composé : *transferts* ou *mouvements incontrôlés*. Dans le même contexte, en Portugais, la suite *de alta velocidade* (*à haut débit*) est à l'origine des deux associations nominales suivantes : *linhas de alta velocidade* (*lignes à haut débit*) et *redes de alta velocidade* (*réseaux à haut débit*).

| Marqueur | Modificateur |
|---|---|
| da televisão digital | êxito |
| | introdução |
| | desenvolvimento |
| de impacto ambiental | aspectos |
| | estudos |
| da taxa extraordinária | quitação |
| | pagamento |
| proibição de experimentações | de cosméticos |
| | em animais |
| dos Negócios Estrangeiros | ministro |
| | ministros |
| uma redução das | metas |
| | tarifas |
| política europeia de | desenvolvimento |
| | subsídios |
| elevado nível de | habilitações |
| | protecção |
| de alta velocidade | linhas |
| | rede |
| direitos fundamentais do | povo |
| | consumidor |

TAB. 8.26 – Marqueurs du Portugais

| Marqueur | Modificateur |
|---|---|
| économique et social | Comité |
| | plan |
| libre circulation des | personnes |
| | travailleurs |
| de service public | notion |
| | L'idée |
| la création d'une | agence |
| | Agence |
| | brigade |
| prêts consentis par | la Banque |
| | la CECA |
| de substances radioactives | transferts |
| | mouvements incontrôlés |
| de petites morues | stocks |
| | quota |
| | quotas |
| faible niveau de | formation |
| | technologie |
| Les pêcheurs de | thon |
| | mon pays |
| marché européen des | télécommuncations |
| | qualifications |

TAB. 8.27 – Marqueurs du Français

Dans le deuxième cas, l'élément modificateur se trouve en queue de l'unité lexicale retenue. Ainsi, le marqueur *marché européen des* suggère la réalisation des deux noms composés *marché européen des qualifications* et *marché européen des télécommunications.* Pareillement, en Portugais, la séquence *direitos fundamentais do* (*droits fondamentaux du*) met en évidence les deux unités lexicales complexes de type nominal *direitos fundamentais do povo* (*droits fondamentaux du peuple*) et *direitos fundamentais do consumidor* (*droits fondamentaux du consommateur*).

Le lecteur averti aura cependant remarqué que les exemples que nous avons fournis dévient de la définition originale donnée par B. Daille[8]. En effet, un bon nombre d'entre eux, plutôt que d'illustrer l'opération de post-modification, démontrent l'incapacité de notre extracteur à identifier de véritables associations lexicales. Par exemple, notre architecture devrait préférer le nom composé *petites morues* à l'association lexicale *de petites morues* réellement extraite. Ce phénomène est facilement explicable. En effet, chaque fois que la suite *petites morues* apparaît dans le corpus, celle-ci est précédée de la préposition *de* comme le montre le résultat du concordanceur — voir Tableau 8.28.

| | | | |
|---|---|---|---|
| stocks | de | *petites morues* | en |
| quotas | de | *petites morues* | en |
| quota | de | *petites morues* | fixé |

TAB. 8.28 – Concordanceur pour *petites morues*

Ainsi, notre extracteur préfère l'association lexicale la plus longue et élit la séquence *de petites morues*. Dans notre cas, c'est-à-dire sans recourir à d'autres informations qui ne sont pas originellement dans les textes, la seule possibilité pour identifier le nom composé *petites morues* plutôt que l'association lexicale *de petites morues* serait d'utiliser une quantité de textes plus importante qui permette de mettre en valeur le concept de *petites morues* dans un contexte lexical plus large. Nous frisons ici un point important des méthodes statistiques basées sur le matériel textuel cru. En effet, les limitations intrinsèques liées à l'utilisation de textes non traités ne permettent pas de dissocier les séquences de formes graphiques de leur contexte. L'usage des mots dépend ainsi du contexte dans lequel ils sont extraits. La seule solution pour éviter ces "effets de bord" serait d'augmenter la diversité du contexte lexical sans pour autant diminuer les régularités qui le caractérise. Cette solution est évidemment utopique car elle met en jeu deux phénomènes antagoniques. Ainsi, nous verrons dans la suite de ce rapport que l'introduction de connaissances linguistiques externes au texte peut être un atout ponctuellement nécessaire pour l'identification d'unités complexes de sens plein.

[8]Sans pour autant en perdre l'esprit.

**Coordination**

Finalement, B. Daille propose une dernière opération qui permet de construire un ensemble de termes complexes à partir de termes de base grâce à la coordination. Dans ce cadre, un terme complexe peut être construit à partir de la fusion par coordination de deux termes de base. Par exemple, si l'on considère les deux termes de base [$N_1$ $PREP_1$ $N_2$] et [$N_2$ $PREP_1$ $N_3$], un terme complexe obtenu par coordination peut prendre la forme suivante [$N_1$ et $N_2$ $PREP_1$ $N_3$]. Suivant cette définition, nous présentons un ensemble d'unités complexes extraites par notre architecture pour le Français et le Portugais dans les deux tableaux 8.29 et 8.30.

| EM | Fréq. | Coordination |
|---|---|---|
| 3.5e-05 | 7 | contre la pauvreté et l'exclusion |
| 1e-05 | 2 | budget rectificatif et supplémentaire |
| 1e-05 | 2 | attitudes racistes et xénophobes |
| 1e-05 | 2 | les uns et les autres |
| 1e-05 | 2 | reconversion écologique et sociale de |
| 1e-05 | 2 | centre d'information et de promotion régionale |
| 8e-06 | 2 | commission économique et monétaire |

TAB. 8.29 – Coordination du Français

Comme le met en évidence B. Daille [21], les termes complexes obtenus par coordination sont rares pour le Français. Ceci se confirme à la lumière des résultats exposés dans le tableau 8.29. En effet, les sept exemples présentés sont les seuls qu'il nous a été possible de repérer parmi l'ensemble de tous les $N$-grams élus. Au contraire, ce phénomène foisonne pour le Portugais. En effet, cette opération est régulièrement utilisée pour éviter la répétition du terme "tête" dans la suite du discours. Ainsi, il nous a été particulièrement facile d'identifier un nombre important d'exemples.

En résumé, après avoir illustré l'ensemble des résultats obtenus par notre architecture selon la classification de G. Gross, nous avons étudié les différentes structures internes des associations lexicales à partir des catégories proposées par B. Daille. Dans les deux cas, notre extracteur s'est montré à la hauteur de la tâche pour laquelle il a été conçu,

| EM | Fréq. | Coordination |
|---|---|---|
| 2.49938e-05 | 5 | Grupo Liberal , Democrático e Reformista |
| 1.49963e-05 | 3 | Europa Central e de Leste |
| 9.9975e-06 | 2 | estruturas de instrução e formação |
| 9.9975e-06 | 2 | por via marítima ou aérea |
| 9.9975e-06 | 2 | funções de acompanhamento e controlo |
| 9.9975e-06 | 2 | defensoras e defensores dos animais |
| 9.9975e-06 | 2 | as vias fluviais e marítimas |
| 9.9975e-06 | 2 | condições de vida e de trabalho |
| 9.9975e-06 | 2 | artigos 92o e 93o do Tratado |
| 6.665e-06 | 2 | formação profissional e contínua |

TAB. 8.30 – Coordination du Portugais

ceci tant pour le Français que pour le Portugais. Néanmoins, certaines lacunes ont pu être mises en évidence notamment en ce qui concerne l'utilisation exclusive de l'information explicitement présente dans les textes. Avant de revenir sur ces problèmes dans une partie ultérieure de notre rapport, nous proposons une étude de phénomènes particulièrement intéressants qui ont été repérés lors de l'analyse de résultats.

### 8.1.3 Autres Types de Patrons

Comme nous l'espérions, un nombre important de phénomènes linguistiques ont été extraits en dehors du cadre spécifique des deux classifications proposées par B. Daille et G. Gross. En effet, comme le souligne F. Smadja [31], un nombre non négligeable d'associations lexicales sont des expressions idiomatiques du domaine qu'il nomme de syntagmes patrons[9]. Comme nous le verrons dans cette partie, le discours de la Commission Européenne foisonne de ce type d'associations. Parallèlement, l'analyse des résultats met en évidence l'extraction de marqueurs syntaxiques tels que la conjonction et la négation. Nous aborderons notre illustration par les deux derniers phénomènes.

#### Adverbes de Négation

L'une des particularités de la négation du Français est d'être un phénomène non contigu par excellence. En effet, sa réalisation, dans le cas de son utilisation non

[9]Traduction de l'anglais *Phrasal Templates*.

prédicative, se fait généralement par la conjugaison de la particule *ne* et d'un adverbe obtenu par la grammaticalisation de substantifs comme dans *ne ... pas* ou *ne ... point*. La particule peut également être suivie d'un forclusif qui donne un sens particulier à la négation. Par exemple, la suite *ne ... rien* représente la négation de la totalité et *ne ... jamais* la négation dans le temps.

Dans le cadre de notre travail, l'utilisation des modèles $N$-gram positionnels nous a permis de déceler ces phénomènes particuliers au Français comme le démontrent les exemples du tableau 8.31. Il est important de noter que cette caractéristique ne s'applique pas au Portugais qui n'utilise pas la composition pour réaliser la négation.

| EM | Fréq. | Négation |
|---|---|---|
| 0.00120537 | 564 | ne ____ pas |
| 3.71391e-05 | 99 | ne ____ ____ pas |
| 3.2766e-06 | 39 | ne ____ que |
| 5.13283e-06 | 32 | ne ____ plus |
| 2.14286e-06 | 2 | ne ____ en rien |

TAB. 8.31 – Négation du Français

Les exemples du tableau 8.31 montrent un aspect intéressant des résultats que nous n'avions pas encore mentionné. En effet, il est possible que deux $N$-grams positionnels, contenant les mêmes formes graphiques mais pour des positions différentes, soient élus en même temps. C'est le cas des deux séquences *ne* ____ *pas* et *ne* ____ ____ *pas* qui représentent le même concept de la négation *ne ... pas* où les pointillés représentent généralement l'occurrence d'une ou deux formes graphiques intermédiaires. Ainsi, il est clair qu'un post-traitement des résultats serait particulièrement intéressant pour permettre de généraliser les données identifiées par l'extracteur. Malheureusement, cette étude n'a pas encore fait l'objet d'un travail exhaustif et par conséquent ne sera mentionné qu'à titre indicatif.

### Conjonction

La conjonction est l'un des problèmes classiques des analyses syntaxiques du discours. En effet, il est souvent difficile de déterminer le lien entre les éléments joints. Ainsi, une

conjonction peut être utilisée pour mettre en rapport deux propositions ou bien deux mots ou groupes de mots de même fonction dans une proposition. Aussi, un travail de désambiguïsation est nécessaire pour repérer les unions correctes. Celui-ci dépend généralement d'heuristiques recensées dans le cadre des travaux des lexicographes. Afin de guider cette analyse, les résultats obtenus par notre architecture mettent en évidence l'extraction d'un vaste ensemble de marqueurs syntaxiques de la conjonction qui utilisent dans la plupart des cas la conjonction de coordination *et* pour le Français et *e* pour le Portugais. Quelques exemples sont illustrés dans les deux tableaux 8.32 et 8.33.

| EM | Fréq. | Conjonction |
|---|---|---|
| 0.000169798 | 82 | des ____ et des |
| 0.000106503 | 63 | les ____ et les |
| 1.76087e-05 | 18 | ni ____ ni |
| 1.19408e-05 | 11 | en ____ et en |
| 1.12281e-05 | 8 | plus ____ et ____ plus |
| 1.21519e-05 | 8 | aux ____ et aux |
| 1.2e-06 | 2 | un ____ et un |
| 1e-06 | 2 | pas ____ et ____ ____ pas |
| 2.4e-06 | 2 | sa ____ et son |
| 2e-06 | 2 | leur ____ et leur |

TAB. 8.32 – Conjonction du Français

Les résultats montrent un foisonnement de combinaisons qui pourraient permettre la résolution d'un nombre important de problèmes d'ambiguïté dans le cadre des conjonctions. Il est clair cependant que des efforts supplémentaires seraient nécessaires pour organiser cette connnaissance. Là encore, nous n'avons pas entrepris de travaux plus importants dans cette direction. Nous sommes toutefois conscients de leur importance et nous espérons continuer nos travaux dans ce sens.

### Syntagmes Patrons

L'un des phénomènes classiques représenté par les associations lexicales est l'expression de formes idiomatiques du langage. Ce sont généralement des suites récurrentes

| EM | Fréq. | Conjonction |
|---|---|---|
| 1.23398e-05 | 12 | os ___ e as |
| 1.77005e-05 | 12 | ao ___ e à |
| 2.75441e-06 | 3 | nem ___ nem |
| 1.66625e-06 | 3 | no ___ ___ e no |
| 2.14232e-06 | 3 | no ___ ___ e ___ ___ no |
| 2.6464e-06 | 3 | seu ___ e ___ sua |
| 3.10267e-06 | 3 | a ___ ___ ou a |
| 1.24969e-06 | 2 | às ___ ___ e às |
| 2.10474e-06 | 2 | o ___ ___ ou o |
| 2.06845e-06 | 2 | menos ___ e mais |

TAB. 8.33 – Conjonction du Portugais

du discours, spécifiques au domaine d'application du corpus. En fait, une bonne partie d'entre elles sont des phrases "toutes faites" qui n'ont de sens que dans leur domaine. Il est apparent qu'un bon nombre de ces séquences ont été repérées par notre extracteur du fait de notre champ d'application particulièrement sujet à ce genre de phénomènes comme le montrent les deux tableaux suivants 8.34 et 8.35.

## 8.2 Analyse par comparaison

Après avoir analysé les résultats du GenLocalMaxs associé à l'Expectative Mutuelle, il est nécessaire de mettre en évidence le bien fondé de notre architecture par rapport à d'autres heuristiques numériques qui ont été utilisées avec succès dans de nombreuses études. Ainsi, nous proposons d'associer le GenLocalMaxs à cinq mesures d'association préalablement normalisées i.e. le coefficient d'association [27], le coefficient Dice [45], la Probabilité Conditionnelle Symétrique [30], le test $\Phi^2$ [29] et le coefficient de vraisemblance LogLike [28].

Dans ce cadre, nous avons utilisé un corpus de taille réduite comme support de notre analyse. En effet, du fait du processus de normalisation, l'effort computationnel requis est particulièrement important. Ainsi, pour un $N$-gram positionnel donné, il faut rechercher

| EM | Fréq. | Syntagme Patron |
|---|---|---|
| 9e-05 | 18 | en vue de l'adoption d'une directive |
| 3e-05 | 6 | (Le Parlement adopte la résolution législative) |
| 1.35e-05 | 3 | Monsieur le Président, je vous demande |
| 1.5e-05 | 3 | (La séance est ouverte à ____ heures) |
| 1e-05 | 2 | mes chers collègues, Monsieur le Commissaire, |
| 1e-05 | 2 | (Le président invite l'orateur à conclure) |
| 1e-05 | 2 | j'appelle simultanément la question n ____ de |
| 1e-05 | 2 | Je suis d'accord avec le rapporteur |
| 1e-05 | 2 | prolonger la séance de ____ minutes pour |
| 1e-05 | 2 | Questions à la Coopération politique européenne |

TAB. 8.34 – Syntagmes Patrons du Français

l'ensemble de tous ses sous-groupes c'est-à-dire tous ceux de rang 1 à $N-1$, contrairement à l'Expectative Mutuelle pour laquelle il suffit de calculer tous les sous-groupes de rang $N-1$ d'un $N$-gram positionnel donné. Nous avons donc extrait de la même base de textes deux corpora, l'un de 33 946 formes graphiques répertorié sous le nom de *qfr006.93.html* pour le Français et son homologue en Portugais de 31 013 mots noté *qpt006.93.html*. A partir de ceux-ci, nous avons appliqué le GenLocalMaxs associé à l'Expectative Mutuelle et aux cinq autres mesures d'association pour un environnement immédiat de cinq unités — i.e. cinq mots à droite et cinq mots à gauche de l'unité pivot. Par conséquent, un ensemble désordonné de $N$-grams positionnels — $\forall N, N = 2..10$ — que nous nous proposons d'analyser rigoureusement a été extrait.

Comme le lecteur l'aura compris, nous avons profité de cet impératif informatique pour changer la taille de l'environnement immédiat ainsi que le domaine d'application du corpus. En effet, même si le texte utilisé est de la responsabilité de la Commission Européenne, nous avons choisi un autre domaine, celui des Questions-Réponses. Le corpus est ainsi un ensemble de questions et de réponses faites à la Commission ce qui implique l'utilisation d'un texte forcément plus hâché que le précédent. L'objectif de ce choix est double. Premièrement, nous avons voulu confronter notre architecture à un type de matériel textuel moins fourni. En effet, les méthodes statistiques ont traditionnellement connu d'énormes difficultés à traiter les textes de petites tailles. Nous verrons dans le prochain chapitre —

| EM | Fréq. | Syntagme Patron |
|---|---|---|
| 6.81433e-05 | 17 | Senhor Presidente , gostaria de |
| 1.9995e-05 | 4 | A votação da matéria de fundo |
| 9.9975e-06 | 2 | tenho ____ observações a fazer |
| 9.9975e-06 | 2 | Regozijo-me pelo facto de |
| 9.08864e-06 | 2 | gostaria de fazer uma observação |
| 9.9975e-06 | 2 | relatório apresentado pelo senhor deputado |
| 8.33125e-06 | 2 | concordo com o senhor deputado |
| 9.9975e-06 | 2 | ( Aplausos da direita ) |
| 9.9975e-06 | 2 | felicitar a senhora deputada ____ pelo seu |
| 9.9975e-06 | 2 | por unanimidade no seio da Comissão |

TAB. 8.35 – Syntagmes Patrons du Portugais

Analyse Quantitative — que notre architecture met en évidence des résultats surprenants même pour ce genre de corpora. Ainsi, nous pourrons comparer sans état d'âme l'Expectative Mutuelle aux cinq autres mesures d'association. Deuxièmement, nous avons décidé d'augmenter la taille de l'environnement immédiat — seul paramètre de notre application — pour vérifier la validité des études lexicographiques qui prônent l'utilisation de contextes de tailles différentes.

### 8.2.1 Comparaison par liste de pertinence

Il est souvent difficile de comparer les résultats d'extraction obtenus à partir de plusieurs heuristiques. En effet, s'il est difficile de déterminer si un $N$-gram positionnel est une association lexicale ou non, il est encore plus compliqué de mettre en évidence les différences entre plusieurs extractions. Dans un premier temps, nous avons donc choisi de présenter une liste des différents $N$-grams extraits, ordonnés selon leur valeur de cohésion. Ainsi, pour chacune des six mesures d'association, nous montrerons les 20 premiers $N$-grams les plus pertinents tant pour le Français que pour le Portugais — voir Tableaux 8.47, 8.48, 8.49, 8.50, 8.51, 8.52. Ceci nous permettra de dégager un certain nombre de caractéristiques propres à chacune des différentes mesures et de guider notre analyse vers d'autres tests — notamment vers une comparaison par intersection.

La première impression qui surgit de l'analyse des résultats est que l'ensemble des

associations lexicales extraites diffère fondamentalement selon la mesure de cohésion utilisée. En effet, par exemple, les résultats obtenus à partir de l'Expectative Mutuelle sont crucialement différents de ceux obtenus à partir du test $\Phi^2$ ou du coefficient Dice. Cependant, un certain nombre d'analogies peuvent être identifiées grâce à une étude détaillée.

Premièrement, le coefficient d'association, la Probabilité Conditionnelle Symétrique et le test $\Phi^2$ démontrent un comportement similaire, du moins à partir des 20 $N$-grams les plus pertinents. Parallèlement, il semblerait que le coefficient de vraisemblance LogLike s'apparente à l'Expectative Mutuelle alors que le coefficient Dice révèle des caractéristiques communes à chacun des deux groupes précédents[10].

Deuxièmement, et plus important encore, toutes les mesures d'association normalisées, à l'exception de l'Expectative Mutuelle, dénoncent le problème classique des formes graphiques fréquentes. Ainsi, le degré d'attraction qui lie entre elles les différentes unités textuelles d'un $N$-gram positionnel est souvent sous-évalué du fait de l'occurrence de particules particulièrement fréquentes. Par exemple, la Probabilité Conditionnelle Symétrique, le test $\Phi^2$ et le coefficent Dice ont permis l'identification du 2-gram positionnel *Unidos* ___ *América* (*Unis* ___ *Amérique*) qui ne forme en aucun cas une association lexicale correcte. En effet, bien que la préposition *da* (*d'*) apparaisse nécessairement entre *Unidos* et *América*, le GenLocalMaxs associé aux cinq mesures testées a permis l'extraction de la séquence la plus courte négligeant ainsi l'occurrence de la préposition. Ainsi, le 2-gram positionnel [*0 Unidos 2 América*] manifeste une valeur de cohésion plus forte que celle du 3-gram [*0 Unidos 1 da 2 América*]. Or, ceci n'est pas supportable puisque chaque fois qu'apparaissent les formes graphiques *Unidos* et *América* séparées d'une unité de longueur, la préposition *da* s'intercale entre elles. Plus encore, dans ce même contexte, le nom commun *Estados* intervient obligatoirement en tête de la séquence pour donner lieu à l'association lexicale *Estados Unidos da América* (*Etats Unis d'Amérique*). Ainsi, nous illustrons nos propos à partir des résultats du concordanceur dans le tableau 8.45.

[10]Nous validerons ces intuitions à partir d'une comparaison par intersection dans la prochaine section. En effet, de telles affirmations ne peuvent être formulées à partir de la seule analyse des 20 premières associations lexicales extraites. Dans cette analyse, nous calculerons donc les intersections et les différences entre les ensembles de $N$-grams positionnels obtenus à partir des diverses mesures d'association.

| | | | | | |
|---|---|---|---|---|---|
| sociais nos | Estados | *Unidos* | da | *América* | e no Japão |
| dólares dos | Estados | *Unidos* | da | *América* | por quilo |

TAB. 8.36 – Concordanceur pour [*0 Unidos 2 América*]

Contrairement à ce phénomène indésirable démontré par les trois mesures d'association précédentes, l'Expectative Mutuelle démontre le seul comportement souhaitable, c'est-à-dire l'extraction de la suite la plus fréquente et la plus longue : *Estados Unidos da América*. Malheureusement, cette caractéristique fâcheuse s'étend également aux deux autres mesures d'association que sont le coefficient de vraisemblance LogLike et le coefficient Dice. En effet, cette dernière permet l'extraction de la séquence *turcs* ___ *kurdes* qui ne constitue pas une association lexicale correcte. Les résultats du concordanceur attestent d'ailleurs cette affirmation — voir Tableau 8.46.

| | | | | | |
|---|---|---|---|---|---|
| de sept | réfugiés politiques | *turcs* | et | *kurdes* | en Grèce |
| faim de | réfugiés politiques | *turcs* | et | *kurdes* | en protestation |
| des sept | réfugiés politiques | *turcs* | et | *kurdes* | qui font |
| Grèce Sept | réfugiés politiques | *turcs* | et | *kurdes* | qui sont |
| que sept | réfugiés politiques | *turcs* | et | *kurdes* | tenus en |

TAB. 8.37 – Concordanceur pour [*0 turcs 2 kurdes*]

Parallèlement, le coefficient de vraisemblance LogLike suggère les mêmes problèmes avec la suite *milhões* ___ *ecus* (*millions* ___ *écus*). En effet, encore une fois, l'espace laissé vide représente une et une seule forme graphique possible i.e. *de* (*d'*). Ainsi, un échantillon du résultat du concordanceur est donné dans le tableau 8.38.

Dans les deux cas, l'Expectative Mutuelle s'est comportée comme prévu et a permis l'identification des deux associations textuelles *réfugiés politiques turcs et kurdes* et *millions d'écus*. Ainsi, nous avons montré que les cinq mesures normalisées sous-évaluaient systématiquement les séquences contenant des formes graphiques fréquentes. Ceci est facilement repérable du fait de l'occurrence de trous inespérés dans les suites repérées. A partir de ces résultats, nous avons donc approfondi notre analyse afin de vérifier s'il existait d'autres phénomènes indésirables dus à la sous-évaluation des unités textuelles

| | | | | |
|---|---|---|---|---|
| 1 537 | *milhões* | de | *ecus* | ) . A Comissão |
| 1 500 | *milhões* | de | *ecus* | até atingir |
| em | *milhões* | de | *ecus* | , das ajudas |
| ... | ... | ... | ... | ... |
| 10 mil | *milhões* | de | *ecus* | para o período |

TAB. 8.38 – Concordanceur pour [*0 milhões 2 ecus*]

fréquentes. En fait, nous avons pu mettre en évidence une autre caractéristique propre à cette sous-évaluation. En effet, toutes les particules fréquentes apparaissant en fin ou en début de séquence sont systématiquement éliminées par le processus d'extraction. Ceci est facilement compréhensible. Prenons l'exemple de la locution prépositive *en matière de*. Il est clair que la sous-évaluation des séquences contenant des particules fréquentes aura pour effet de minorer l'importance de la préposition *de*. Ainsi, la séquence *en matière* se verra attribuer une valeur de cohésion plus forte que celle mise en évidence par la suite *en matière de*. Le coefficient Dice, la Probabilité Conditionnelle Symétrique, le test $\Phi^2$ et le coefficient de vraisemblance LogLike ont donc identifié la suite incomplète *en matière* comme le démontre le résultat du concordanceur — voir Tableau 8.39.

| | | | |
|---|---|---|---|
| une expérience | *en matière* | de | batteries et |
| ses activités | *en matière* | de | contrôle de |
| ce domaine | *en matière* | de | fiscalité des |
| ... | ... | ... | ... |
| communautaires | *en matière* | de | marchés |

TAB. 8.39 – Concordanceur pour *en matière*

Plus surprenant est le comportement du coefficient d'association qui rejette les deux suites *en matière* et *en matière de* et élit la séquence *en matière* ___ *politique* bien moins fréquente que les deux précédentes. Nous reviendrons sur ce résultat intéressant dans la partie suivante de notre rapport. En effet, il semblerait que le coefficient d'association, comme l'ont déjà montré certaines études [21], tend à sélectionner un nombre important de séquences de faible fréquence.

Cependant, cette caractéristique de sous-évaluation peut éventuellement montrer quelques atouts. En effet, comme nous l'avons vu précédemment, le GenLocalMaxs associé à l'Expectative Mutuelle permet l'identification d'un certain nombre de marqueurs dans le cadre de la post-modification. Or, même si ces marqueurs sont souvent intéressants, ils cachent néanmoins l'occurrence de termes de base. Par exemple, le marqueur élu *de substances radioactives* interdit le repérage par le GenLocalMaxs du terme de base *substances radioactives*. Le même phénomène est évidemment présent pour le Portugais. Dans ce contexte, la sous-évaluation des particules fréquentes peut présenter un aspect positif. Ainsi, l'association lexicale *Nations unies* a été repérée par la Probabilité Conditionnelle Symétrique, le coefficient Dice et le test $\Phi^2$ alors que l'Expectative Mutuelle a identifié le marqueur *des Nations unies*, plus fréquent et plus long. Encore une fois, nous nous aiderons des résultats du concordanceur pour illustrer nos propos — voir Tableau 8.40. Une nouvelle fois, le coefficient d'association n'a repéré aucune association lexicale contenant *Nations*[11]. Parallèlement, le coefficient de vraisemblance LogLike a identifié une séquence de formes graphiques plus longue, en l'occurrence *Nations unies et*. Cependant, dans d'autres cas, ces deux mesures se sont marginalisées par rapport à l'Expectative Mutuelle. Ainsi, pour le coefficient de vraisemblance LogLike, la séquence *Groupe consultatif* a été préférée au 3-gram *du Groupe consultatif* mis en évidence par l'expectative Mutuelle. Identiquement, le coefficient d'association a permis l'identification de l'association lexicale *résolution du Conseil* à l'instar de la séquence *la résolution du Conseil* repérée par l'Expectative Mutuelle.

| | | | |
|---|---|---|---|
| pertinentes | des | *Nations unies* | , afin que |
| générale | des | *Nations unies* | de 1992 |
| observateurs | des | *Nations unies* | et d'autres |
| ... | ... | ... | ... |
| recommendations | des | *Nations unies* | sur le |

TAB. 8.40 – Concordanceur pour *Nations unies*

L'utilisation de différentes mesures de cohésion implique donc l'identification de différentes associations lexicales. Ceci n'est pas une découverte en soi. Par contre, nous avons mis en

[11] La faible sensibilité du coefficient d'association aux fréquences élevées est la principale cause de ce phénomène.

évidence un certain nombre de caractéristiques communes à un certain nombre d'entre elles. Ainsi, il est évident que toutes les mesures normalisées à l'exception de l'Expectative Mutuelle sous-évaluent les associations contenant des particules fréquentes. Ce phénomène est facilement explicable à partir de l'analyse des normalisations effectuées pour chacune d'entre elles. Alors que la normalisation de l'Expectative Mutuelle ne prend en compte que la probabilité de voir apparaître une unité textuelle sachant l'occurrence des $N-1$ autres, la normalisation des autres mesures implique forcément le calcul de toutes les probabilités possibles au sein de ce même $N$-gram. Un exemple s'impose pour mieux comprendre cette affirmation. Supposons le 3-gram [*0 surface 1 du 2 globe*]. Son calcul de l'Expectative Mutuelle peut être représentée par l'équation suivante 8.1 où $T$ est la taille de l'énoncé et $F$ la taille de l'environnement immédiat.

$$ME([0 \quad surface \quad 1 \quad du \quad 2 \quad globe]) = \frac{\frac{f([0 \quad surface \quad 1 \quad du \quad 2 \quad globe])}{T-2F} \times \frac{f([0 \quad surface \quad 1 \quad du \quad 2 \quad globe])}{\frac{1}{3}\begin{pmatrix} f([0 \quad surface \quad 1 \quad du])+ \\ f([0 \quad surface \quad 2 \quad globe])+ \\ f([0 \quad du \quad 1 \quad globe]) \end{pmatrix}} \tag{8.1}$$

Parallèlement, pour le coefficient d'association[12], la formule utilisée serait la suivante 8.2.

$$I([0 \quad surface \quad 1 \quad du \quad 2 \quad globe]) = \log_2 \frac{f([0 \quad surface \quad 1 \quad du \quad 2 \quad globe])}{\frac{1}{3}\begin{pmatrix} (f([0 \quad surface \quad 1 \quad du]) \times f([globe]))+ \\ (f([0 \quad surface \quad 2 \quad globe]) \times f([du]))+ \\ (f([0 \quad du \quad 1 \quad globe]) \times f([surface])) \end{pmatrix}} \tag{8.2}$$

A la lumière des deux formules présentées, il est clair que la différence principale est à mettre au crédit des dénominateurs. En effet, alors que pour l'Expectative Mutuelle, il n'est pas nécessaire de faire appel aux probabilités marginales des unités textuelles, le coefficient d'association — comme d'ailleurs toutes les autres mesures — introduit

[12]Nous avons pris pour exemple le coefficient d'association, mais les mêmes conclusions peuvent être formulées pour les quatre autres mesures d'association normalisées.

nécessairement ces dernières dans les calculs. Ainsi, la probabilité de voir apparaître deux formes graphiques ensemble doit être pondérée par la probabilité marginale de la troisième unité espérée. Dans ce contexte, il est facile de comprendre pourquoi la suite *surface* ___ *globe* obtiendrait probablement une valeur de cohésion plus forte que celle de la séquence *surface du globe*. En effet, dans ce cas, la probabilité marginale très faible de la préposition *du* n'apparaitraît pas dans la formule ce qui ne pénaliserait pas autant la valeur de cohésion comme le montre l'équation 8.3.

$$I([0 \quad surface \quad 2 \quad globe]) =$$

$$\log_2 \frac{f([0 \quad surface \quad 2 \quad globe])}{(f([surface]) \times f([globe])} \tag{8.3}$$

Contrairement à cette affirmation, il est fort probable que l'Expectative Mutuelle de la suite *surface* ___ *globe* dénote une valeur plus faible que la suite plus longue *surface du globe*. En effet, comme le montre l'équation suivante 8.4, les probabilités marginales des deux formes graphiques *surface* et *globe* sont utilisées minorant ainsi la valeur de cohésion du 2-gram par rapport au 3-gram.

$$ME([0 \quad surface \quad 2 \quad globe]) =$$

$$\frac{\frac{f([0 \quad surface \quad 2 \quad globe])}{T-2F}}{\frac{f([0 \quad surface \quad 2 \quad globe])}{\frac{1}{2}(f([surface])+f([globe]))}} \tag{8.4}$$

Le problème majeur des mesures d'association préalablement énoncées peut être résumé par le fait que ces dernières tentent de mesurer une double expectative en une seule mesure. Ainsi, pour un $N$-gram donné, elles s'appliquent à calculer la probabilité d'apparition d'une unité textuelle sachant que les $N-1$ autres l'entourent ainsi que la probabilité d'occurrence du $N-1$-gram complémentaire connaissant l'occurrence de l'unique forme graphique. Il est évident que c'est cette dernière expectative qui fausse les résultats de cohésion. En effet, si cette hypothèse est valide pour les modèles 2-gram, ceci n'est pas le cas pour les modèles $N$-gram i.e $\forall N, N > 2$. De fait, dans l'exemple précédent, calculer l'expectative de voir apparaître *surface* et *globe* autour de la préposition *du* ne présente aucun intérêt sinon celui de fausser les résultats. En effet, seule l'information contenue dans *surface* et *globe* sur l'apparition de la préposition *du* peut être considérée pertinente. Dans ce

cadre, l'introduction de l'Expectative Mutuelle s'est avérée nécessaire pour la définition de valeurs de cohésion cohérentes. Dans le même ordre d'idée, le lecteur intéressé pourra se référer au travail de J. Silva [30] qui propose une normalisation originale pour le cas des modèles $N$-gram classiques qui lui permet de minimiser l'importance des fragments fréquents et par conséquent d'éviter la sous-évaluation de séquences les contenant.

### 8.2.2 Comparaison par intersection

Comme nous l'avons mentionné dans la partie précédente, il semblerait qu'il existe certaines similitudes entre mesures d'association. Cependant, une simple intuition basée sur les 20 premiers candidats les plus pertinents ne fait pas office de vérité scientifique. Afin de mieux analyser ce phénomène, nous proposons une analyse par intersection qui consiste à évaluer le nombre de $N$-grams élus que partagent deux à deux les six mesures d'association proposées. Ainsi, nous introduisons les premières données quantitatives de notre évaluation. Cette étude servira donc d'introduction au prochain chapitre et nous permettra de mettre en évidence les relations entre mesures de cohésion de façon plus naturelle.

Les tableaux 8.41 et 8.42 — respectivement pour le Français et le Portugais — présentent les pourcentages d'associations lexicales communes aux mesures d'association proposées. Ainsi, le pourcentage mentionné dans la case $ij$ correspondant à la ligne $i$ et à la colonne $j$ représente la proportion d'associations lexicales extraites par la mesure d'association de la ligne $i$ qui est également repérée par la mesure d'association de la colonne $j$. Pour simplifier la présentation des résultats, nous utiliserons les abréviations suivantes pour chacune des mesures d'association : I pour le coefficient d'association, DICE pour le coefficient Dice, SCP pour la Probabilité Conditionnelle Symétrique, LOGLIKE pour le coefficient de vraisemblance LogLike et ME pour l'Expectative Mutuelle.

La première évidence est à mettre au crédit de l'équivalence des résultats entre le Français et le Portugais. En effet, les deux tableaux révèlent des résultats semblables. Ainsi, les propos que nous porterons au sujet du Français pourront être étendus au Portugais. Comme nous l'avions prévu, certaines de nos hypothèses ont pu être validées alors que d'autres ne se sont pas confirmées à la lueur des résultats plus complets. Ainsi, trois groupes distincts de mesures d'association peuvent être dégagés.

| | PHI | DICE | LOGLIKE | ME | I | SCP |
|---|---|---|---|---|---|---|
| PHI | - | 18.85% | 27.14% | 24.41% | 74.56% | 54.75% |
| DICE | 22.24% | - | 27.89% | 11.74% | 20.49% | 18.56% |
| LOGLIKE | 4.78% | 4.16% | - | 2.93% | 4.88% | 3.54% |
| ME | 45.11% | 18.41% | 30.85% | - | 52.21% | 45.25% |
| I | 52.08% | 12.13% | 23.06% | 19.73% | - | 41.27% |
| SCP | 70.17% | 20.17% | 25.82% | 31.38% | 75.73% | - |

TAB. 8.41 – Comparaison par intersection pour le Français

| | PHI | DICE | LOGLIKE | ME | I | SCP |
|---|---|---|---|---|---|---|
| PHI | - | 20.88% | 29.74% | 32.48% | 84.30% | 75.33% |
| DICE | 18.76% | - | 26.81% | 9.91% | 17.03% | 20.11% |
| LOGLIKE | 3.98% | 3.99% | - | 3.06% | 4.52% | 3.31% |
| ME | 45.20% | 15.35% | 31.86% | - | 51.83% | 42.24% |
| I | 45.34% | 10.20% | 18.17% | 20.03% | - | 38.37% |
| SCP | 73.76% | 21.91% | 24.25% | 29.73% | 69.86% | - |

TAB. 8.42 – Comparaison par intersection pour le Portugais

Le premier groupe, déjà repéré par l'analyse succinte des résultats, rassemble le test $\Phi^2$, le coefficient d'association et la Probabilité Conditionnelle Symétrique. En effet, le tableau 8.41 montre clairement qu'un nombre important d'associations lexicales sont communes aux trois mesures de cohésion. En particulier, on notera que la différence majeure entre le coefficient d'association et les deux autres heuristiques est due au fait que celui-ci extrait un nombre plus important d'associations lexicales[13]. Ainsi, les pourcentages mis en évidence par la ligne du coefficient d'association montrent des valeurs inférieures à celles du test $\Phi^2$ et de la Probabilité Conditionnelle Symétrique.

Le deuxième ensemble peut rassembler le coefficient Dice et le coefficient de vraisemblance LogLike. Ce regroupement n'est pas flagrant. Cependant, ces deux mesures font découvrir un nombre maximum d'associations lexicales identiques par rapport aux

[13]Nous reviendrons sur ce résultat dans le prochain chapitre.

autres mesures. En effet, soit le coefficient Dice, soit le coefficient de vraisemblance LogLike révèlent un pourcentage maximum d'unités semblables. Là encore, la grande différence tient au fait que le coefficient LogLike extrait un nombre considérablement plus important de $N$-grams positionnels candidats par rapport au coefficient Dice et plus généralement par rapport à toutes les mesures d'association. Si le regroupement antérieur était prévisible, celui-ci ne l'était pas par l'analyse des 20 premières unités polylexicales candidates. En effet, il apparaissait que le LogLike démontrait des caractéristiques plus proches de celles de l'Expectative Mutuelle, ce qui ne s'est pas vérifié.

Finalement, l'Expectative Mutuelle montre un comportement plus singulier qui ne permet pas de l'associer à une autre mesure. En effet, on pourrait dire que l'Expectative Mutuelle se positionne au milieu des deux autres groupes bénéficiant de caractéristiques communes aux cinq autres mesures d'association. Ainsi, l'Expectative Mutuelle met en évidence un nombre important d'unités lexicales complexes communes au test $\Phi^2$, au coefficient d'association et à la Probabilité Conditionnelle Symétrique mais aussi au coefficient de vraisemblance LogLike pour lequel 30.85% des associations sont identiques. La mesure la plus éloignée de l'Expectative Mutuelle est certainement le coefficient Dice, ce qui du reste est surprenant puisque l'Expectative Mutuelle et le coefficient Dice définissent la même formule de cohésion pour les modèles 2-gram positionnels. Mais les chiffres attestent clairement cette distance mettant en évidence les pourcentages les plus faibles pour chacune des deux mesures.

Pour ne pas nous éloigner du thème de cette partie qui concerne exclusivement l'analyse qualitative des résultats, il nous a semblé intéressant de montrer l'ensemble des associations lexicales les plus pertinentes communes à toutes les mesures d'association. En effet, il serait souhaitable que celles-ci démontrent un certain degré de qualité. Les résultats sont cependant mitigés. Ainsi, même si le tableau 8.43 illustre un ensemble d'unités complexes de qualité élévée pour le Français, le tableau 8.44, pour le Portugais, ne satisfait pas nos hypothèses de départ. En effet, pour le Portugais, la plupart des associations lexicales repérées sont des syntagmes patrons peu fréquents. Au contraire, pour le Français, les résultats mettent en évidence l'extraction d'un ensemble important de noms composés. Ces résultats sont surprenants. Nous nous attendions en effet à l'occurrence de résultats semblables pour le Portugais et le Français. Plusieurs raisons

peuvent être avancées. La première tient au fait que le nombre d'unités lexicales complexes communes aux six mesures d'association est particulièrement faible — 73 pour le Français et 48 pour le Portugais. Ainsi, la représentativité des résultats n'est pas assurée et par conséquent, leur analyse doit être effectuée avec une certaine précaution. La deuxième raison principale tient au fait que la langue Portugaise fait appel à un nombre important de prépositions qui accentuent les différences entre mesures d'association, notamment entre l'Expectative Mutuelle et le reste des heuristiques numériques.

| Freq. | Associations Lexicales |
|---|---|
| 250 | la Commission |
| 102 | États membres |
| 101 | la Communauté |
| 65 | ? Réponse |
| 43 | ne ____ pas |
| 34 | a été |
| 21 | Parlement européen |
| 19 | État membre |
| 16 | milions d'écus |
| 14 | pays tiers |

TAB. 8.43 – 10 Meilleures associations lexicales communes aux Mesures d'association pour le Français

### 8.2.3 Comparaison du processus d'extraction

Jusqu'à présent, nous nous sommes intéressés à comparer les performances de différentes mesures d'association. Cependant, notre architecture propose une nouvelle méthode d'extraction qu'il convient de comparer aux méthodes qui préconisent l'application de valeurs seuil. Dans ce cadre, nous proposons d'abord d'illustrer les 10 $N$-grams positionnels les plus fortement valués par l'Expectative Mutuelle et de les compararer à ceux obtenus à partir du GenLocalMaxs. Le tableau 8.45 montre ces résultats pour le corpus de 200 000 formes graphiques.

La particularité des modèles $N$-gram positionnels ne permet effectivement pas l'appli-

| Freq. | Associations Lexicales |
|---|---|
| 3 | prestação de serviço |
| 2 | artigo ___ do Tratado |
| 2 | é compatível com |
| 2 | igualdade ___ tratamento ___ homens ___ mulheres |
| 2 | na comunicação da Comissão |
| 2 | a ___ ___ ___ de acordo com as |
| 2 | , como por exemplo ___ ___ , |
| 2 | da ___ de ___ entre ___ e |
| 2 | em ___ com os representantes dos |
| 2 | Estados-membros que ___ ___ uma taxa |

TAB. 8.44 – 10 Meilleures associations lexicales communes aux Mesures d'association pour le Portugais

cation des valeurs seuil comme processus d'extraction. En effet, comme le montrent les données du tableau 8.45, cette méthode ne permet pas de délimiter l'ensemble des unités polylexicales de forme précise. De fait, il est clair que si une séquence est fortement valuée par la mesure d'association, il est fort probable qu'une de ses sur- ou sous-séquences soit fortement pondérée également. Par exemple, pour le Français, si la suite "*Monsieur le Président ,*" peut être considérée comme une association lexicale complexe, ceci n'est pas le cas pour les séquences "*Monsieur* ___ *Président ,*" et "*Monsieur le*" bien que leur pondération soit particulièrement élevée. Dans ce cadre, le GenLocalMaxs démontre des résultats d'extraction sans comparaison possible.

L'un des problèmes des valeurs seuil réside dans leur définition. En effet, la valeur doit être assez permissive pour permettre un rappel suffisamment grand et assez rigide pour que les résultats soient les plus précis possibles. Cependant, comme nous l'avons déjà dit, ces valeurs seuils sont suceptibles de laisser de côté un certain nombre d'associations lexicales correctes du fait qu'il existe un grand nombre de ces unités qui sont peu fréquentes et par conséquent faiblement valuées. A titre d'exemple, nous illustrons cette hypothèse à partir du tableau 8.46 dans lequel nous montrons la distribution des unités lexicales complexes extraites par le GenLocalMaxs.

| Français |
|---|
| de la |
| Monsieur le Président , |
| Monsieur ___ Président , |
| ne ___ pas |
| le Président , |
| la ___ de |
| ce qui concerne |
| Monsieur le |
| la Commission |
| à la |
| Etats membres |
| **Portugais** |
| e Monetários e da Polıt́ica Industrial , |
| Monetários e da Polıt́ica Industrial , |
| Pescas ___ ___ Desenvolvimento Rural |
| Pescas ___ do ___ Rural |
| . Em segundo ___ , |
| . ___ segundo lugar , |
| . ___ ( EN ) Senhora Presidente |
| , tal como |
| Comissão da ___ , ___ Pescas |
| da Comissão ___ Assuntos Económicos e |

TAB. 8.45 – 10 meilleures associations lexicales pour l'Expectative Mutuelle sans GenLocalMaxs

Une explication s'impose pour mieux comprendre le tableau 8.46. La première colonne correspond à la puissance de pondération de l'Expectative Mutuelle. Par exemple, pour la première ligne nous considérons tous les $N$-grams positionnels dont la valeur d'EM est supérieure à un exposant de $10^{-5}$. La deuxième colonne correspond au nombre total de $N$-grams positionnels pour une puissance d'EM donnée. Finalement, la troisième colonne définit le nombre de $N$-grams positionnels extraits par le GenLocalMaxs pour une puissance d'EM donnée. Ainsi, les résultats montrent que la plupart des

| Puisance de l'EM | Total de N-grams | N-grams GenLocalMaxs |
|---|---|---|
| > e-05 | 1709 | 54 |
| = e-05 | 27961 | 1793 |
| = e-06 | 122528 | 10631 |
| = e-07 | 65030 | 86 |
| = e-08 | 32619 | 0 |
| = e-09 | 11309 | 0 |

TAB. 8.46 – Distribution des associations lexicales par EM

*N*-grams positionnels extraits par le GenLocalMaxs se trouvent dans une puissance d'EM moyenne. Il est clair qu'une méthode d'extraction par valeur seuil pourrait difficilement s'appliquer. En effet, premièrement, si l'on voulait considérer une valeur limite de l'ordre de $10^{-6}$, un très grand nombre d'associations candidates ne seraient pas correctes. Cette valeur peut être approximativement quantifiée par le calcul suivant : $(1709 - 54) + (27961 - 1793) + (122528 - 10631) = 139720$. Cette valeur n'est qu'une approximation grossière mais montre effectivement l'ampleur des déchets potentiels. D'autre part, dans ces mêmes conditions, au moins 86 associations lexicales potentiellement correctes ne seraient pas identifiées.

Les chiffres ainsi que les exemples proférés sont suffisamment clairs pour déduire que le GenLocalMaxs propose une meilleure solution que les valeurs seuil pour le processus d'extraction. Il est également vrai que le GenLocalMaxs n'est pas une solution idéale. En effet, comme nous l'avons déjà remarqué, la notion de marqueur cache des insuffisances notoires d'extraction qu'il conviendra d'analyser dans un futur proche.

## 8.3 Conclusion

Nous concluons ainsi notre analyse qualitative des résultats obtenus par le GenLocalMaxs associé aux six mesures d'association normalisées. Comme nous l'avons démontré, cette architecture propose une solution intéressante pour l'extraction d'un ensemble important de phénomènes de figement. En particulier, l'Expectative Mutuelle a mis en évidence certaines caractéristiques propres qui la distingue fondamentalement des cinq autres heuristiques. Ainsi, l'Expectative Mutuelle ne sous-évalue pas les associations

lexicales qui contiennent des formes graphiques dont la fréquence est élevée.

Parallèlement, les comparaisons réalisées entre valeurs seuil et GenLocalMaxs montrent une supériorité notoire de notre architecture.

Cependant, même si l'analyse qualitative des données est une étape nécessaire à une évaluation exhaustive, celle-ci n'est pas suffisante. En effet, comme nous en avons fait l'expérience dans le paragraphe précédent, un certain nombre de caractéristiques ne peuvent être mises en évidence que par l'analyse quantitative des résultats. C'est ainsi que nous proposerons dans la prochaine partie de notre rapport une analyse quantitative exhaustive des résultats d'extraction ■

| Français |
|---|
| de la |
| États membres |
| QUESTION ÉCRITE N |
| la Commission |
| des Communautés européennes |
| la Commission des Communautés européennes |
| Réponse donnée par |
| Communauté et ses États membres |
| l'honorable parlementaire |
| Coopération politique européenne |
| la Coopération politique |
| ne ____ pas |
| par M. ____ au nom de la |
| dans le |
| la ____ de |
| droits de l'homme |
| la Communauté |
| ce qui concerne |
| État membre |
| en matière de |
| **Portugais** |
| PERGUNTA ESCRITA N |
| das Comunidades Europeias |
| Comissão das Comunidades Europeias |
| senhor deputado |
| Resposta dada |
| cooperação política europeia |
| a Comissão |
| milhões de ecus |
| Resposta dada pela ____ ____ ____ em nome |
| a ____ de |
| Parlamento Europeu |
| Países Baixos |
| pergunta escrita n |
| parceiros sociais |
| Livro Verde |
| no âmbito da |
| África do Sul |
| no sentido de |
| em matéria de |
| direitos humanos |

TAB. 8.47 – 20 meilleures associations lexicales pour la mesure d'Expectative Mutuelle

| Français |
|---|
| l'adaptation ___ modules ___ situations ___ ___ ___ visé |
| approuvées ___ ___ ___ ___ D'autres ___ ___ ___ l'étude |
| 6,7 ___ ___ ___ ___ garde ___ 2,9 |
| susmentionnées ___ approuvées ___ ___ Commission ___ D'autres |
| modules ___ situations ___ ___ public visé |
| l'Organisation mondiale ___ ___ santé ___ OMS |
| d'absorption ___ ___ ___ dans ___ ___ zones ___ déclin |
| tonnage ___ ___ puissance ___ ___ port d'immatriculation |
| D'autres ___ sont ___ l'étude ___ feront |
| conjoint ___ ministres ___ l'Énergie ___ ___ l'Environnement |
| postaux ___ ___ ___ éloignées et ___ ___ ___ rentabilité |
| d'immatriculation ___ ___ ___ ___ le ___ d'engin ___ pêche |
| changement ___ ___ ___ surface du globe |
| surface ___ globe |
| organisés précédemment |
| objections formulées |
| nome ___ Kozani |
| main-d'oeuvre ___ ___ ___ ___ la CES représente-t-elle |
| L'Algemeen Burgerlijk Pensioenfonds |
| fondée ___ ___ ___ technologiques |
| **Portugais** |
| reflecte ___ ___ supracitadas e ___ ___ ___ ___ Estão |
| exprimir ___ seu ___ ___ ___ remoção |
| protocolo ___ ___ procedimento ___ ___ défices excessivos |
| partilham ___ ___ ___ senhor ___ acerca |
| analisados ___ ___ ___ serão ___ proximamente |
| potência ___ porto ___ matrícula ___ ___ ___ navio |
| correspondentes ___ ___ espanhola ___ em ___ ___ ___ portos |
| pesca ___ ___ ___ N ___ ___ ___ Miguel Arias |
| Unidos ___ América |
| Registo ___ ___ ___ ___ procedeu ___ informatização |
| purificação étnica |
| Públicas ___ ___ ___ Internos |
| porto ___ ___ ___ nome ___ navio |
| perseguidos ___ ___ turcas |
| navio ___ ___ ___ ___ de pesca |
| navio ___ ___ ___ arte ___ pesca |
| goze ___ ___ ___ generalizado |
| fronteira ___ ___ Indonésia ___ ___ ___ Guiné |
| equilíbrio ___ ___ oferta ___ ___ procura |
| energias renováveis |

TAB. 8.48 – 20 meilleures associations lexicales pour le coefficient d'association

| Français |
|---|
| tonnage ___ ___ puissance ___ ___ port d'immatriculation |
| l'adaptation ___ modules ___ situations ___ ___ ___ visé |
| Doc ___ COM( ___ ___ ___ final |
| approuvées ___ ___ ___ ___ D'autres ___ ___ ___ l'étude |
| l'Organisation mondiale ___ ___ ___ ___ OMS |
| 1992 ___ ___ 93/C ___ ___ Objet : |
| Communautés européennes ___ ___ ___ 1992 ___ ___ 93/C |
| Vlaamse Gemeenschap |
| turcs ___ kurdes |
| Sotiris Kostopoulos |
| Sérgio Ribeiro |
| rayonnements ___ ionisants |
| QUESTION ÉCRITE N |
| Nations unies |
| Miguel Arias Cañete |
| l'ex-Union soviétique |
| Leon Brittan |
| L'Algemeen Burgerlijk Pensioenfonds |
| Jaak Vandemeulebroucke ___ ARC |
| flotte espagnole immatriculée |
| **Portugais** |
| potência ___ porto ___ matrícula ___ ___ ___ navio |
| incluídos ___ ___ ___ ___ frota espanhola matriculada |
| Registo ___ ___ ___ ___ procedeu ___ informatização |
| exprimir ___ ___ consentimento ___ ___ remoção |
| reflecte ___ ___ supracitadas ___ aprovadas ___ ___ ___ Estão |
| 1992 ___ ___ 93/C ___ ___ Objecto : |
| Vida Selvagem |
| turcos ___ curdos |
| Sotiris Kostopoulos |
| Sérgio Ribeiro |
| Públicas ___ ___ ___ Internos |
| protocolo ___ ___ procedimento ___ ___ défices excessivos |
| portos ___ província ___ ___ ___ especificando |
| Pian Martino |
| PERGUNTA ESCRITA |
| perfeitamente cientes |
| Países Baixos |
| Negócios Estrangeiros |
| Nações Unidas |
| Miguel Arias Cañete |

TAB. 8.49 – 20 meilleures associations lexicales pour le coefficient Dice

| Français |
|---|
| tonnage ___ ___ puissance ___ ___ port d'immatriculation |
| l'adaptation ___ modules ___ situations ___ ___ ___ visé |
| Fichier ___ ___ ___ ___ a informatisé |
| Doc ___ COM( ___ ___ ___ final |
| approuvées ___ ___ ___ ___ D'autres ___ ___ ___ l'étude |
| 6,7 ___ ___ ___ ___ garde ___ 2,9 |
| Communautés européennes ___ ___ ___ 1992 ___ ___ 93/C |
| l'importance ___ ___ ___ et ___ ___ ___ attachent |
| susmentionnées ___ approuvées ___ ___ Commission ___ D'autres |
| modules ___ situations ___ ___ public visé |
| l'Organisation mondiale ___ ___ santé ___ OMS |
| d'absorption ___ ___ ___ dans ___ ___ zones ___ déclin |
| D'autres ___ sont ___ l'étude ___ feront |
| conjoint ___ ministres ___ l'Énergie ___ ___ l'Environnement |
| 1992 ___ ___ 93/C ___ ___ Objet : |
| O'Hagan ___ ED ___ ___ Conseil ___ ___ ___ ___ 14 |
| immatriculée ___ ___ ports ___ la province |
| postaux ___ ___ ___ éloignées et ___ ___ ___ rentabilité |
| changement ___ ___ ___ surface du globe |
| montants attribués ___ programmes ___ ___ l'initiative |
| **Portugais** |
| potência ___ porto ___ matrícula ___ ___ ___ navio |
| incluídos ___ ___ ___ ___ frota espanhola matriculada |
| correspondentes ___ frota ___ ___ em ___ ___ ___ portos |
| reflecte ___ ___ supracitadas e ___ ___ ___ ___ Estão |
| Registo ___ ___ ___ ___ procedeu ___ informatização |
| porto ___ ___ ___ nome ___ navio |
| navio ___ ___ ___ ___ de pesca |
| navio ___ ___ ___ arte ___ pesca |
| exprimir ___ ___ consentimento ___ ___ remoção |
| portos ___ província ___ ___ ___ especificando |
| protocolo ___ ___ procedimento ___ ___ défices excessivos |
| partilham ___ ___ ___ ___ deputado acerca |
| analisados ___ ___ ___ serão ___ proximamente |
| 1992 ___ ___ 93/C ___ ___ Objecto : |
| Vlaamse Gemeenschap |
| Vida Selvagem |
| Unidos ___ América |
| turcos ___ curdos |
| transportador ___ utilizador |
| Tamer Erkots |

TAB. 8.50 – 20 meilleures associations lexicales pour la Probabilité Conditionnelle Symétrique

| Français |
|---|
| Toutes ___ ___ veillent à ___ ___ modules ___ situations |
| tonnage ___ ___ puissance ___ ___ port d'immatriculation |
| l'adaptation ___ modules ___ situations ___ ___ ___ visé |
| 6,7 ___ ___ ___ ___ garde ___ 2,9 |
| veillent ___ ___ des ___ ___ situations |
| Fichier ___ ___ ___ ___ a informatisé |
| Doc ___ COM( ___ ___ ___ final |
| Communautés européennes ___ ___ ___ 1992 ___ ___ 93/C |
| flotte espagnole immatriculée ___ ___ ports ___ ___ province |
| Fichier ___ bateaux La ___ ___ informatisé ___ fichier |
| précisant ___ longueur ___ leur tonnage ___ ___ puissance |
| l'importance ___ ___ ___ et ___ ___ ___ attachent |
| susmentionnées ___ approuvées ___ la ___ ___ D'autres |
| modules ___ situations ___ du ___ visé |
| susmentionnées ___ approuvées ___ ___ Commission ___ D'autres |
| d'absorption ___ ___ ___ dans ___ ___ zones ___ déclin |
| l'Organisation mondiale ___ ___ santé ___ OMS |
| d'Afrique ___ ___ Caraïbes et ___ Pacifique ___ ACP |
| longueur ___ perpendiculaires ___ qui ___ repris ___ ___ fichier |
| turcs ___ kurdes |
| **Portugais** |
| matrícula ___ ___ ___ navio e ___ ___ arte ___ pesca |
| reflecte ___ ___ supracitadas e ___ ___ ___ ___ Estão |
| navio ___ ___ ___ arte ___ pesca |
| navio ___ ___ ___ ___ de pesca |
| exprimir ___ ___ consentimento ___ ___ remoção |
| Registo ___ ___ ___ ___ procedeu ___ informatização |
| porto ___ ___ ___ nome ___ navio |
| tonelagem ___ ___ ___ porto ___ matrícula |
| Orientação ___ Garantia Agrícolas ___ FEOGA ___ ___ secção |
| portos ___ província ___ ___ ___ especificando |
| protocolo ___ ___ ___ ___ aos défices excessivos |
| partilham ___ ___ ___ ___ deputado acerca |
| analisados ___ ___ ___ serão ___ proximamente |
| dióxido ___ carbono ___ ___ melhorar ___ eficácia energética |
| 1992 ___ ___ 93/C ___ ___ Objecto : |
| nomes ___ a ___ ___ ___ ___ nomes |
| fronteira ___ ___ ___ ___ a ___ Guiné |
| equilíbrio ___ a ___ ___ ___ procura |
| Vlaamse Gemeenschap |
| Unidos ___ América |

TAB. 8.51 – 20 meilleures associations lexicales pour le test $\Phi^2$

| Français |
|---|
| la Commission |
| États membres |
| de la |
| 1992 ___ ___ 93/C |
| Communautés européennes |
| QUESTION ÉCRITE N ___ ___ M. |
| QUESTION ÉCRITE N ___ de |
| ? Réponse |
| Commission ___ Communautés européennes |
| Réponse donnée |
| à la |
| la Communauté |
| au nom ___ ___ Commission |
| La Commission |
| septembre 1992 |
| 1er ___ 1992 |
| ne ___ pas |
| 1er ___ ___ ___ ___ 93/C ___ ___ Objet dans le |
| Coopération politique européenne |
| donnée par ___ ___ au nom |
| **Portugais** |
| das Comunidades Europeias |
| 1992 ___ ___ 93/C |
| PERGUNTA ESCRITA ___ ___ ___ Sr. |
| PERGUNTA ESCRITA N ___ do |
| à ___ ___ Comunidades Europeias |
| Comissão ___ Comunidades Europeias |
| Resposta dada |
| ? Resposta |
| em nome ___ Comissão |
| a Comissão |
| cooperação política europeia |
| Setembro ___ 1992 |
| Comunidade ___ ___ seus Estados-membros |
| os seus |
| 1 ___ JO n |
| S ___ ___ ___ ___ Comunidades Europeias |
| milhões ___ ecus |
| Parlamento Europeu |
| JO n L |
| bem como |

TAB. 8.52 – 20 meilleures associations lexicales pour le coefficient de vraisemblance LogLike

# Chapitre 9

# Analyse Quantitative

Dans le chapitre précédent, nous avons abordé l'aspect qualitatif des résultats d'extraction. Bien que cette étude soit nécessaire — nous devrions même dire primordiale, elle n'est cependant pas suffisante pour donner une vision juste et éclairée du processus d'extraction. Parallèlement, elle n'est pas non plus suffisante pour permettre une caractérisation complète du phénomène de figement. En effet, de nombreuses questions restent encore sans réponse. Par exemple, quelle est la performance — précision et rappel — de notre architecture ? Quelles sont les valeurs moyennes de la taille et de la fréquence d'une association lexicale ? Il est clair que seule une analyse quantitative peut répondre à toutes ces questions importantes. Traditionnellement, les études quantitatives se sont limitées à calculer le taux de précision et le taux de rappel du système analysé[1]. Ces deux indicateurs ont pour objectif principal d'évaluer la performance de l'architecture proposée. Cependant, bien que ceux-ci soient d'une extrême importance, d'autres analyses sont nécessaires pour comprendre encore mieux le concept d'association lexicale. Ainsi, nous étudierons de forme exhaustive et détaillée toutes les données relatives à la taille des associations lexicales, leur fréquence ainsi que la proportion de suites contiguës *versus* non contiguës. Les deux axes principaux de l'étude quantitative i.e. la performance de l'extraction et la caractérisation des unités lexicales complexes seront ainsi traités. Afin de ne pas rompre avec la dynamique de notre analyse précédente, nous commencerons par étudier les caractéristiques propres aux unités polylexicales élues par notre architecture, pour, dans un deuxième temps, conclure sur une évaluation impartiale de la performance de notre système d'extraction.

[1] D'autres indicateurs comme la F-mesure ont également été proposés.

## 9.1 Analyse de la Taille des Associations Lexicales Candidates

Très peu d'études ont été réalisées sur la taille des associations lexicales. En fait, la seule que nous ayons recensée a été réalisée par J. Justeson et S. Katz [22] dans le cadre des noms composés de l'Anglais. Ainsi, ils ont mis en évidence que la plupart des associations lexicales de type nominal contiennent entre deux et trois mots. Nous rappelons ces résultats dans le tableau 9.1 où, pour chaque dictionnaire considéré, 200 termes ont été extraits aléatoirement.

| Dictionnaire | Taille=1 | Taille=2 | Taille=3 | Taille $\geq 4$ |
|---|---|---|---|---|
| Fibre optique | 43 | 109 | 36 | 12 |
| Médecine | 88 | 80 | 22 | 10 |
| Physique et Mathématique | 41 | 125 | 29 | 5 |
| Psychologie | 64 | 120 | 12 | 4 |

TAB. 9.1 – Taille des Termes analysés par J. Justeson et S. Katz

Les résultats montrent clairement qu'à partir de quatre formes graphiques, les unités polylexicales repérées sont peu nombreuses voire même insignifiantes[2]. Cependant, il est important de noter que cette étude a été réalisée exclusivement pour l'Anglais. Or, il est évident que le phénomène de figement est fondamentalement différent dans sa réalisation selon que l'on considère les langues d'origine latine ou bien anglo-saxonne. B. Daille [21] met en évidence cette caractéristique en utilisant lors de son processus d'extraction, les patrons syntaxiques $ADJ \quad N$ et $N \quad N$ pour l'Anglais, et $N \quad ADJ$, $N \quad N$, $N$ *de* $(DET) \quad N$, $N \quad PREP \quad N$ et $N$ *à* $(DET) \quad N$ pour le Français. Ainsi, il semblerait que la réalisation des associations lexicales du Français suggère des suites qui contiendraient entre 2 et 4 formes graphiques alors que pour l'Anglais une forte proportion des unités polylexicales serait formée d'au plus deux formes graphiques. Cette analyse est renforcée par les travaux de M.L. Herviou-Picard [25] qui propose les patrons syntaxiques suivants dans le

[2]Du fait de son support, cette étude s'affranchit d'une certaine légitimité linguistique. Cependant, comme l'ont montré G. Dias *et al.* [111], les dictionnaires peuvent contenir des erreurs. Ainsi, des associations lexicales de petite taille devraient souvent être étendues pour donner lieu à des unités polylexicales plus longues et moins générales. Par conséquent, les études basées sur des ressources manuellement compilées doivent être utilisées avec précaution.

cadre du Français : $N \quad ADJ$, $ADJ \quad N$, $N \quad N$, $N \quad PREP \quad N$, *N à N*, *N en N*, *N de le/la/l'/les N* et *N à le/la/l'/les N*. Il est par conséquent fort probable que les résultats d'extraction proposés par notre architecture soient plus proches en moyenne d'une taille de trois formes graphiques plutôt que de deux unités lexicales comme le démontrent J. Justeson et S. Katz pour l'Anglais. Nous nous proposons de valider cette hypothèse dans les deux paragraphes suivants.

### 9.1.1 Première Expérience

Dans le cadre de notre architecture, la taille des associations lexicales repérées dépend de l'environnement immédiat considéré. Ainsi, pour un environnement de taille $F$, un ensemble de $N$-grams positionnels tels que $N = 2..2F$ peut être extrait. Lors de notre première expérience i.e. à partir du corpus de 200000 formes graphiques, nous avons donc repéré un ensemble de 2-grams, 3-grams, 4-grams, 5-grams et 6-grams candidats à association lexicale. A partir de ces données, il est facile de déterminer la taille moyenne d'une association lexicale candidate et de calculer la distribution des $N$-grams extraits suivant le nombre de formes graphiques qu'ils contiennent. Nous rappellerons que tous les $N$-grams non contigus ont été considérés comme contenant exactement $N$ formes graphiques. Nous n'avons donc pas pris en compte le nombre de mots qui peuvent intervenir à l'intérieur de la séquence non contiguë. La taille moyenne — TM — d'une association lexicale candidate a donc été calculée selon la formule énoncée dans l'équation 9.1 où $NI$ tel que $I = 2, 3, ..., 2F$ correspond au nombre de $I$-grams qui ont été extraits et NB au nombre total de $N$-grams repérés.

$$TM = \frac{N2 \times 2 + N3 \times 3 + ... + N(2F) \times (2F)}{NB} \tag{9.1}$$

Nous proposons donc dans le tableau 9.2 le récapitulatif du calcul de la taille moyenne des associations lexicales candidates telles qu'elles ont été extraites par le GenLocalMaxs associé à l'Expectative Mutuelle.

| Français | Portugais |
|---|---|
| 3.56 | 3.59 |

TAB. 9.2 – Taille Moyenne des Associations Lexicales

Comme prévu, les résultats présentés dans le tableau 9.2 ne correspondent pas à ceux proposés par J. Justeson et S. Katz pour l'Anglais. Alors que ces derniers défendent que la majorité des associations lexicales contiennent deux formes graphiques, les tailles moyennes calculées pour le Français et le Portugais mettent en évidence le fait que la plupart des unités polylexicales candidates sont des 3-grams. Ces résultats vont directement à l'encontre des analyses proposées par B. Daille et M.L. Herviou-Picard et sont particulièrement faciles à expliquer. En effet, alors qu'un nombre important d'associations lexicales de l'Anglais n'utilisent pas la conjonction par préposition des deux termes pleins qui le composent, cette caractéristique est pratiquement inévitable pour le Français et le Portugais. Ainsi, on parlera de *Droits de l'Homme* et de *Direitos do Homem* par opposition au nom composé sans préposition en Anglais *Human Rights*.

Ces résultats ne peuvent toutefois pas être considérés comme une vérité absolue. En effet, cette caractéristique pourrait être propre à l'utilisation de l'Expectative Mutuelle et pourrait ne pas représenter la réalité du phénomène de figement. Pareillement, pour nous assurer de ce fait, nous devrions croiser nos résultats avec ceux obtenus à partir de textes en Anglais. Dans ces conditions, nous avons montré que le GenLocalMaxs continue à élire de préférence des 3-grams bien que la différence entre 3-grams et 2-grams soit plus faible que pour le Portugais et le Français. Il paraîtrait donc que notre architecture tende à favoriser l'élection de suites contenant trois formes graphiques. Cependant, nous verrons dans la prochaine partie que les résultats obtenus à partir de l'Expectative Mutuelle sont particulièrement encourageants par comparaison aux autres mesures d'association. En effet, la taille moyenne calculée pour l'Expectative Mutuelle est celle qui dévie le moins de la taille moyenne absolue calculée à partir de l'ensemble des six mesures d'association.

Avant de passer à cette analyse, il est important de vérifier quel a été le comportement de notre architecture par rapport aux deux langues utilisées. Les résultats montrent des similitudes prévisibles. Nous proposons donc dans la figure 9.1, la distribution des $N$-grams élus selon le nombre de formes graphiques qu'ils contiennent pour le Portugais et le Français.

La similitude entre le Français et le Portugais est flagrante. La courbe d'extraction se comporte exactement de la même manière montrant un maximum d'associations lexicales

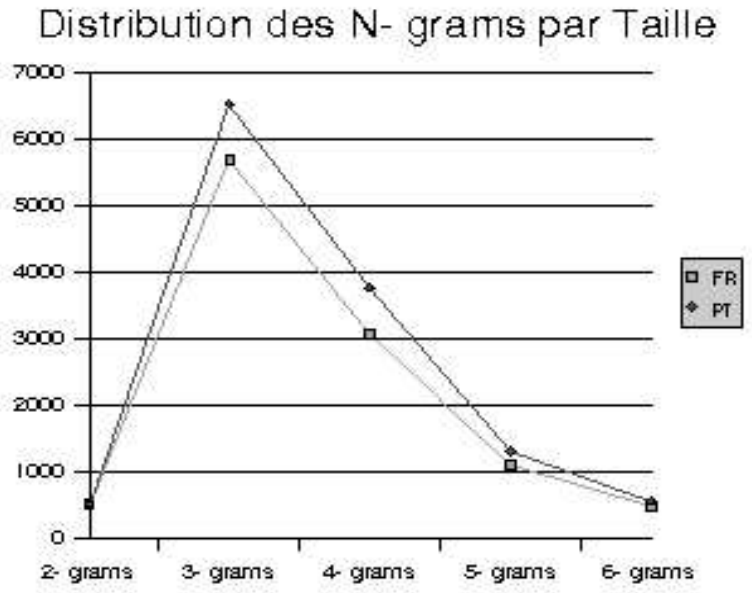


FIG. 9.1 – Distribution des $N$-grams par Taille

de trois formes graphiques. Malgré tout, il est possible de vérifier que le nombre d'associations lexicales extraites pour le Portugais est légèrement plus important que pour le Français. Nous verrons cependant que cette caractéristique n'est pas généralisable et que ces résultats doivent être imputés aux caractéristiques propres du texte traité.

### 9.1.2 Deuxième Expérience

Les résultats que nous avons présentés précédemment sont bien évidemment contestables. En effet, rien ne nous garantit qu'une association lexicale candidate contienne en moyenne un peu plus de trois unités textuelles. Cette caractéristique pourrait être due soit au texte en question, soit à l'environnement immédiat choisi, soit à la mesure d'association utilisée, soit à la langue utilisée comme nous l'avons déjà remarqué. D'autres tests sont par conséquent nécessaires pour s'assurer du bien fondé de ces résultats préliminaires. C'est ce que nous nous proposons de faire ici.

Ainsi, à partir de notre deuxième expérience sur un corpus d'environ 30000 formes graphiques, nous avons mené le même type d'analyse que précédemment afin de déterminer la taille moyenne d'une association lexicale candidate ainsi que la distribution des $N$-grams élus selon leur taille, et ceci, pour chacune des six mesures d'association. Ainsi, pour un environnement immédiat de taille 5, les tableaux 9.3 et 9.4 mettent en évidence les valeurs

de la taille moyenne d'une unité polylexicale candidate pour le coefficient d'association [27], le coefficient Dice [45], la Probabilité Conditionnelle Symétrique [30], le test $\Phi^2$ [29], le coefficient de vraisemblance LogLike [28] et bien évidemment l'Expectative Mutuelle. On remarquera que dans le tableau 9.3, seuls les $N$-grams tels que $N = 2..6$ ont été pris en compte alors que dans le tableau 9.4, tous les $N$-grams i.e. $\forall N, N = 2..10$ ont été comptés. Ceci nous permettra d'analyser ces nouveaux résultats comparativement à ceux que nous avons mis en évidence dans l'expérience précédente.

| Mesure | Français | Portugais |
|---|---|---|
| ME | 3.49 | 3.40 |
| I | 3.75 | 3.67 |
| SCP | 3.87 | 3.58 |
| PHI | 3.59 | 3.42 |
| DICE | 2.75 | 2.60 |
| LOGLIKE | 2.88 | 2.76 |

TAB. 9.3 – Taille Moyenne des Associations Lexicales pour $N = 2..6$

| Mesure | Français | Portugais |
|---|---|---|
| ME | 3.72 | 3.58 |
| I | 3.96 | 3.85 |
| SCP | 4.17 | 3.84 |
| PHI | 3.89 | 3.66 |
| DICE | 3.05 | 2.82 |
| LOGLIKE | 3.01 | 2.88 |

TAB. 9.4 – Taille Moyenne des Associations Lexicales pour $N = 2..10$

A partir des données présentées dans le tableau 9.3, la première remarque qu'il convient de mentionner tient au fait que le GenLocalMaxs associé à l'Expectative Mutuelle démontre une certaine stabilité par rapport à la taille du corpus. En effet, la taille moyenne varie peu entre la première et la deuxième expérience que nous avons réalisées. Cette donnée est particulièrement intéressante suite aux différentes "attaques" qui ont été faites aux méthodes purement statistiques suivant lesquelles ces dernières démontreraient

des résultats insatisfaisants lorsque la taille du corpus diminue. Or, en ce qui concerne la taille moyenne d'extraction, ceci ne semble pas se vérifier puisque la taille moyenne se montre particulièrement stable relativement à la taille du corpus utilisé. Nous verrons dans la suite de cette étude que notre architecture se distingue par une certaine stabilité face à la taille des énoncés considérés.

Dans un deuxième temps, les résultats sont unanimes en ce qui concerne les différences entre les mesures d'association : différentes mesures proposent différentes tailles moyennes. Dans le cadre des données du tableau 9.3, la taille moyenne d'une association lexicale candidate, en considérant toutes les mesures d'association, serait de 3.38 pour le Français et 3.23 pour le Portugais c'est-à-dire très proche des trois unités textuelles. Parallèlement, le tableau 9.4 montre qu'agrandir la taille de l'environnement immédiat influe légèrement sur les valeurs moyennes. Ainsi, pour le Français, la taille moyenne d'une unité polylexicale serait de 3.63 et de 3.43 pour le Portugais. Un nombre non marginal d'associations lexicales candidates de taille élevée a donc été repéré ce qui montre tout l'intérêt de ne pas définir *a priori* une longueur idéale pour une association lexicale.

Finalement, il apparaît clairement que les valeurs exposées par l'Expectative Mutuelle pour le Français et pour le Portugais sont très proches de la valeur moyenne absolue. Nous confirmons cette affirmation à partir des tableaux 9.5 et 9.6 qui illustrent la différence entre la taille moyenne calculée pour chaque mesure d'association et la taille moyenne absolue calculée à partir de l'ensemble des heuristiques.

| Mesure | Français | Portugais |
|---|---|---|
| ME | 0.11 | 0.17 |
| I | 0.37 | 0.44 |
| SCP | 0.49 | 0.35 |
| PHI | 0.21 | 0.19 |
| DICE | 0.63 | 0.63 |
| LOGLIKE | 0.50 | 0.47 |

TAB. 9.5 – Distance par rapport à la Taille moyenne pour $N = 2..6$

Les distances qui séparent les mesures d'association de la valeur moyenne d'une unité

| Mesure | Français | Portugais |
|---|---|---|
| ME | 0.09 | 0.15 |
| I | 0.33 | 0.42 |
| SCP | 0.54 | 0.41 |
| PHI | 0.26 | 0.23 |
| DICE | 0.58 | 0.61 |
| LOGLIKE | 0.62 | 0.55 |

TAB. 9.6 – Distance par rapport à la Taille moyenne pour $N = 2..10$

polylexicale candidate sont les plus faibles pour l'Expectative Mutuelle, et ceci dans tous les cas de figures. Cet indice est particulièrement intéressant car il nous permet de penser que l'Expectative Mutuelle est capable de proposer un ensemble diversifié d'associations lexicales au contraire des autres mesures dont le comportement d'extraction est fortement biaisé. Par exemple, le coefficient Dice et le coefficient de vraisemblance LogLike se caractérisent par la négative montrant régulièrement des résultats d'extraction distants de la moyenne absolue. Ainsi, ces heuristiques tendent à élire principalement des digrams positionnels. Il convient donc de comparer plus en détail le comportement de chaque mesure d'association. Nous proposons alors les résultats de l'analyse de la distribution des $N$-grams élus selon leur taille tant pour le Français — voir Figure 9.2 — que pour le Portugais — voir Figure 9.3.

Cette analyse plus détaillée des résultats d'extraction permet de distinguer deux grands groupes de mesures d'association. En effet, il est clair que le coefficient de vraisemblance LogLike et le coefficient Dice démontrent un comportement similaire extrayant un nombre particulièrement important de 2-grams par rapport aux autres tailles disponibles. On se doit ici d'interrompre notre raisonnement pour remarquer que le coefficient de vraisemblance LogLike a régulièrement été considéré par la communauté scientifique anglo-saxonne comme la mesure d'association de prédilection. Le fait que le coefficient de vraisemblance LogLike élise préférablement des associations lexicales de petite taille n'est certainement pas étranger à ce fait. En effet, cette caractéristique s'accorde parfaitement avec la réalisation des unités polylexicales de l'Anglais. Plus étonnant est le fait que B. Daille propose cette heuristique pour le Français. Une analyse précise de sa méthode

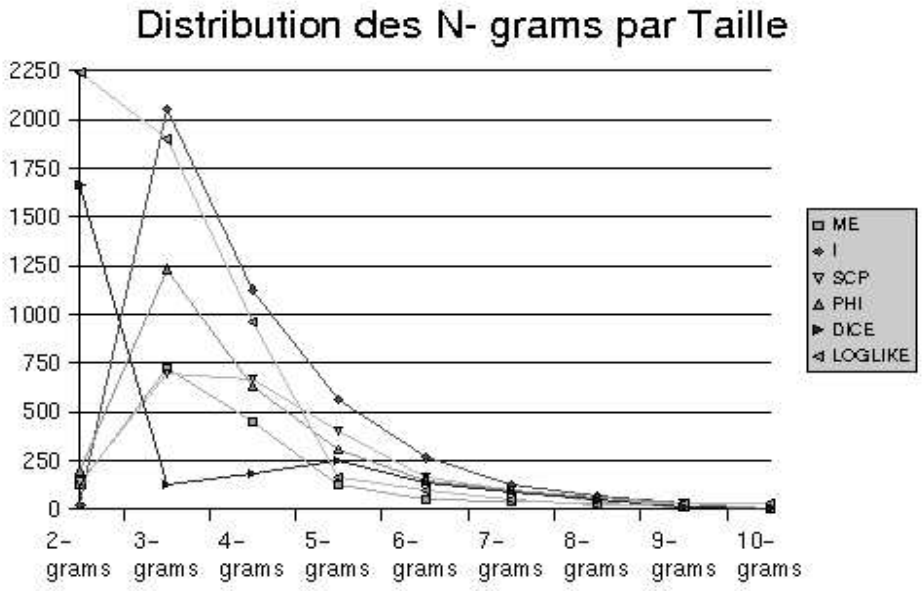


FIG. 9.2 – Distribution de la Taille pour le Français

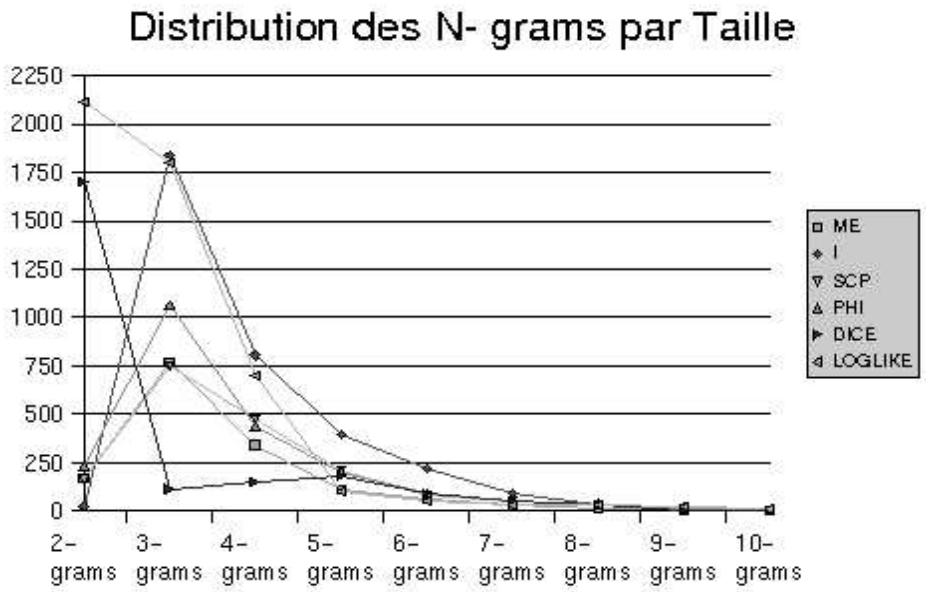


FIG. 9.3 – Distribution de la Taille pour le Portugais

de calcul peut facilement élucider ce "mystère". En effet, toutes les séquences retenues dans une première étape sont préalablement transformées en paires de termes pleins afin d'évaluer leur cohésion. Il est par conséquent facile de vérifier que le coefficient de vraisemblance LogLike est à même de proposer des résultats comparativement plus intéressants.

Revenant à notre idée directrice, le second groupe de mesures d'association est formé par l'Expectative Mutuelle, le coefficient d'association, la Probabilité Conditionnelle Symétrique et le test $\Phi^2$. En particulier, ce groupe met en évidence l'extraction préférentielle de 3-grams et de 4-grams positionnels par opposition aux deux heuristiques précédentes.

Cependant, bien que les courbes de chacun des groupes aient la même forme, des différences existent entre chacune d'entre elles. Par exemple, le coefficient d'association n'extrait pratiquement pas de 2-grams — 19 pour le Français et 29 pour le Portugais — alors que l'Expectative Mutuelle, la Probabilité Conditionnelle Symétrique et le test $\Phi^2$ élisent respectivement 124, 141 et 192 associations candidates pour le Français et 169, 167 et 206 pour le Portugais. Parallèlement, le coefficient d'association élit un nombre très important de 3-grams, 4-grams, 5-grams, 6-grams et 7-grams comparativement aux autres mesures de son groupe. Nous verrons dans la prochaine section que cette aptitude est la conséquence de la forte sensibilité du coefficient d'association aux $N$-grams de faible fréquence. En effet, plus la taille d'un $N$-gram augmente, plus sa fréquence est faible et plus il est probable que le coefficient d'association le repère.

Finalement, comme nous l'avions déjà annoncé dans la partie précédente, il n'existe pas de relation entre la taille moyenne des associations lexicales repérées et les langues considérées. En effet, comparativement à ce qui se passait dans l'expérience précédente, le corpus du Français met en évidence une taille moyenne supérieure au corpus du Portugais. Les résultats obtenus dépendent donc du corpus utilisé et non pas de la langue.

## 9.2 Analyse de la Fréquence des Associations Lexicales

Encore une fois, très peu d'études ont été réalisées sur la fréquence des associations lexicales. Du point de vue des études statistiques, cette donnée peut même être considérée comme un tabou. En effet, la plupart des systèmes préconisent l'utilisation de valeurs seuil. C'est le cas de F. Smadja [31] qui propose de n'élire que des suites de formes graphiques dépassant 50 occurrences. Cette valeur est considérablement élevée et ne correspond certainement pas à la réalité du phénomène de figement.

Comme nous l'avons déjà mentionné, plus une suite de formes graphiques fortement liées est fréquente, plus il est probable que celle-ci soit une association lexicale correcte. Cette caractéristique des unités polylexicales n'est pas nouvelle et a remarquablement été mise en évidence par B. Daille [21]. Cependant, cette donnée a plus à voir avec la précision d'une architecture donnée qu'autre chose. Ainsi, c'est la certitude qu'une séquence d'unités textuelles soit une association lexicale qui est mesurée et moins le fait qu'une association lexicale doive nécessairement démontrer une fréquence élevée. Nous croyons que, comme pour les mots simples, il existe une grande proportion d'unités polylexicales dont la fréquence est relativement faible dans le corpus. Par exemple, il est clair qu'un article de journal est relativement court mais contient néanmoins une foison d'expressions figées. Si l'on se réfère aux résultats du parseur de B. Daille illustrés dans son article [21], un bon nombre des suites de formes graphiques formées par les patrons syntaxiques $N \quad ADJ$ et $N \quad PREP \quad (DET) \quad N$ n'apparaissent qu'une seule ou deux fois dans les deux corpora *Satellite Communication Handbook* et *Communication Blue Book*. Nous rappelons ces résultats dans le tableau 9.7.

| Corpora | *N ADJ* | *N PREP (DET) N* |
|---|---|---|
| Satellite Communication Handbook | | |
| Freq = 1 | 3144 | 6834 |
| Freq = 2 | 655 | 1503 |
| Freq > 2 | 684 | 1616 |
| Communication Blue Book | | |
| Freq = 1 | 5201 | 12167 |
| Freq = 2 | 1507 | 3481 |
| Freq > 2 | 2113 | 6288 |

TAB. 9.7 – Fréquence des Patrons Syntaxiques mesurée par B. Daille

A partir de ces données, il est prévisible qu'un grand nombre d'associations lexicales n'apparaissent qu'une seule voire deux fois dans un corpus donné. En effet, il serait du domaine du miracle si aucune des suites ayant un patron syntaxique reconnu comme propre aux expressions figées et ayant une fréquence relativement faible ne soit pas une unité polylexicale. Nous croyons donc que comme l'a montré G.K.Zipf [74] pour les mots simples, il existe peu d'expressions figées très fréquentes et un nombre élevé d'associations

lexicales peu fréquentes. Nous proposons donc de vérifier ces hypothèses à partir des résultats illustrés par notre architecture pour les deux expériences considérées.

### 9.2.1 Première Expérience

A partir du corpus de 200000 formes graphiques, nous avons d'abord calculé la fréquence moyenne des $N$-grams positionnels repérés par le GenLocalMaxs. Ces résultats sont résumés dans le tableau 9.8.

| Français | Portugais |
|---|---|
| 5.89 | 6.86 |

TAB. 9.8 – Fréquence Moyenne des Associations Lexicales Candidates

A première vue, il semble que l'Expectative Mutuelle tende à élire des suites de formes graphiques dont la fréquence est raisonnablement élevée i.e. pratiquement 6 pour le Français et 7 pour le Portugais. Encore une fois, les résultats entre le Portugais et le Français diffèrent. Nous verrons cependant que ces variations sont propres au texte considéré et non aux langues utilisées. Ces deux valeurs sont toutefois peu informatives. En effet, la moyenne arithmétique est une mesure statistique intéressante mais qui démontre un certain nombre d'inconvénients. En particulier, elle tend à dissimuler les variations entre les éléments de la suite testée. Ainsi, nous détaillons dans le tableau 9.9 la fréquence moyenne des $N$-grams élus selon leur type i.e. pour tout $N = 2..6$.

| N-gram | Français | Portugais |
|---|---|---|
| 2-gram | 68.16 | 97.23 |
| 3-gram | 2.90 | 3.26 |
| 4-gram | 2.52 | 2.84 |
| 5-gram | 2.68 | 3.16 |
| 6-gram | 2.68 | 3.83 |

TAB. 9.9 – Fréquence Moyenne par Type de N-gram

Les résultats sont intéressants. En effet, ils démontrent que la fréquence moyenne est fortement biaisée par la fréquence moyenne des digrams positionnels repérés. Ainsi, pour

tout $N$ tel $N = 3..6$, la fréquence moyenne se rapproche de nos expectatives c'est-à-dire proche de 2 ou 3 occurrences.

Parallèlement, il semblerait que la loi de Zipf soit respectée suggérant que plus une suite de formes graphiques est courte plus elle est fréquente. Cette caractéristique est due à l'occurrence de fragments fonctionnels tels que *de la* pour le Français ou *de um* pour le Portugais. Nous remarquerons également que la fréquence moyenne tend à augmenter à partir des associations lexicales candidates formées de 5 formes graphiques. Une analyse détaillée des résultats montre que cette caractéristique tient du fait que plus une séquence s'allonge plus il est probable d'avoir à faire à un syntagme patron qui comme nous le savons sont récurrents dans les énoncés de la Commission Européenne.

Cependant, nous sommes en droit de nous poser certaines questions en ce qui concerne les fortes fréquences moyennes illustrées dans le cas des digrams et ceci pour les deux langues. En effet, il semblerait à première vue que tous les digrams positionnels repérés soient particulièrement fréquents. Or, ceci n'est pas forcément vrai. Encore une fois, les caractéristiques de la moyenne arithmétique peuvent fausser nos conclusions. Pour être encore plus précis, nous proposons donc une analyse détaillée de la distribution de la fréquence des $N$-grams élus selon leur taille. Les figures 9.4 et 9.5 illustrent cette étude pour le Français et le Portugais.

Comme nous l'espérions, les $N$-grams positionnels apparaissant exactement deux fois ont été préférablement élus et ceci dans tous les cas de figure tant pour le Français que pour le Portugais. Ces résultats sont parfaitement en accord avec nos hypothèses initiales ce qui nous réconforte sur le bien fondé de notre architecture. Ainsi, tant pour le Français que pour le Portugais, environ 80% des associations lexicales candidates apparaissent 2 fois dans le corpus, 10% apparaissent 3 fois et 10% sont utilisées au moins 4 fois.

Les résultats que nous venons de mettre en évidence remettent en cause les techniques qui prônent l'utilisation de valeurs seuil de fréquence élevées comme c'est souvent le cas dans les architectures statistiques qui ne traitent que les textes bruts i.e. sans pré-traitement linguistique. Il est certes vrai que la précision de notre extracteur sera d'autant plus forte que la fréquence d'une suite de formes graphiques sera élevée[3] mais nous verrons

[3]Nous confirmerons cette affirmation dans la suite de notre évaluation

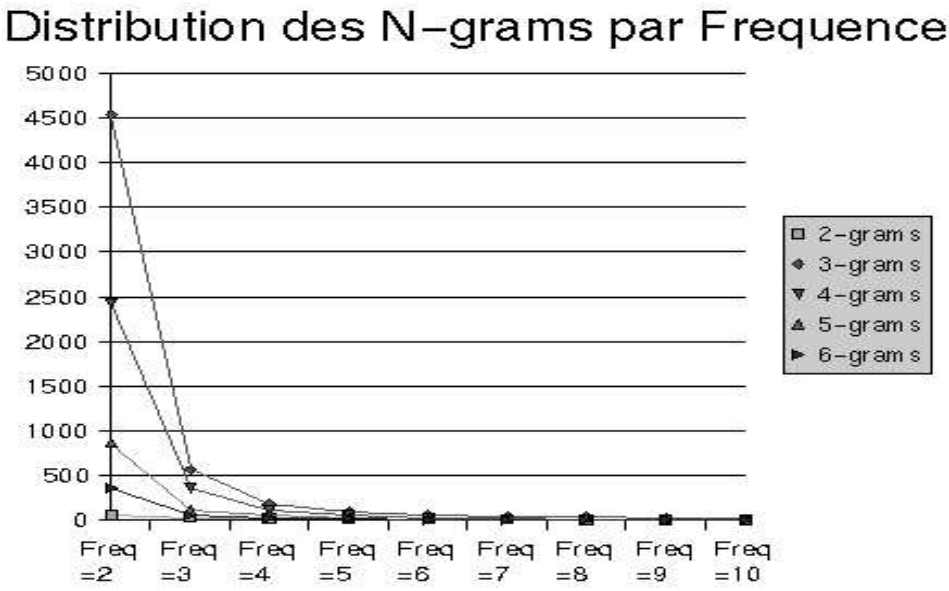


FIG. 9.4 – Distribution selon la Fréquence pour le Français

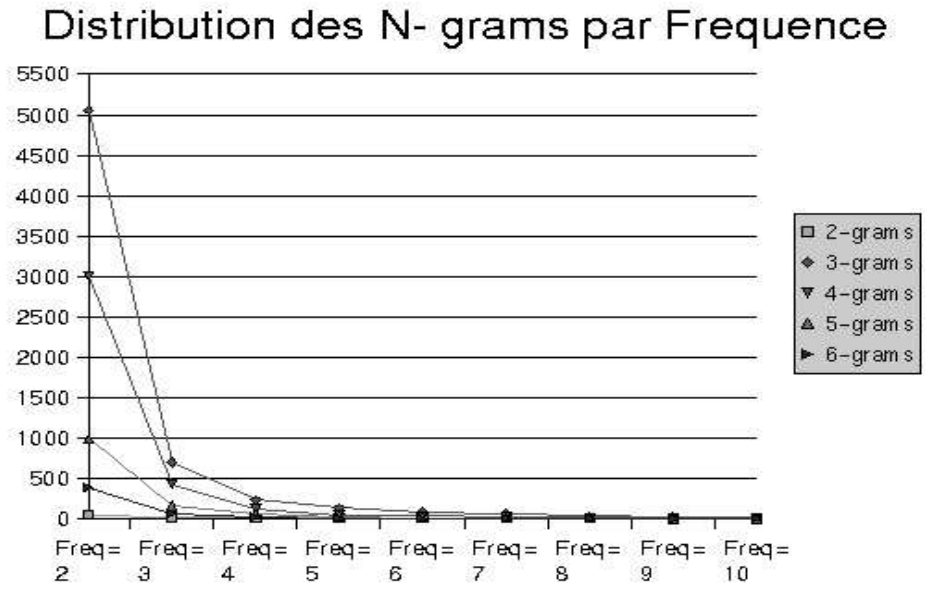


FIG. 9.5 – Distribution selon la Fréquence pour le Portugais

que dans sa globalité notre architecture démontre un fort taux de précision ce qui implique forcément une qualité d'extraction intéressante même pour des fréquences peu élevées.

Comme nous l'avions remarqué dans la partie précédente, ces résultats, bien qu'encourageants, ne peuvent être considérés comme des valeurs de vérité. En effet, il convient

d'analyser le comportement des autres mesures d'association pour dresser un tableau le plus impartial possible sur les valeurs idéales de fréquence des associations lexicales. Nous nous proposons donc d'étudier les résultats obtenus à partir des six mesures d'association normalisées et de les comparer dans la prochaine partie.

### 9.2.2 Deuxième Expérience

Nous proposons d'étudier dans cette partie les indices propres aux fréquences des unités polylexicales calculées à partir du second corpus d'environ 30000 formes graphiques sur lequel nous avons appliqué le GenLocalMaxs associé au coefficient d'association [27], au coefficient Dice [45], à la Probabilité Conditionnelle Symétrique [30], au test $\Phi^2$ [29], au coefficient de vraisemblance LogLike [28] et bien évidemment à l'Expectative Mutuelle. Nous commençons donc par analyser la fréquence moyenne des $N$-grams positionnels qui ont été retenus. Ces résultats sont résumés dans le tableau 9.10.

| Mesure | Français | Portugais |
|---|---|---|
| ME | 4.87 | 4.16 |
| I | 2.19 | 2.19 |
| SCP | 3.80 | 3.38 |
| PHI | 3.47 | 3.19 |
| DICE | 7.66 | 7.69 |
| LOGLIKE | 3.74 | 3.74 |

TAB. 9.10 – Fréquence Moyenne des Associations Lexicales Candidates

La première remarque qu'il convient de mettre en évidence est relative à l'Expectative Mutuelle qui comme prévu démontre son aptitude à s'adapter au texte considéré. En effet, elle extrait en moyenne des associations lexicales candidates de plus faible fréquence à partir du deuxième corpus qu'à partir de l'énoncé de 200 000 formes graphiques. Ceci n'est pas étonnant en soi, bien au contraire. Simplement, il est important de mentionner cette caractéristique qui pourrait ne pas être vérifiée.

Dans un deuxième temps, même si la moyenne arithmétique souffre de certains inconvénients, elle permet néanmoins de mettre en évidence quelques caractéristiques propres à chacune des mesures d'association et ainsi de guider notre analyse. A partir des

résultats du tableau 9.10, il est clair que les différentes heuristiques analysées suggèrent des caractéristiques différentes d'extraction. En effet, il est fort probable qu'il existe une énorme différence de comportement entre le coefficient d'association et le coefficient Dice. Nous démontrerons que cette première analyse se vérifie effectivement. En effet, le coefficient d'association tend à favoriser l'extraction de $N$-grams n'apparaissant que deux fois dans le corpus alors que le coefficient Dice démontre un comportement atypique élisant des unités plus fréquentes. Nous nous proposons donc d'analyser plus en détail l'ensemble des indices de fréquence mis en évidence par l'ensemble des heuristiques testées.

Comme nous l'avons déjà remarqué, les valeurs proposées par les différentes mesures d'association sont particulièrement disparates. Ainsi, parallèlement à ce que nous avons fait précédemment pour l'analyse de la taille des associations lexicales candidates, nous proposons de calculer la distance qui sépare chaque valeur mise en évidence par une mesure d'association donnée de la fréquence moyenne absolue calculée à partir de toutes les heuristiques. Ainsi, en considérant toutes les mesures d'association, la fréquence moyenne absolue d'une unité polylexicale candidate serait de 4.28 pour le Français et 4.05 pour le Portugais. Nous présentons dans le tableau 9.11 les différentes distances qui séparent la fréquence moyenne d'une mesure d'association donnée de la fréquence moyenne absolue.

| Mesure | Français | Portugais |
|---|---|---|
| ME | 0.59 | 0.11 |
| I | 2.09 | 1.86 |
| SCP | 0.48 | 0.67 |
| PHI | 0.81 | 0.86 |
| DICE | 3.38 | 3.64 |
| LOGLIKE | 0.54 | 0.31 |

TAB. 9.11 – Distance par rapport à la Fréquence Moyenne

Le coefficient d'association et le coefficient Dice sont les deux heuristiques qui se distinguent par la négative de l'ensemble des mesures d'association. En effet, elles dévient le plus de l'axe moyen dressé par l'ensemble des heuristiques en compétition : négativement pour le coefficient d'association et positivement pour le coefficient Dice. En ce qui concerne les autres mesures, les résultats sont raisonnablement similaires. On notera

cependant que le test $\Phi^2$ tend à extraire des unités moins fréquentes que la moyenne. En ce qui concerne l'Expectative Moyenne, les résultats sont plutôt encourageants. En effet, pour le Français, même si elle est la troisième mesure la plus distante, elle reste très proche des deux "meilleures" heuristiques. Pour le Portugais, l'Expectative Mutuelle démontre le meilleur comportement i.e. le plus proche de la fréquence moyenne absolue. Ainsi, parallèlement à ce qui s'est passé pour la taille des expressions figées, l'Expectative Mutuelle suggère une certaine capacité à proposer un ensemble d'associations lexicales diversifié et représentatif de la réalité du phénomène de figement.

Malgré leur intérêt évident, ces résultats manquent de précision. En effet, comme nous l'avons déjà mentionné, la simple moyenne arithmétique n'est pas suffisante pour rendre compte des subtilités des données exposées. Nous proposons donc une étude détaillée de la distribution des $N$-grams positionnels selon leur fréquence, ceci tant pour le Français que pour le Portugais. Mais, contrairement à la partie précédente, et afin de ne pas surcharger notre exposé, nous ne détaillerons pas cette distribution par type de $N$-gram. En effet, ces données seront suffisantes pour bien comprendre les différents comportements qui caractérisent chacune des mesures d'association. Les figures 9.6 et 9.7 illustrent ainsi cette étude.

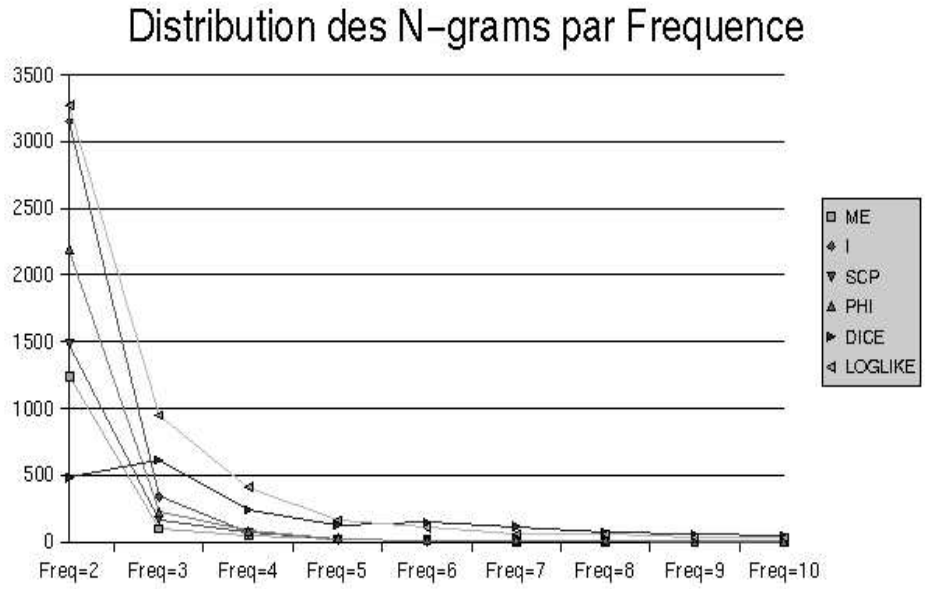


FIG. 9.6 – Distribution selon la Fréquence pour le Français

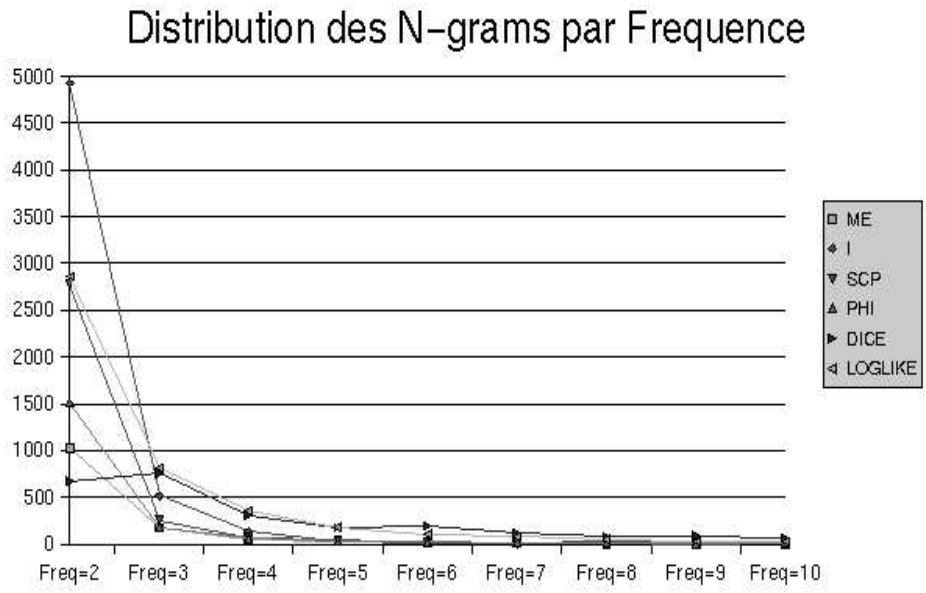


FIG. 9.7 – Distribution selon la Fréquence pour le Portugais

La première remarque importante est relative au comportement atypique de la courbe du coefficient Dice. En effet, alors que toutes les mesures d'association élisent préférablement des associations lexicales candidates de fréquence égale à deux unités, le coefficient Dice démontre une préférence pour les unités polylexicales apparaissant trois fois dans le corpus. Afin de mieux comprendre ce qui se passe réellement, nous proposons donc d'analyser ces distributions par type de $N$-gram. Les deux figures 9.8 et 9.9 présentent ces résultats respectivement pour le Français et le Portugais.

L'analyse des courbes de distribution montre en fait que seule l'extraction des 2-grams positionnels candidats présente un comportement atypique. En effet, pour le cas des $N$-grams positionnels tels que $N$=3..10, le coefficient Dice élit généralement un maximum d'associations lexicales de fréquence égale à deux. Cependant, ceci n'est pas vrai pour le cas des 2-grams. En effet, dans ce cas, le coefficient Dice présente la particularité d'élire préférablement les associations lexicales candidates qui apparaissent trois fois dans le corpus. Or, comme le nombre de 2-grams positionnels repérés par le coefficient Dice est particulièrement élevé par rapport aux autres types de $N$-grams, cette caractéristique tend à se répercuter fortement sur les valeurs moyennes.

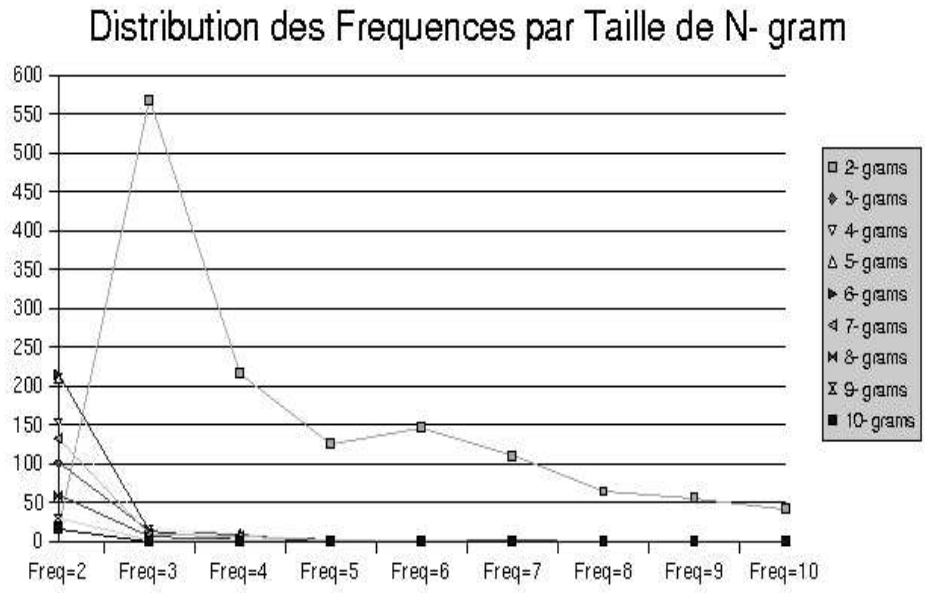


FIG. 9.8 – Distribution selon la Fréquence pour le Français

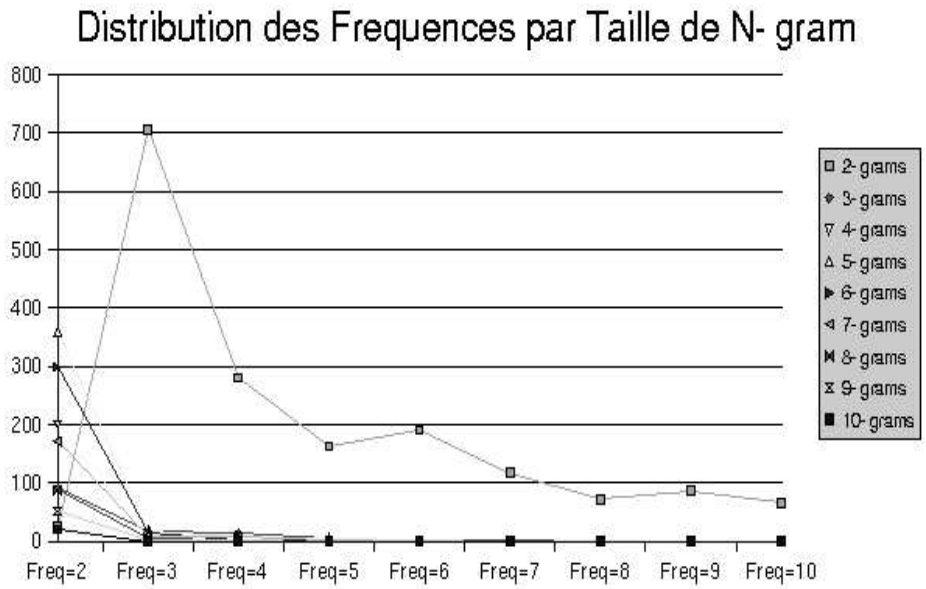


FIG. 9.9 – Distribution selon la Fréquence pour le Portugais

Parallèlement, les figures 9.6 et 9.7 montrent un certain nombre de caractéristiques propres à chacune des heuristiques comparées. En effet, même si les courbes présentent des formes similaires, il existe néanmoins des différences qu'il convient de noter. La première concerne l'Expectative Mutuelle qui extrait le plus petit nombre d'associations lexicales candidates alors que le coefficient de vraisemblance LogLike et le coefficient

d'association sont les plus prolifiques. Nous reviendrons sur ces données dans l'étude de la performance de notre système.

Finalement, comme nous l'avions déjà annoncé dans la partie précédente, il n'existe pas de relation entre la fréquence moyenne des associations lexicales repérées et les langues considérées. En effet, dans cette dernière expérience, la fréquence moyenne d'une association lexicale est supérieure pour le Français comparativement au Portugais. Encore une fois, les résultats obtenus dépendent donc du corpus utilisé et aucune relation entre la fréquence moyenne et la langue considérée ne peut être établie.

Nous procédons maintenant à l'analyse de la diversité des associations lexicales extraites. Nous analyserons en particulier les quantités de $N$-grams élus selon leur type i.e. contigus ou non contigus.

## 9.3 Analyse de la Diversité des Associations Lexicales

Après avoir étudié la taille et la fréquence des associations lexicales élues par notre architecture, nous analysons dans cette partie les données relatives à leur type c'est-à-dire suivant qu'elles représentent des suites lexicales contiguës ou non contiguës. Le fait que peu d'études linguistiques se soient intéressées à ce problème implique de notre part une certaine attention par rapport aux conclusions à dresser. Nous aborderons donc cette partie comme une analyse exploratoire et non pas comme la recherche d'un patron prédéfini. Ainsi, nous essaierons d'éclairer le lecteur sur le genre de résultats qu'il est en droit d'espérer de notre système.

### 9.3.1 Première Expérience

A partir du corpus de 200000 formes graphiques, nous avons étudié les proportions de suites complexes contiguës et non contiguës qui ont été extraites. Les résultats sont, somme toute, étonnants comme le démontre le tableau 9.12.

En effet, il apparaît qu'une partie importante des associations lexicales sont des suites non contiguës de formes graphiques. Ceci est particulièrement remarquable en tenant compte de l'analyse qualitative que nous avons effectuée précédemment. En effet, dans ce cadre, la plupart des exemples retenus sont des suites non contiguës. Or, les résultats

| Type | Français | Portugais |
|---|---|---|
| Contiguës | | |
| Nombre de Contiguës | 4257 | 4061 |
| Nombre Total de N-grams élus | 10819 | 12670 |
| Pourcentage | 39.35% | 32.06% |
| Non Contiguës | | |
| Nombre de Non Contiguës | 6562 | 8609 |
| Nombre Total de N-grams élus | 10819 | 12670 |
| Pourcentage | 60.65% | 67.94% |

TAB. 9.12 – Proportion des Associations Lexicales selon leur Type

d'extraction suggèrent le contraire. Une attention toute particulière devra donc être portée lors du calcul de la performance de notre architecture. En effet, il faudra vérifier la véracité de ces résultats face aux caractéristiques propres des associations lexicales. Il convient donc d'approfondir l'analyse de ces phénomènes pour mieux guider notre étude.

Les séquences non contiguës sont constituées de diverses interruptions. Nous rappelons qu'une interruption correspond à l'occurrence d'au moins deux formes graphiques pour une position donnée du $N$-gram positionnel considéré. Or, pour un environnement immédiat de taille 3, un $N$-gram positionnel peut mettre en évidence une, deux, trois ou quatre interruptions au maximum. Nous avons donc recherché à illustrer la distribution des suites non contiguës selon leur nombre d'interruptions. Les résultats sont résumés dans le tableau 9.13.

| Nb d'Interruptions | Français | Portugais |
|---|---|---|
| 1 | 34.99% | 28.96% |
| 2 | 44.38% | 49.59% |
| 3 | 16.07% | 16.52% |
| 4 | 4.56% | 4.93% |

TAB. 9.13 – Distribution des Associations Lexicales non Contiguës selon le Nombre d'Interruptions

Les résultats obtenus sont plutôt encourageants. En effet, la plupart des suites non contiguës contiennent soit une soit deux interruptions alors que les séquences contenant trois et quatre interruptions sont moins nombreuses. Ce comportement semble être en adéquation avec les caractéristiques du Français et du Portugais qui utilisent peu les associations lexicales distantes comparativement à ce qui se passe pour l'Anglais ou l'Allemand. Ainsi, il est fort probable qu'un nombre significatif de $N$-grams élus soient réellement des associations lexicales correctes. D'ailleurs, les exemples choisis pour l'analyse qualitative vont dans ce sens.

Nous devons cependant prendre certaines précautions. En effet, cette capacité de notre extracteur à extraire des associations lexicales non contiguës en grand nombre est un fait inespéré qu'il convient d'analyser en détail. Nous nous attacherons donc à comprendre les caractéristiques propres des unités non contiguës relativement aux associations lexicales contiguës. Dans ce cadre, nous proposons d'analyser la fréquence des $N$-grams positionnels selon leur type. Les résultats sont résumés dans les figures 9.10 et 9.11.

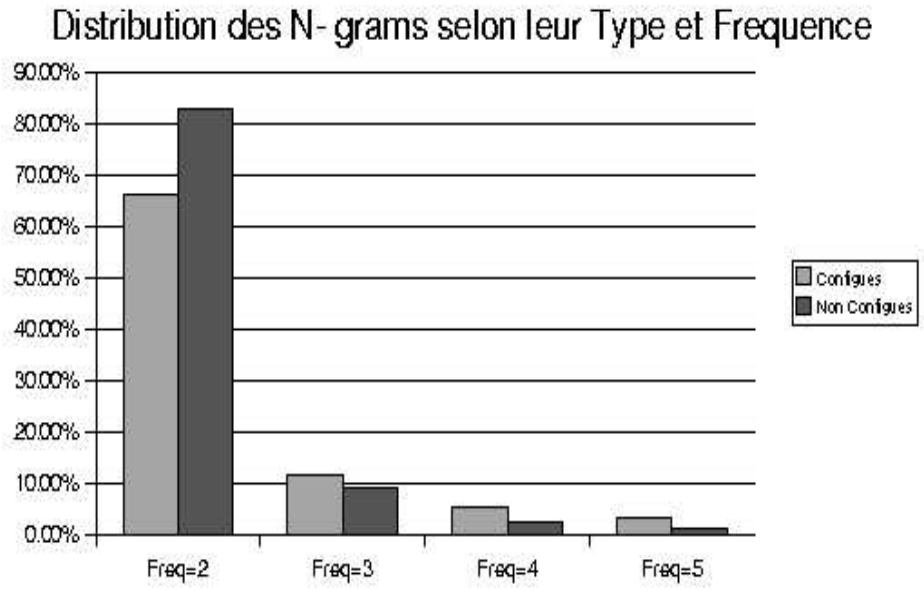


FIG. 9.10 – Distribution selon le Type pour le Français

Les résultats sont clairs. En effet, alors que pour les fréquences supérieures à 2 les suites non contiguës sont moins nombreuses que les séquences contiguës, cette situation s'inverse

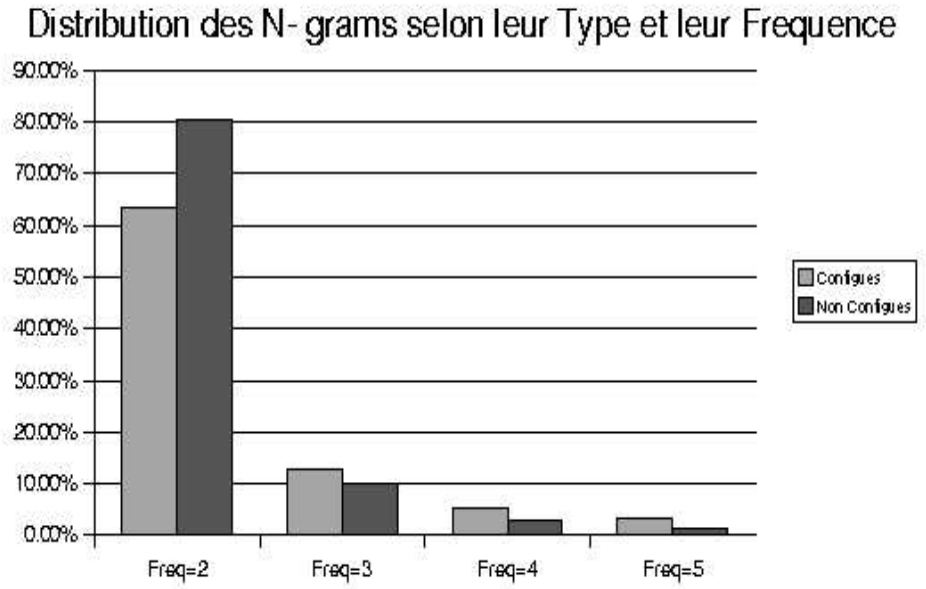


FIG. 9.11 – Distribution selon le Type pour le Portugais

pour une fréquence de deux. Plusieurs remarques peuvent être faites sur ce phénomène. En particulier, le fait de travailler avec des positions fixes implique que deux associations lexicales contenant les mêmes formes graphiques mais à des positions différentes soient comptées séparément. Or, cette remarque est tout à fait pertinente pour le cas des suites non contiguës. En effet, le propre d'une association lexicale non contiguë est de permettre une certaine flexibilité et par conséquent l'usage de différentes positions. Par exemple, l'association lexicale *recourir à* peut tout à fait être identifiée suivant différentes formes. Ainsi, les deux $N$-grams positionnels [*0 recourir 1 à*] et [*0 recourir 2 à*] peuvent être élus représentant respectivement des suites telles que *recourir forcément à* et *recourir très souvent à*. Par conséquent, il n'est pas surprenant que les fréquences des suites non contiguës soient plus faibles en moyenne que celles des suites contiguës.

Le revers de la médaille est bien entendu le fait que plus la fréquence d'un $N$-gram positionnel est faible, moins il est probable que celui-ci soit une association lexicale valide. Ainsi, il est prévisible que les résultats de précision des suites non contiguës soient plus faibles que ceux des suites contiguës. Nous verrons si cette hypothèse se confirme dans la prochaine partie. Cependant, afin de ne pas formuler des conclusions trop hâtives, nous proposons une analyse plus profonde des résultats précédents en prenant en compte la diversité des suites non contiguës i.e. le nombre d'interruptions qu'elles contiennent.

En effet, nous voulons vérifier si le phénomène que nous venons de mettre en évidence se répartit de forme égale selon qu'une suite de formes graphiques contient une, deux, trois ou quatre interruptions. Les résultats sont présentés dans les deux figures 9.12 et 9.13.

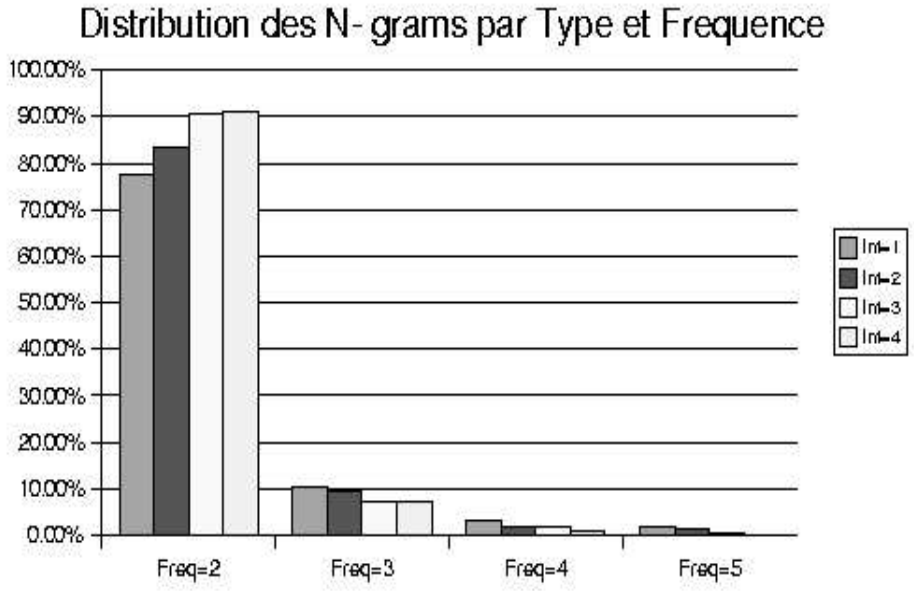


FIG. 9.12 – Distribution selon le Type pour le Français

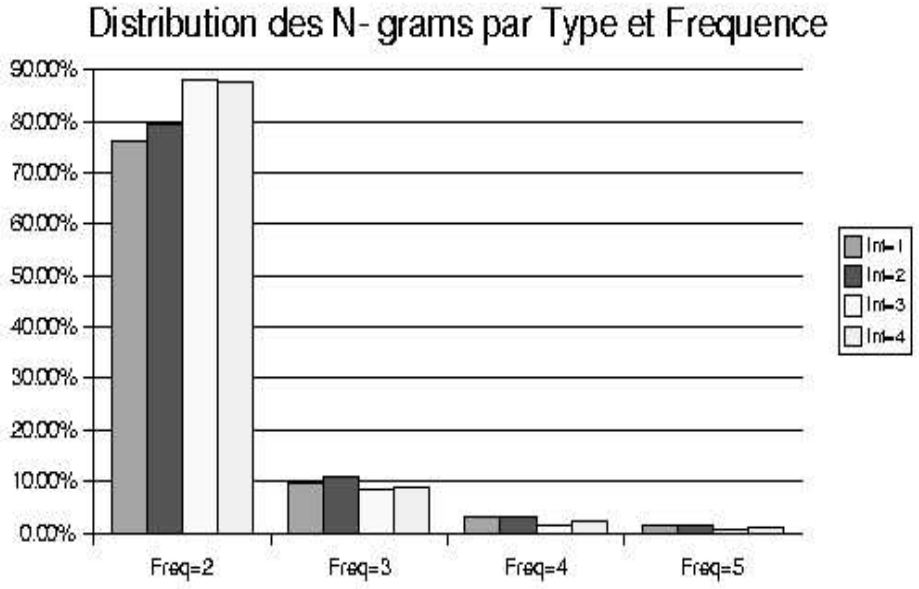


FIG. 9.13 – Distribution selon le Type pour le Portugais

Comme nous l'avons remarqué préalablement, une grande proportion des suites non contiguës contient soit une soit deux interruptions. Or, d'après les valeurs illustrées dans les figures 9.12 et 9.13, ces mêmes suites sont celles dont la fréquence moyenne est la plus élevée. Ceci est de bonne augure sachant que plus il y a d'interruptions dans une suite, plus il est probable que le $N$-gram positionnel candidat ne soit pas une association lexicale correcte. Ainsi, nous sommes à même de croire, d'après leur valeur de fréquence et le nombre d'interruptions qu'elles contiennent, qu'un nombre important non négligeable de suites non-contiguës contenant peu d'interruptions devra illustrer des résultats de précision favorables.

Comme nous l'avons fait dans les études précédentes, nous analysons maintenant les résultats obtenus à partir du corpus de 30000 formes graphiques pour les six mesures d'association considérées.

### 9.3.2 Deuxième Expérience

Comme nous l'avons déjà remarqué dans la partie précédente, un nombre significatif de suites non contiguës met en évidence de "fausses" associations lexicales. En effet, la sous-évaluation des formes graphiques fréquentes comprises dans les $N$-grams positionnels implique généralement la sélection de séquences non complètes. Il est donc fort probable que le pourcentage de suites non contiguës soit particulièrement élevé pour les cinq mesures normalisées par rapport à l'Expectative Mutuelle. Nous confirmons cette hypothèse grâce aux tableaux 9.14 et 9.15.

L'Expectative Mutuelle démontre le pourcentage de suites contiguës le plus élevé de toutes les mesures d'association. En effet, le fait que l'Expectative Mutuelle ne sous-évalue pas les associations contenant des particules fréquentes implique nécessairement des valeurs qui sont le plus proche possible de la réalité des phénomènes de figement. Paradoxalement, le coefficient Dice démontre également un pourcentage non négligeable de suites contiguës bien qu'il mette en évidence les mêmes problèmes que les autres heuristiques normalisées. Cette caractéristique peut être cependant facilement expliquée. En effet, nous avons montré précédemment que le coefficient Dice tend à élire des $N$-grams positionnels qui apparaissent trois fois dans le corpus. Or, nous savons d'après notre étude sur le corpus de 200000 formes graphiques que la plupart des suites non contiguës sélectionnées

| Type | PHI | DICE | LOGLIKE | MI | SCP | ME |
|---|---|---|---|---|---|---|
| Contiguës | | | | | | |
| Nb | 332 | 637 | 2848 | 273 | 321 | 481 |
| Total | 2633 | 2230 | 14935 | 3765 | 2052 | 1423 |
| Perc. | 12.60 | 28.56 | 19.06 | 7.25 | 15.64 | 33.80 |
| Non Contiguës | | | | | | |
| Nb | 2301 | 1593 | 12087 | 3492 | 1731 | 942 |
| Total | 2633 | 2230 | 14935 | 3765 | 2052 | 1423 |
| Perc. | 87.40 | 71.44 | 80.94 | 92.75 | 84.36 | 66.20 |

TAB. 9.14 – Proportion des Associations Lexicales selon leur Type pour le Français

| Type | PHI | DICE | LOGLIKE | MI | SCP | ME |
|---|---|---|---|---|---|---|
| Contiguës | | | | | | |
| Nb | 262 | 641 | 2808 | 402 | 301 | 475 |
| Total | 1930 | 2148 | 14421 | 3588 | 1971 | 1387 |
| Perc. | 13.57 | 29.84 | 19.47 | 11.20 | 15.27 | 34.24 |
| Non Contiguës | | | | | | |
| Nb | 1668 | 1507 | 11613 | 3186 | 1670 | 912 |
| Total | 1930 | 2148 | 14421 | 3588 | 1971 | 1387 |
| Perc. | 86.43 | 70.16 | 80.53 | 88.80 | 84.73 | 65.76 |

TAB. 9.15 – Proportion des Associations Lexicales selon leur Type pour le Portugais

démontrent une fréquence de deux. Ainsi, il est clair que le pourcentage de suites non contiguës retenues par le coefficient Dice est particulièrement faible relativement aux autres mesures à l'exception de l'Expectative Mutuelle.

En ce qui concerne l'Expectative Mutuelle, les valeurs calculées pour ce corpus sont très proches de celles présentées dans l'expérience précédente bien que l'environnement immédiat soit considérablement plus grand. En effet, nous nous attendions à vérifier l'extraction d'un nombre plus important de suites non contiguës. Or, ceci ne s'est pas vérifié ni pour le Français, ni pour le Portugais. Ces résultats sont particulièrement encourageants. En effet, il semblerait que notre architecture se soit montrée capable —

du moins en apparence — de discerner à l'intérieur d'un nombre de $N$-grams positionnels non contigus plus important, un nombre équivalent d'unités potentiellement correctes. Une étude plus complète est bien entendue nécessaire pour confirmer ces premières impressions. Nous commencerons donc par analyser les proportions de suites contiguës suivant le nombre d'interruptions qu'elles contiennent. Ces résultats sont illustrés dans le tableau 9.16. Nous rappelons que dans le cadre d'un environnement immédiat de taille 5, une association lexicale candidate peut contenir un maximum de huit interruptions.

| Nb d'Interruptions | Français | Portugais |
|---|---|---|
| 1 | 19.33% | 15.14% |
| 2 | 12.53% | 14.04% |
| 3 | 18.05% | 20.17% |
| 4 | 21.02% | 21.60% |
| 5 | 13.16% | 12.72% |
| 6 | 8.81% | 8.77% |
| 7 | 5.94% | 5.04% |
| 8 | 1.16% | 2.52% |

TAB. 9.16 – Distribution des Associations Lexicales non Contiguës selon le Nombre d'Interruptions

Contrairement ce à quoi nous nous attendions, la distribution des suites non contiguës ne suit pas celle mise en évidence pour un environnement immédiat de taille 3. En effet, les résultats montrent que la plupart des séquences retenues contiennent entre trois et quatre interruptions. Cette information est loin d'être réconfortante sachant que plus les interruptions sont nombreuses dans une séquence, plus il est probable que celle-ci ne forme pas une association lexicale correcte. Il faudra donc confirmer cette hypothèse lors de notre étude sur les performances de notre architecture.

Parallèlement, il semblerait qu'il existe une relation entre l'environnement immédiat considéré et le type de $N$-grams positionnels extraits. Ainsi, pour un environnement de taille 3, le nombre d'interruptions prédominant se situe entre un et deux alors que pour un environnement de taille 5, ce nombre passe à 3 et 4. Il est donc fort probable, si cette tendance se maintient, que pour un environnement immédiat de taille 4, les résultats

suggèrent l'extraction de suites non contiguës contenant entre 2 et 3 interruptions. Afin de confirmer cette hypothèse, nous avons donc testé notre système sur le même corpus pour un environnement immédiat de taille 4. Les résultats sont présentés dans le tableau 9.17. Nous rappelons que dans ce cas le nombre maximum d'interruptions est de 6.

| Nb d'Interruptions | Français | Portugais |
|---|---|---|
| 1 | 25.40% | 22.77% |
| 2 | 26.60% | 22.19% |
| 3 | 27.00% | 31.08% |
| 4 | 12.04% | 14.70% |
| 5 | 7.06% | 6.90% |
| 6 | 1.90% | 2.36% |

TAB. 9.17 – Distribution des Associations Lexicales non Contiguës selon le Nombre d'Interruptions

Effectivement, nos hypothèses paraissent se confirmer. En effet, pour le Français comme pour le Portugais, une grande partie des suites non contiguës contient entre 2 et 3 interruptions. Cependant, il faut noter que le pourcentage de séquences contenant exactement une interruption est particulièrement important ce qui fausse légèrement nos conclusions anticipées. Ainsi, il faut reconnaître que les résultats obtenus par notre architecture sont biaisés par la valeur de l'environnement immédiat choisi. En effet, il est impossible que les trois distributions illustrées précédemment soient en adéquation avec le phénomène de figement. L'une d'entre elles devra nécessairement se distinguer par rapport aux autres. Or, une première analyse des résultats qualitatifs semble suggérer que plus le nombre d'interruptions est faible, plus il est probable que la suite non contiguë soit une association lexicale correcte. Dans ces conditions, il est fort probable que l'environnement immédiat de taille 3 produise les meilleurs résultats de performance.

Dans cette partie, nous avons régulièrement fait référence à la performance de notre architecture. Ainsi, nous avons atteint le stade de notre étude où l'analyse de la précision et le rappel du système proposés s'imposent. En effet, nous avons besoin de savoir si notre architecture a un comportement moyen plutôt favorable ou défavorable et quels sont les critères qui peuvent influencer ce comportement. Nous nous proposons donc de mener à

bien cette étude dans la partie suivante.

## 9.4 Analyse de la Performance

La performance d'un système d'extraction est généralement évaluée à partir des mesures de précision et de rappel. Dans le premier cas, le taux de précision permet de mesurer l'efficacité du système. On teste alors sa capacité à fournir les données pour lesquelles il a été construit — dans notre cas les associations lexicales. Dans le deuxième cas, le taux de rappel permet de mesurer la couverture de l'architecture. Ainsi, on évalue la proximité des résultats obtenus par l'architecture proposée par rapport à la situation d'extraction idéale c'est-à-dire pour laquelle toutes les données recherchées ont été identifiées.

Le taux de précision et le taux de rappel se basent nécessairement sur le concept de correction des résultats. Ainsi, nous devons dans un premier temps définir les caractéristiques des associations lexicales correctes afin de pouvoir évaluer la performance de notre système. Dans ce cadre, de nombreuses études ont été proposées. Nous en mentionnerons deux, [31] et [32], qui mettent clairement en évidence les problèmes rencontrés pour mesurer correctement les résultats d'extraction. D'une part, F. Smadja [31] propose trois critères d'évaluation. Ainsi, il utilise le sigle YY pour une association lexicale de grande qualité, Y pour une suite de moins bonne qualité mais représentant un concept polylexical et finalement N pour les séquences qui ne sont pas clairement des associations lexicales. Parallèlement, S. Shimohata [32] suggère quatre critères d'évaluation : CS pour une phrase complète, GU pour une unité grammaticale, MU pour une séquence non grammaticale mais dont le sens se rapproche d'un concept et F pour un fragment fonctionnel. Ainsi, toutes les suites de formes graphiques qui ne sont pas cataloguées sous le sigle F sont considérées comme étant des associations lexicales correctes. Il est clair que ces définitions ne sont en aucun cas satisfaisantes. De fait, il existe un fort degré d'incertitude et de flou en ce qui concerne les catégories Y de F. Smadja et MU de S. Shimohata. Nous dirons même que celles-ci peuvent être considérées comme des classes "fourre-tout". De plus, l'analyse de l'évaluateur est forcément subjective et par conséquent conduit nécessairement à des résultats biaisés.

Dans notre évaluation, nous avons essayé d'éviter l'aspect partial et d'en protéger

son aspect intrinsèque. Ainsi, nous avons décidé de guider notre analyse suivant les informations recueillies par G. Gross [51] sur les phénomènes de figement. En effet, G. Gross propose une analyse exhaustive de ce phénomène qui peut réellement servir de base pour un travail d'analyse impartiale. Ainsi, il définit pour chacune des catégories un ensemble de règles qui permettent de vérifier avec précision si une suite de formes graphiques est ou non une association lexicale.

A partir de ce guide, nous pouvons donc légitimement argumenter contre la nécessité d'une analyse croisée externe. En effet, les règles du jeu sont claires. Ainsi, seuls les $N$-grams positionnels dont la structure s'apparente à une règle définie par G. Gross seront considérés comme étant des associations lexicales correctes. Dans ces conditions, la partialité du jugement est considérablement réduite. Une suite de formes graphiques sera donc une association lexicale si elle met en évidence une structure de nom composé, de déterminant composé, de locution verbale, de locution adjectivale ou de locution prépositive ou conjonctive. Afin de rendre compte de la diversité des résultats obtenus, nous ajouterons à ces catégories, toutes les structures qui ont été mises en évidence dans la partie précédente de l'analyse qualitative. En particulier, nous considérerons les adverbes de négation, les suites représentant des conjonctions et les syntagmes patrons.

Finalement, il est particulièrement important de remarquer que les résultats de notre évaluation ne pourront en aucun cas être comparés à ceux de F. Smadja et S. Shimohata du fait de la séparation évidente des définitions de correction. Comme nous l'avons déjà mis en évidence, les résultats que nous présenterons dans cette étude devront donc être analysés avec précaution comparativement à d'autres études moins précises.

Avant de présenter les résultats proprement dits, il nous semble indispensable d'expliquer d'abord comment nous avons mené notre analyse. En effet, il n'est en aucun cas possible de vérifier la correction de chaque $N$-gram positionnel un à un. Pour cette raison, nous avons dû recourir à l'échantillonnage de l'ensemble des suites extraites. Ainsi, pour chaque caractéristique que nous nous sommes fixés d'analyser, nous avons systématiquement calculé trois échantillons aléatoires de même taille à partir de l'ensemble des $N$-grams positionnels extraits. Cependant, il est important de remarquer que du fait du nombre important de tests que nous avons réalisés, la taille de chaque échantillon est parti-

culièrement faible. Nous sommes conscients des problèmes que ces impératifs ont sur la représentativité des résultats, mais comme nous le verrons les valeurs obtenues sont raisonnablement échelonnées de telle forme que les grandes lignes de l'évaluation sont clairement représentées.

Nous passons donc aux premiers résultats. Dans un premier temps, nous avons recherché à calculer le taux de précision et le taux de rappel de notre extracteur dans sa globalité. Ainsi, nous nous sommes attachés à mesurer la performance absolue de notre architecture. Comme nous l'avons déjà mentionné, alors qu'il est relativement simple de calculer le taux de précision d'une extraction donnée — Equation 9.2 —, il est plus difficile de mesurer le taux de rappel.

$$\text{Précision} = \frac{\text{Nb de } N\text{-grams Positionnels extraits correctement}}{\text{Nb de } N\text{-grams Positionnels extraits}} \tag{9.2}$$

En effet, comme il n'existe pas de texte normalisé pour lequel toutes les associations lexicales ont été identifiées, il n'est pas possible de calculer le taux de rappel de notre extracteur. Le lecteur comprendra bien qu'il nous est impossible de relever toutes les associations d'un corpus de 200000 formes graphiques ! Cependant, il est nécessaire de proposer un indicateur qui permette de mesurer le nombre d'expressions que l'on est à même d'espérer du processus d'extraction. Dans ce cadre, nous proposons de calculer la proportion des associations lexicales extraites correctement par rapport à la taille du texte considéré. Nous appellerons cet indicateur : le taux de couverture — Equation 9.3.

$$\text{Couverture} = \frac{\text{Nb de } N\text{-grams Positionnels corrects}}{\text{Nb de formes graphiques du corpus}} \tag{9.3}$$

A partir de ces explications, nous proposons donc les résultats d'extraction de notre architecture à partir du corpus de 200 000 formes graphiques, autant pour le Français que pour le Portugais à partir des tableaux 9.18 et 9.19.

Les taux de précision globale obtenus sont particulièrement faibles par rapport aux chiffres annoncés par les méthodologies hybrides c'est-à-dire qui utilisent des patrons linguistiques prédéterminés associés à des mesures statistiques. Ce résultat n'est cependant pas surprenant. En effet, les études travaillant exclusivement sur les formes graphiques

| | échantillon 1 | échantillon 2 | échantillon 3 | Moyenne |
|---|---|---|---|---|
| Corrects | 143 | 140 | 140 | 141 |
| Total | 270 | 270 | 270 | 270 |
| Précision | 52.96% | 51.85% | 51.85% | 52.22% |
| Couverture | 2.86% | 2.8% | 2.8% | 2.82% |

TAB. 9.18 – Précision et Couverture Globale du Français

| | échantillon 1 | échantillon 2 | échantillon 3 | Moyenne |
|---|---|---|---|---|
| Corrects | 141 | 149 | 155 | 148 |
| Total | 315 | 315 | 315 | 315 |
| Précision | 44.76% | 47.30% | 49.20% | 46.98% |
| Couverture | 2.82% | 2.98% | 3.1% | 2.96% |

TAB. 9.19 – Précision et Couverture Globale du Portugais

n'ont jamais démontré des résultats très précis. A titre d'exemples, F. Smadja prétend atteindre un taux de précision de 40% pour un texte d'un million de mots [31] et S. Shimohata, dans son article [32], présente des valeurs de l'ordre de 64%. Or, comme nous l'avons déjà mentionné, les largesses au niveau des définitions de correction ne permettent pas de comparer directement ces trois chiffres. Néanmoins, elles permettent de donner une notion de l'horizon dans lequel nous travaillons. Donc, plutôt que de tenter comparer notre architecture aux autres systèmes proposés dans la littérature, nous nous attacherons plutôt à comparer la mesure d'Expectative Mutuelle par rapport aux autres mesures d'association normalisées. Ceci fera l'objet d'une attention toute particulière dans la suite de notre exposé. De plus, pour comparer plusieurs systèmes encore faudrait-il qu'ils extraient les mêmes phénomènes ce qui n'est pas le cas.

Parallèlement, les chiffres concernant la couverture de notre système montrent qu'entre 5000 et 6000 associations lexicales correctes peuvent être extraites pour un corpus de 200 000 formes graphiques. Ces résultats mettent clairement en évidence la dure tâche que les lexicographes, traducteurs et autres doivent réaliser. *A fortiori*, il est utile de souligner que notre extracteur n'est capable de repérer que des unités polylexicales qui apparaissent au moins deux fois dans le corpus. Ceci implique bien évidemment que le

nombre d'associations lexicales réellement présentes dans le texte est largement supérieur. Des outils performants sont donc forcément utiles pour ce travail laborieux.

Bien que les résultats globaux soient intéressants, ils cachent souvent des particularités importantes du système analysé. Dans ce contexte, nous nous attacherons dans un premier temps à décortiquer sous tous les angles les résultats obtenus à partir du corpus de 200 000 formes graphiques. Ainsi, nous commencerons par analyser les résultats obtenus selon le type des $N$-grams positionnels extraits dans les tableaux 9.20 et 9.21 c'est-à-dire selon qu'ils sont contigus ou non. Nous verrons que ces résultats seront particulièrement importants pour la suite de nos analyses.

| Type | échantillon 1 | échantillon 2 | échantillon 3 | Moyenne |
|---|---|---|---|---|
| ***Cont.*** | | | | |
| Corrects | 118 | 123 | 135 | 125 |
| Total | 170 | 170 | 170 | 170 |
| Précision | 69.41% | 72.35% | 79.41% | 73.72% |
| Couverture | 1.47% | 1.53% | 1.68% | 1.56% |
| ***Non Cont.*** | | | | |
| Corrects | 53 | 50 | 51 | 51 |
| Total | 164 | 164 | 164 | 164 |
| Précision | 32.31% | 30.48% | 31.09% | 31.29% |
| Couverture | 1.06% | 1.00% | 1.02% | 1.02% |

TAB. 9.20 – Précision/Couverture du Français selon le type de $N$-gram

Les résultats sont clairs. Il existe une grande disparité entre les $N$-grams positionnels contigus et non contigus. En ce qui concerne le taux de précision, il est largement plus élevé pour les unités contiguës comparativement aux suites non contiguës. Ceci provient en particulier du fait que la plupart des $N$-grams positionnels non contigus ont une fréquence de deux ce qui a tendance à diminuer la qualité de l'extraction comme nous le verrons dans la suite de cette analyse. Ces chiffres sont cependant encourageants. En effet, d'une part, les résultats pour les suites contiguës sont particulièrement élevés et amenuisent ainsi les effets négatifs des suites non contiguës sur les résultats globaux. En effet, les unités non contiguës sont les principales responsables des résultats précédents.

| Type | échantillon 1 | échantillon 2 | échantillon 3 | Moyenne |
|---|---|---|---|---|
| ***Cont.*** | | | | |
| Corrects | 138 | 122 | 136 | 132 |
| Total | 161 | 161 | 161 | 161 |
| Précision | 85.71% | 75.77% | 84.47% | 81.98% |
| Couverture | 1.72% | 1.52% | 1.7% | 1.65% |
| ***Non Cont.*** | | | | |
| Corrects | 45 | 37 | 37 | 39 |
| Total | 172 | 172 | 172 | 172 |
| Précision | 26.16% | 21.51% | 21.51% | 23.06% |
| Couverture | 1.12% | 0.92% | 0.92% | 0.99% |

TAB. 9.21 – Précision/Couverture du Portugais selon le type de $N$-gram

L'utilisation des positions comportait effectivement un risque évident dès le départ de notre recherche. En effet, très peu d'études se sont aventurées dans ce sens. Ceci est facilement compréhensible. Il est clair que cette tâche pose un certain nombre de problèmes propres. En particulier, l'utilisation des positions permet l'extraction d'associations entre verbe et préposition qui sont souvent appelées attachements[4]. Or, nous savons que les prépositions sont des fragments fréquents qui compliquent l'analyse des associations lexicales. Ainsi, un bon nombre de $N$-grams positionnels sont extraits anormalement à cause de relations fortuites entre une préposition et le reste des constituants du propre $N$-gram. Il est donc évident que cette tâche qui est souvent analysée librement comporte des risques qui se répercutent sur les résultats globaux. En ce qui concerne le taux de couverture, bien que le nombre de $N$-grams positionnels non contigus soit plus grand que le nombre de suites contiguës, le nombre d'associations lexicales contiguës correctes dépasse largement le nombre de suites non contiguës. Ainsi, il faut espérer de notre architecture qu'elle produise un nombre important d'éléments contigus.

Nous savons qu'une association lexicale non contiguë peut contenir entre 2 et 4 interruptions. Or, comme nous l'avons mentionné dans la partie précédente, il semblerait que la correction des $N$-grams positionnels ne soit pas égale selon le nombre d'interruptions qu'ils contiennent. Nous avons donc analysé ces différences dans les tableaux 9.22 et 9.23.

[4]PP-attachment en Anglais.

| Interruptions | échantillon 1 | échantillon 2 | échantillon 3 | Moyenne |
|---|---|---|---|---|
| *=1* | | | | |
| Corrects | 29 | 25 | 34 | 29 |
| Total | 57 | 57 | 57 | 57 |
| Précision | 50.87% | 43.85% | 59.64% | 51.45% |
| Couverture | 0.58% | 0.5% | 0.68% | 0.58% |
| *=2* | | | | |
| Corrects | 10 | 14 | 16 | 13 |
| Total | 58 | 58 | 58 | 58 |
| Précision | 17.24% | 24.13% | 27.58% | 22.98% |
| Couverture | 0.25% | 0.35% | 0.4% | 0.33% |
| *=3* | | | | |
| Corrects | 4 | 4 | 5 | 4 |
| Total | 52 | 52 | 52 | 52 |
| Précision | 7.69% | 7.69% | 9.61% | 8.33% |
| Couverture | 0.04% | 0.04% | 0.05% | 0.04% |
| *=4* | | | | |
| Corrects | 2 | 3 | 0 | 1 |
| Total | 53 | 53 | 53 | 53 |
| Précision | 3.77% | 5.66% | 0.00% | 3.14% |
| Couverture | 0.005% | 0.008% | 0.00% | 0.004% |

TAB. 9.22 – Précision/Couverture du Français selon les interruptions des $N$-grams

Là encore, les résultats sont particulièrement clairs. Plus le nombre d'interruptions est grand, plus il est probable qu'un $N$-gram positionnel candidat soit une association lexicale incorrecte. La raison de ces résultats est simple à déterminer. En effet, plus le nombre d'interruptions d'un $N$-gram positionnel est grand, plus l'espace couvert par celui-ci est grand et plus il est possible qu'une particule fréquente se trouve plus d'une fois à la même position par rapport aux autres éléments du $N$-gram par le simple fruit du hasard. Par exemple, le $N$-gram positionnel [*0 période 1 transitoire 4 à*] a été extrait par le GenLocalMaxs alors qu'il ne détermine pas une association lexicale correcte du fait de l'occurrence imprévue de la préposition *à*. Cette situation est illustrée dans le tableau

| Interruptions | échantillon 1 | échantillon 2 | échantillon 3 | Moyenne |
|---|---|---|---|---|
| *=1* | | | | |
| Corrects | 18 | 25 | 21 | 21 |
| Total | 49 | 49 | 49 | 49 |
| Précision | 36.73% | 51.02% | 42.85% | 43.53% |
| Couverture | 0.45% | 0.62% | 0.52% | 0.53% |
| *=2* | | | | |
| Corrects | 7 | 1 | 5 | 4 |
| Total | 42 | 42 | 42 | 42 |
| Précision | 16.66% | 2.38% | 11.90% | 10.31% |
| Couverture | 0.35% | 0.05% | 0.25% | 0.21% |
| *=3* | | | | |
| Corrects | 1 | 1 | 4 | 2 |
| Total | 42 | 42 | 42 | 42 |
| Précision | 2.38% | 2.38% | 9.52% | 4.76% |
| Couverture | 0.01% | 0.01% | 0.06% | 0.02% |
| *=4* | | | | |
| Corrects | 1 | 2 | 0 | 1 |
| Total | 42 | 42 | 42 | 42 |
| Précision | 2.38% | 4.76% | 0.00% | 2.38% |
| Couverture | 0.005% | 0.01% | 0.00% | 0.005% |

TAB. 9.23 – Précision/Couverture du Portugais selon les interruptions des $N$-grams

9.24 grâce à l'utilisation du concordanceur.

Cette caractéristique est malheureusement fréquente et est à l'origine d'un nombre important d'associations incorrectes. Ces résultats sont particulièrement importants en ce qui concerne la taille de l'environnement immédiat considéré. En effet, ils démontrent que plus l'environnement immédiat est grand, plus il est probable que les résultats d'extraction empirent. De fait, plus l'espace couvert est grand, plus les associations lexicales peuvent être aléatoires. Nous vérifierons cette hypothèse dans la suite de notre évaluation à partir de l'expérience sur les corpora d'environ 30000 formes graphiques.

| | | | | | |
|---|---|---|---|---|---|
| où la | *période transitoire* | est | venue | *à* | échéance |
| plus longue | *période transitoire* | pour | permettre | *à* | l'Agence |

TAB. 9.24 – Concordanceur pour [ *0 période 1 transitoire 4 à*]

Avant de passer à la comparaison entre les différentes mesures d'association normalisées, nous devons encore analyser deux scenari d'extraction. En effet, nous devons prendre en compte l'analyse des taux de précision et de couverture selon la taille des $N$-grams positionnels extraits et selon leur fréquence. Nous commencerons donc par l'analyse des résultats d'extraction par taille. Les résultats sont illustrés dans les tableaux 9.25 et 9.26.

Les résultats obtenus sont particulièrement intéressants. En effet, les valeurs de précision et de couverture diffèrent fondamentalement selon le nombre de constituants des $N$-grams positionnels. Une explication est cependant possible pour chacun de ces cas de figure. Premièrement, le fait que les unités contenant exactement deux formes graphiques soient plus précises que les autres se doit à deux caractéristiques principales. D'une part, du fait de la structure des digrams positionnels, le nombre d'interruptions maximum qu'ils contiennent en leur sein est de deux. Ainsi, en accord avec ce que nous avons dit préalablement, il est fort probable que ces digrams soient des associations lexicales correctes car contenant peu d'intérruptions. D'autre part, la pression du GenLocalMaxs sur les digrams positionnels est particulièrement forte du fait de la comparaison de mesures d'association normalisées sur la base de fréquences marginales pour les digrams et de valeurs de cohésion basées sur une moyenne de fréquences jointes pour les trigrams positionnels. Ainsi, pour être élu, un digram positionnel doit manifester une valeur de cohésion particulièrement forte ce qui a pour conséquence d'assurer sa correction. Cette même pression numérique est également à l'origine du faible poids des digrams positionnels par rapport aux autres tailles de $N$-grams. Ainsi, le taux de couverture des digrams est l'un des plus faibles pour $N = 2$. En effet, peu d'unités polylexicales démontrent de très hautes valeurs de cohésion.

En ce qui concerne les trigrams positionnels, les résultats sont aussi clairs que pour ceux des digrams. En effet, le taux de précision faible s'explique de la même façon que précédemment. D'une part, le nombre d'interruptions d'un $N$-gram positionnel est maximum pour $N = 3$. En effet, à l'extrême, un trigram positionnel peut contenir quatre

| Taille | échantillon 1 | échantillon 2 | échantillon 3 | Moyenne |
|---|---|---|---|---|
| ***N=2*** | | | | |
| Corrects | 33 | 41 | 39 | 37 |
| Total | 49 | 49 | 49 | 49 |
| Précision | 67.34% | 83.67% | 79.59% | 76.86% |
| Couverture | 0.165% | 0.205% | 0.195% | 0.18% |
| ***N=3*** | | | | |
| Corrects | 25 | 23 | 31 | 26 |
| Total | 56 | 56 | 56 | 56 |
| Précision | 44.64% | 41.07% | 55.35% | 47.02% |
| Couverture | 1.25% | 1.15% | 1.55% | 1.31% |
| ***N=4*** | | | | |
| Corrects | 27 | 19 | 22 | 22 |
| Total | 45 | 45 | 45 | 45 |
| Précision | 60.00% | 42.22% | 48.88% | 50.36% |
| Couverture | 0.9% | 0.63% | 0.73% | 0.75% |
| ***N=5*** | | | | |
| Corrects | 33 | 27 | 32 | 30 |
| Total | 54 | 54 | 54 | 54 |
| Précision | 61.11% | 50.00% | 59.25% | 56.78% |
| Couverture | 0.33% | 0.27% | 0.32% | 0.30% |
| ***N=6*** | | | | |
| Corrects | 23 | 23 | 32 | 26 |
| Total | 46 | 46 | 46 | 46 |
| Précision | 50.00% | 50.00% | 69.56% | 56.52% |
| Couverture | 0.11% | 0.11% | 0.16% | 0.13% |

TAB. 9.25 – Précision/Couverture du Français selon la taille de $N$-gram

interruptions ce qui implique nécessairement une grande couverture de l'espace du texte qui conduit à de faibles résultats de précision. D'autre part, inversement au cas des digrams positionnels, la pression du GenLocalMaxs est beaucoup plus faible pour les trigrams. Par conséquent, cette caractéristique implique des valeurs de cohésion relativement faibles pour l'extraction de trigrams. En ce qui concerne le taux de couverture,

| Taille | échantillon 1 | échantillon 2 | échantillon 3 | Moyenne |
|---|---|---|---|---|
| ***N=2*** | | | | |
| Corrects | 32 | 33 | 37 | 34 |
| Total | 50 | 50 | 50 | 50 |
| Précision | 64.00% | 66.00% | 74.00% | 68.00% |
| Couverture | 0.16% | 0.16% | 0.18% | 0.17% |
| ***N=3*** | | | | |
| Corrects | 15 | 26 | 21 | 20 |
| Total | 65 | 65 | 65 | 65 |
| Précision | 23.07% | 40.00% | 32.30% | 31.79% |
| Couverture | 0.75% | 1.30% | 1.05% | 1.03% |
| ***N=4*** | | | | |
| Corrects | 27 | 18 | 17 | 20 |
| Total | 56 | 56 | 56 | 56 |
| Précision | 48.21% | 31.14% | 30.35% | 52.63% |
| Couverture | 0.9% | 0.60% | 0.56% | 0.68% |
| ***N=5*** | | | | |
| Corrects | 20 | 25 | 29 | 24 |
| Total | 51 | 51 | 51 | 51 |
| Précision | 39.21% | 49.01% | 56.86% | 48.36% |
| Couverture | 0.25% | 0.31% | 0.36% | 0.30% |
| ***N=6*** | | | | |
| Corrects | 31 | 23 | 24 | 26 |
| Total | 56 | 56 | 56 | 56 |
| Précision | 55.35% | 41.07% | 42.85% | 46.42% |
| Couverture | 0.15% | 0.11% | 0.12% | 0.13% |

TAB. 9.26 – Précision/Couverture du Portugais de la taille de $N$-gram

celui-ci se met en évidence par son importance. En effet, il est de loin le plus fort pour les trigrams positionnels. Ce résultat est la conséquence logique des résultats d'extraction de notre système. En effet, une grande partie des associations lexicales candidates proposées par notre architecture sont des trigrams positionnels. Il n'est donc pas surprenant que le taux de couverture soit particulièrement élevé pour le cas des trigrams.

Finalement, les résultats ont tendance à se stabiliser autour de la moyenne i.e. 50% en ce qui concerne le taux de précision pour les $N$-grams restants. En effet, le nombre d'interruptions diminue au fur et à mesure que la taille du $N$-gram augmente, mais en contrepartie les suites repérées perdent en précision incarnant des suites toutes faites sans sens réel du point de vue lexical.

Afin de poursuivre notre étude des résultats d'extraction, il nous reste à examiner en détail la distribution des résultats suivant les fréquences manifestées par les suites de formes graphiques extraites. Nous illustrons ces résultats dans les tableaux 9.27 et 9.28.

Les résultats obtenus sont conformes à nos expectatives. En effet, plus la fréquence d'un $N$-gram positionnel candidat est élevée, plus il est probable que celui-ci soit une association lexicale correcte. En particulier, ceci est dû à nos hypothèses de départ qui suggèrent que la fréquence joue un rôle primordial pour le repérage d'unités lexicales complexes. Cependant, bien que valide, cette caractéristique montre clairement son insuffisance. En effet, si le seul critère de sélection se limitait à la fréquence, la plupart des associations lexicales ne seraient pas identifiées. En effet, le taux de couverture montre qu'un nombre significatif de suites lexicales n'apparaît que deux fois dans le corpus. Ce résultat est tout à fait encourageant. En effet, jusqu'à présent, tous les extracteurs proposés dans la littérature se basaient sur l'étude de textes de grandes tailles afin de leur permettre des résultats d'extraction satisfaisants. Par exemple, F. Smadja [32] applique une valeur de fréquence limite supérieure à 50 pour extraire ses associations lexicales. Cette démarche n'est en aucun cas satisfaisante comme le démontrent nos résultats. En effet, dans ces conditions, nous n'osons imaginer quel serait le taux de couverture obtenu. Notre architecture propose donc une solution originale qui permet effectivement l'extraction d'associations lexicales peu fréquentes. Cette caractéristique est des plus importantes car elle nous permet de tester notre système sur des énoncés de petite taille ce qui s'est avéré impossible pour la plupart des architectures recensées. En particulier, nous verrons que les résultats obtenus à partir du corpus d'environ 30000 formes graphiques sont des plus satisfaisants et que les différences tant au niveau du taux de précision que du taux de couverture ne dévient pas drastiquement d'une expérience à l'autre.

| Fréquence | échantillon 1 | échantillon 2 | échantillon 3 | Moyenne |
|---|---|---|---|---|
| *=2* | | | | |
| Corrects | 88 | 102 | 86 | 92 |
| Total | 206 | 206 | 206 | 206 |
| Précision | 42.71% | 49.51% | 41.74% | 44.65% |
| Couverture | 1.76% | 2.04% | 1.72% | 1.84% |
| *=3* | | | | |
| Corrects | 27 | 32 | 34 | 31 |
| Total | 55 | 55 | 55 | 55 |
| Précision | 49.09% | 58.18% | 61.81% | 53.36% |
| Couverture | 0.27% | 0.32% | 0.34% | 0.31% |
| *=4* | | | | |
| Corrects | 34 | 34 | 39 | 35 |
| Total | 54 | 54 | 54 | 54 |
| Précision | 62.96% | 62.96% | 72.22% | 66.04% |
| Couverture | 0.11% | 0.11% | 0.13% | 0.11% |
| $\geq$ *5* | | | | |
| Corrects | 37 | 36 | 44 | 39 |
| Total | 53 | 53 | 53 | 53 |
| Précision | 69.81% | 67.92% | 83.01% | 73.58% |
| Couverture | 0.37% | 0.36% | 0.44% | 0.39% |

TAB. 9.27 – Précision/Couverture du Français selon la fréquence des $N$-grams

Nous abordons finalement notre évaluation comparative entre les différentes mesures normalisées que sont le coefficient d'association [27], le coefficient Dice [45], la Probabilité Conditionnelle Symétrique [30], le test $\Phi^2$ [29], le coefficient de vraisemblance LogLike [28] et bien sûr l'Expectative Mutuelle. Contrairement à ce que nous avons fait dans les parties précédentes où une analyse exhaustive des résultats a été menée, nous ne présenterons ici que les résultats globaux de précision et de couverture. En effet, comme nous l'avons mentionné dans la partie précédente, une forte proportion des $N$-grams positionnels non contigus extraits à partir des heuristiques autres que l'Expectative Mutuelle ne sont pas des associations lexicales correctes. Ainsi, il est clair que les résultats mis en évidence selon le type des $N$-grams positionnels i.e. contigus ou non

| Fréquence | échantillon 1 | échantillon 2 | échantillon 3 | Moyenne |
|---|---|---|---|---|
| *=2* | | | | |
| Corrects | 62 | 65 | 61 | 62 |
| Total | 188 | 188 | 188 | 188 |
| Précision | 32.97% | 34.57% | 32.44% | 33.32% |
| Couverture | 1.55% | 1.62% | 1.52% | 1.56% |
| *=3* | | | | |
| Corrects | 26 | 32 | 24 | 27 |
| Total | 54 | 54 | 54 | 54 |
| Précision | 48.14% | 59.25% | 44.44% | 50.61% |
| Couverture | 0.32% | 0.4% | 0.3% | 0.34% |
| *=4* | | | | |
| Corrects | 26 | 32 | 35 | 31 |
| Total | 54 | 54 | 54 | 54 |
| Précision | 48.14% | 59.25% | 64.81% | 57.04% |
| Couverture | 0.10% | 0.13% | 0.14% | 0.12% |
| $\geq$ *5* | | | | |
| Corrects | 31 | 29 | 36 | 32 |
| Total | 52 | 52 | 52 | 52 |
| Précision | 59.61% | 55.76% | 69.23% | 61.53% |
| Couverture | 0.31% | 0.21% | 0.36% | 0.32% |

TAB. 9.28 – Précision/Couverture du Portugais selon la fréquence des $N$-grams

contigus, seraient forcément moins intéressants que ceux de l'Expectative Mutuelle. Parallèlement, d'après notre première étude, les résultats de précision et de couverture suivent de très près les caractéristiques d'extraction mises en évidence par chacune des mesures d'association. Nous nous limiterons donc à proposer les résultats globaux d'extraction en alertant le lecteur que ceux-ci sont suffisants pour comprendre les caractéristiques de chacune des heuristiques. Nous illustrons donc les taux de précision et de couverture pour l'ensemble des heuristiques proposées à partir des tableaux 9.29 et 9.30.

Nous allons d'abord analyser les résultats obtenus par l'Expectative Mutuelle. Nous rappelons que pour cette expérience la taille de l'environnement immédiat a été fixée à

5. Par conséquent, l'espace couvert par chaque $N$-gram positionnel est potentiellement supérieur à celui défini pour un environnement immédiat de taille 3. Ainsi, il est fort probable que les résultats de précision soient plus faibles dans cette expérience. Or, c'est exactement ce que nous vérifions. En effet, pour les deux langues, la précision d'extraction a diminué. Néanmoins, les valeurs mises en évidence par l'Expectative Mutuelle ne dévient que d'environ 7% relativement à la première expérience. Ceci est particulièrement encourageant du fait de la taille considérée de l'énoncé et de l'environnement immédiat élargi. Dans ces conditions, notre extracteur suggère des valeurs de précision tout à fait acceptables. Comme nous l'avions déjà remarqué, notre architecture est parfaitement capable de proposer des associations lexicales correctes même dans des conditions d'expérience adverses i.e. pour des énoncés de petite taille et pour des environnements immédiats des plus larges. En ce qui concerne le taux de couverture, celui-ci démontre une baisse significative par rapport à l'expérience précédente. Cette baisse est d'environ 1%. Ceci est dû en grande partie au type du texte considéré i.e. à la forme dont il a été écrit. En effet, ce corpus est un agglomérat de questions-réponses de la Commission Européenne et par conséquent se distingue par un certain hâchage du texte qui ne permet pas de rehausser les régularités du langage. Finalement, les différences illustrées dans l'expérience précédente entre le Français et le Portugais semblent se répéter. En effet, dans les deux cas, le taux de précision est supérieur dans le cas du Français alors que le taux de couverture est supérieur pour le Portugais. Les faibles résultats de précision pour le Portugais sont dus en particulier à la forte proportion d'associations lexicales non contiguës incorrectes. En effet, l'utilisation très fréquente de la préposition *de* s'est révélée néfaste au point d'impliquer l'extraction d'un nombre important de suites fortement liées par le simple fait du hasard. Inversement, cette caractéristique a pour conséquence l'extraction d'un nombre plus important d'associations lexicales correctes. En effet, parmi toutes les suites qui contiennent la préposition *de*, toutes ne sont pas incorrectes !

Après avoir analysé les résultats de l'Expectative Mutuelle, nous nous attachons maintenant à l'étude des résultats mis en évidence par les autres mesures d'association. Le coefficicient Dice s'est révélé être la meilleure seconde heuristique après l'Expectative Mutuelle. Ceci n'est pas une surprise. En effet, nous avons remarqué dans la partie précédente que le coefficient Dice tend à élire des $N$-grams positionnels dont la fréquence est égale à 3. Or, nous savons que plus la fréquence d'une suite de formes graphiques

est élevée, plus il est probable que la séquence soit une association lexicale correcte. C'est exactement ce qui se passe ici. Ainsi, le coefficient Dice démontre une certaine aptitude à repérer des associations lexicales fréquentes. Néanmoins, ceci n'explique pas complètement ces chiffres. En effet, le fait que le coefficient Dice élise préférablement des digrams positionnels implique également une hausse relative du taux de précision. En effet, comme nous l'avons dit précédemment, les digrams positionnels contiennent un nombre réduit d'interruptions ce qui favorise leur pertinence. Ainsi, les bons résultats de précision sont également dus à l'extraction d'un nombre important de digrams positionnels.

Après le coefficient Dice, le coefficient de vraisemblance LogLike démontre en moyenne le meilleur comportement en terme de précision. Cette mesure a mis en évidence certaines caractéristiques que l'on peut comparer à celles du coefficient Dice notamment en ce qui concerne l'extraction préférentielle de digrams positionnels. Parallèlement au coefficient Dice, le coefficient de vraisemblance LogLike démontre donc l'extraction d'une forte proportion de digrams corrects ayant pour conséquence la définition d'un taux de précision global comparativement intéressant. En ce qui concerne le taux de couverture, le coefficient LogLike est de loin le plus prolifique. En effet, il démontre un taux au moins 4 fois supérieur au deuxième meilleur résultat qui appartient à l'Expectative Mutuelle. Cette caractéristique est due principalement à l'extraction d'un nombre particulièrement important de digrams positionnels par rapport aux autres mesures d'associations.

La Probabilité Conditionnelle Symétrique et le test $\Phi^2$ sont respectivement les deux mesures d'association les plus satisfaisantes après les trois heuristiques que nous venons d'analyser. Leur comportement similaire au niveau de l'extraction des associations lexicales candidates leur procure des caractéristiques semblables au niveau de la précision et de la couverture bien que celles-ci soient meilleures pour la Probabilité Conditionnelle Symétrique. Leurs faibles résultats de précision s'expliquent par leur aptitude à extraire préférablement des trigrams positionnels qui contiennent un nombre d'interruptions potentielles maximum. Or, comme nous le savons, cette caractéristique est fortement préjudiciable à la précision des résultats. Parallèlement, la différence qui existe entre les deux mesures tient au fait que la Probabilité Conditionnelle Symétrique tend à élire un nombre plus important de tetragrams positionnels qui augurent une plus grande précision par rapport aux trigrams. Ainsi, les différences de précision et de couverture

sont fondamentalement imputables à cette caractéristique.

Finalement, le coefficient d'association met en évidence les résultats les moins satisfaisants de toutes les mesures d'association testées. En effet, son taux de précision moyen est le plus faible et son taux de couverture moyen l'un des moins élevés. L'une des particularités évidente du coefficient d'association est sa forte tendance à repérer des unités polylexicales candidates de faible fréquence. Cette caractéristique est particulièrement préjudiciable à la précision des unités extraites. En effet, les fréquences faibles sont le plus souvent le témoin de l'extraction de suites non contiguës particulièrement rares et par conséquent dues au hasard. C'est ce qui se passe ici. En effet, la plupart des associations retenues par le coefficient d'association sont des $N$-grams non contigus contenant de nombreuses interruptions. Il est évident que dans ces conditions, il est très difficile de mettre en évidence des résultats d'extraction acceptables. En ce qui concerne le taux de couverture moyen, celui-ci s'approche de ceux du test $\Phi^2$ et de la Probabilité Conditionnelle Symétrique du fait du nombre important d'associations candidates extraites et en aucun cas grâce au taux de précision.

## 9.5 Conclusion

Nous concluons ainsi notre analyse quantitative en démontrant que l'Expectative Mutuelle se distingue positivement de l'ensemble des mesures d'associations normalisées. En effet, tant pour un texte de "grande" taille que pour un texte de petite taille associé à un environnement immédiat de taille élevée, le GenLocalMaxs et l'Expectative Mutuelle ont démontré des résultats de précision et de couverture satisfaisants. Nous avons également montré que les résultats les moins encourageants sont certainement dus aux suites non contiguës dont le taux de précision est particulièrement faible. Néanmoins, notre architecture propose une approche différente qui est capable d'élire avec un degré de confiance important des suites contiguës de formes graphiques et de proposer un ensemble d'associations lexicales non contiguës qui seront utiles pour un bon nombre d'applications du traitement du langage naturel ■

| Mesure | échantillon 1 | échantillon 2 | échantillon 3 | Moyenne |
|---|---|---|---|---|
| ***PHI*** | | | | |
| Corrects | 14 | 14 | 10 | 12 |
| Total | 105 | 105 | 105 | 105 |
| Précision | 13.33% | 13.33% | 9.52% | 12.06% |
| Couverture | 1.03% | 1.03% | 0.73% | 0.93% |
| ***DICE*** | | | | |
| Corrects | 32 | 27 | 31 | 30 |
| Total | 111 | 111 | 111 | 111 |
| Précision | 28.82% | 24.32% | 27.92% | 27.02% |
| Couverture | 1.88% | 1.59% | 1.82% | 1.76% |
| ***LOGLIKE*** | | | | |
| Corrects | 17 | 25 | 17 | 19 |
| Total | 104 | 104 | 104 | 104 |
| Précision | 16.34% | 21.92% | 16.34% | 18.02% |
| Couverture | 7.15% | 10.51% | 7.15% | 8.27% |
| ***ME*** | | | | |
| Corrects | 46 | 45 | 45 | 45 |
| Total | 99 | 99 | 99 | 99 |
| Précision | 46.46% | 45.45% | 45.45% | 45.78% |
| Couverture | 1.93% | 1.89% | 1.89% | 1.90% |
| ***I*** | | | | |
| Corrects | 11 | 10 | 14 | 11 |
| Total | 112 | 112 | 112 | 112 |
| Précision | 9.82% | 8.92% | 12.50% | 12.50% |
| Couverture | 1.07% | 0.98% | 1.37% | 1.14% |
| ***SCP*** | | | | |
| Corrects | 21 | 17 | 21 | 19 |
| Total | 102 | 102 | 102 | 102 |
| Précision | 20.58% | 16.66% | 20.58% | 19.27% |
| Couverture | 1.23% | 1.00% | 1.23% | 1.15% |

TAB. 9.29 – Précision/Couverture du Français selon les Mesures d'Association

| Mesure | échantillon 1 | échantillon 2 | échantillon 3 | Moyenne |
|---|---|---|---|---|
| ***PHI*** | | | | |
| Corrects | 16 | 9 | 18 | 14 |
| Total | 96 | 96 | 96 | 96 |
| Précision | 16.66% | 9.37% | 18.75% | 14.92% |
| Couverture | 1.03% | 0.58% | 1.16% | 0.92% |
| ***DICE*** | | | | |
| Corrects | 32 | 32 | 33 | 32 |
| Total | 107 | 107 | 107 | 107 |
| Précision | 29.90% | 29.90% | 30.84% | 30.21% |
| Couverture | 2.06% | 2.06% | 2.12% | 2.08% |
| ***LOGLIKE*** | | | | |
| Corrects | 24 | 17 | 22 | 21 |
| Total | 100 | 100 | 100 | 100 |
| Précision | 24.00% | 17.00% | 22.00% | 21.00% |
| Couverture | 11.05% | 7.82% | 10.13% | 9.66% |
| ***ME*** | | | | |
| Corrects | 42 | 40 | 35 | 39 |
| Total | 97 | 97 | 97 | 97 |
| Précision | 43.29% | 41.23% | 36.08% | 40.20% |
| Couverture | 1.93% | 1.84% | 1.61% | 1.79% |
| ***I*** | | | | |
| Corrects | 13 | 13 | 14 | 13 |
| Total | 107 | 107 | 107 | 107 |
| Précision | 12.14% | 12.14% | 13.08% | 12.45% |
| Couverture | 1.39% | 1.39% | 1.50% | 1.42% |
| ***SCP*** | | | | |
| Corrects | 15 | 22 | 20 | 19 |
| Total | 98 | 98 | 98 | 98 |
| Précision | 15.30% | 22.44% | 20.40% | 19.38% |
| Couverture | 0.96% | 1.41% | 1.28% | 1.21% |

TAB. 9.30 – Précision/Couverture du Portugais selon selon les Mesures d'Association

# Quatrième partie

# Conclusions

**"Les linguistiques de corpus se révéleront fructueuses comme domaine de recherche si l'on accepte l'imparfait, c'est-à-dire des ressources toujours impures et si s'affirment des collaborations soutenues entre linguistes et informaticiens"**

---

*Habert, Nazarenko et Salem [17]*

# Chapitre 10

# Conclusions et Perspectives

Après avoir présenté les différents concepts de notre architecture, nous avons réalisé une évaluation exhaustive de ses caractéristiques ainsi que de ses performances. Par conséquent, il nous reste maintenant à définir les différents axes de travail qui s'offrent à nous. Dans ce but, nous allons d'abord résumer l'ensemble des conclusions que nous avons dressées tout au long de notre exposé afin d'aider le lecteur dans son analyse. Nous introduirons ensuite la réalisation de trois nouvelles expériences prometteuses. Ainsi, nous mettrons en évidence la flexibilité de notre extracteur qui permet l'usage de textes composés de différents types d'unités textuelles. Parmi celles-ci, nous compterons avec les caractères [107] [112] et les étiquettes morpho-syntaxiques [113]. Finalement, face aux problèmes rencontrés par notre architecture initiale, nous conclurons notre rapport par la présentation d'une nouvelle mesure d'association basée sur l'usage de connaissances linguistiques spécifiquement introduites pour cet effet dans les énoncés. Nous verrons que l'enrichissement des modèles statistiques peut se faire tout naturellement sans la définition de patrons pré-déterminés.

## 10.1 Conclusions

Notre recherche s'est appuyée initialement sur le principe de base selon lequel l'énoncé ne doit en aucun cas être modifié et que toute l'information qui le compose doit être utilisée. Ainsi, le texte est simplement mis en forme pour des raisons évidentes de traitement informatique mais ne souffre d'aucun changement fondamental. Ainsi, il n'est ni lemmatisé, ni étiqueté morpho-syntaxiquement et encore moins épuré à partir de listes de mots "soi-disant" vides. Dans ce cadre, nous nous sommes souvent référés au principe d'intégrité du corpus. Cette spécificité est clairement orientée vers la création

d'un système totalement flexible et versatile. En effet, l'application de notre système à des corpora de différentes langues est directe comme nous l'avons montré avec le Français et le Portugais mais aussi avec d'autres langues comme l'Italien [103], le Slovène [100], l'Anglais [106] [104] [105] et l'Estonien [101]. Aucun traitement spécifique n'est donc nécessaire.

A partir de ce principe, il nous a fallu segmenter le texte en portions révélatrices du concept d'association lexicale. Or, dans ce cadre, la plupart des études se sont limitées à l'usage des modèles digrams contigus ou non, ou à la représentation en $N$-grams contigus — modèles $N$-gram classiques. Cependant, il est clair que ces modèles sont insuffisants pour représenter toutes les particularités du phénomène d'unité polylexicale. Ainsi, nous avons introduit les modèles $N$-gram positionnels qui sont capables de représenter les structures les plus complexes des phénomènes de figement à partir de la définition d'un environnement immédiat. Le découpage du corpus en vecteurs ordonnés d'unités textuelles permet ainsi de représenter les associations lexicales contiguës mais également distantes. Ces modèles souffrent cependant d'une calculabilité d'ordre exponentiel qui limite fortement leur utilisation. Ainsi, la taille du corpus à utiliser dépend nécessairement de la taille de l'environnement considéré. Par exemple, il n'est pas pensable d'imaginer de segmenter un corpus de 200000 formes graphiques pour un environnement immédiat de taille 5. Cependant, cet inconvénient n'est pas des plus génants. En effet, les résultats de performance de notre architecture montrent qu'il est préférable d'utiliser un environnement immédiat qui réduise au maximum le nombre d'interruptions possibles pour une suite de formes graphiques. Ainsi, notre architecture montre un comportement idéal — précision et couverture — pour un environnement de taille 3. Dans ces conditions, nous pouvons envisager le traitement d'un texte d'un million de formes graphiques grâce notamment à une codification optimisée pour notre application. Ainsi, l'ensemble de notre système peut traiter un texte de cette taille en trois jours à partir d'un ordinateur personnel de type Pentium à 166 Mhz avec 128 Mo de mémoire vive et 20 Go de disque rigide.

Parallèlement, nous nous sommes vu confrontés au problème de la définition d'une mesure d'association normalisée. En effet, d'une part, la plupart des mesures d'association ne sont définies que pour deux unités textuelles. D'autre part, les mesures d'association $N$-aires existantes mettent en évidence des insuffisances représentatives évidentes. Dans un premier temps, nous avons donc proposé de normaliser un ensemble de mesures d'as-

sociation binaires couramment utilisées dans le domaine du traitement automatique des langues. Seulement, les résultats obtenus ont montré leur insuffisance pour traiter toute l'information présente dans les textes. En effet, ces heuristiques tendent à sous-évaluer les associations contenant des unités textuelles fréquentes impliquant irrémédiablement l'extraction de suites non contiguës incorrectes. Ainsi, nous avons défini une nouvelle mesure d'association, l'Expectative Mutuelle, qui s'appuie sur les concepts de fréquence relative et d'Expectative Normalisée. En particulier, cette nouvelle heuristique démontre un comportement réellement supérieur comparativement à toutes les mesures normalisées testées. Ainsi, toutes les suites non contiguës extraites par l'Expectative Mutuelle mettent en évidence l'occurrence de plusieurs unités textuelles pour chacune de leurs interruptions. Dans ces conditions, l'Expectative Mutuelle s'est avérée être un atout important de la réussite de notre architecture globale.

Dans le cadre proprement dit de l'extraction des associations lexicales, nous présentons également une nouvelle approche sur laquelle nous manifestons une grande confiance. La tâche spécifique à l'extraction des unités de sens a souvent été laissée de côté, les méthodes proposées s'attachant même quasiment exclusivement à définir de nouvelles mesures d'association toujours plus performantes. Ainsi, la phase d'extraction revenait à définir une valeur seuil de fréquence et/ou de mesure d'association marquant une frontière entre les unités polylexicales correctes et incorrectes. Cependant, cette technique est particulièrement peu fiable du fait de son analyse trop grossière. En effet, comme le soulignent K. Frantzi et S. Ananiadou [33], cette méthodologie impose un post-traitement des résultats. Par exemple, si la suite *Conseil des Nations Unies* dépasse la valeur seuil stipulée, il est très probable que la suite *Conseil des Nations* dépasse également cette limite. Il est alors nécessaire d'éliminer la dernière suite de formes graphiques qui ne forme pas une association lexicale. Dans ce domaine, l'algorithme de maxima locaux, le Gen-LocalMaxs, propose une solution originale basée sur l'analyse locale de chaque $N$-gram positionnel disponible. Ainsi, nous proposons une analyse spécifique de l'environnement immédiat de chaque $N$-gram. Cette approche particulièrement fine nous permet en autre d'extraire des associations lexicales dont la fréquence est relativement faible et qui ne seraient certainement pas extraites à partir de valeurs seuil globales nécessairement élevées.

Finalement, les résultats de notre architecture se sont révélés particulièrement intéressants.

En ce qui concerne son aptitude à extraire des phénomènes de figement, notre extracteur a démontré un comportement excellent. En effet, une vaste diversité de phénomènes ont pu être mis en évidence : noms composés, locutions prépositives et conjonctives en passant par des structures obtenues par composition ou coordination. Ainsi, nous pouvons dire que nos objectifs initiaux ont été atteints avec succès. En ce qui concerne la performance de notre architecture, les résultats absolus calculés sont comme prévu moins satisfaisants. Cependant, tant au niveau de la précision qu'au niveau de la couverture, l'Expectative Mutuelle a démontré des résultats comparativement bons : première heuristique au niveau de la précision et seconde au niveau de la couverture. En moyenne, la combinaison Expectative Mutuelle et GenLocalMaxs est donc de loin la meilleure. Les résultats moins satisfaisants sont certainement à mettre au crédit des suites non contiguës. En effet, un nombre important des suites non contiguës retenues ne forment pas des associations lexicales correctes du fait de l'occurrence fortuite répétée de fragments fonctionnels. Afin de réduire cet effet de bord, nous présenterons dans la dernière partie de cette conclusion une proposition de solution.

Il convient enfin de faire référence à l'implémentation du système. Celui-ci a été réalisé dans sa globalité — i.e. segmentation, mesures d'association et GenLocalMaxs — à partir de scripts GAWK et de programmes en C sur une plateforme LINUX pour ordinateur personnel.

Finalement, comme nous l'avons déjà répété à plusieurs reprises, notre architecture se distingue par sa forte flexibilité et versatilité. Dans ce contexte, nous proposons dans la prochaine section les résultats de trois expériences qui ont été réalisées à partir d'unités textuelles de base différentes i.e. les caractères et les étiquettes morpho-syntaxiques.

## 10.2 Perspectives

A partir d'un texte transformé en une série de caractères, nous nous intéressons à l'extraction d'associations morphologiques telles que les morphèmes affixaux, les thèmes et les associations entre préfixes et suffixes. En effet, ces unités représentent des suites de caractères qui se trouvent plus fréquemment associés qu'ils ne le seraient vraiment par le simple fruit du hasard. La flexibilité de notre extracteur nous a donc permis de réaliser un certain nombre d'expériences pour le Français et le Portugais à partir

d'une simple transformation du corpus initial. Ainsi, plutôt que de travailler à partir de $N$-grams positionnels de formes graphiques, notre extracteur utilise des $N$-grams positionnels de caractères. Cette expérience a fait l'objet d'une publication aux 5$^{\text{èmes}}$ Journées Internationales d'Analyse de Données Textuelles [107]. Nous reprenons dans les deux tableaux 10.1 et 10.2 certains des résultats décrits dans cet article afin d'illustrer le genre d'unités morphologiques extraites. Pour des raisons évidentes de clarté, le caractère espace a été remplacé par le caractère "#".

| ME | Fréquence | Association Morphologique | Occurrences |
|---|---|---|---|
| 0.00015 | 472 | égal | #égale<br>illégales |
| 0.00011 | 376 | tif# | préventif#<br>consultatif# |
| 2.8e-05 | 73 | huma | inhumaines<br>humanitaire |
| 0.00068 | 1453 | eurs# | demandeurs#<br>donateurs# |
| 4.4e-05 | 100 | #rédu | #réduction<br>#réduire |

TAB. 10.1 – Associations Morphologiques du Français

Cette étude est encore au niveau embryonnaire. En effet, une analyse exhaustive serait nécessaire pour déterminer l'efficacité réelle de notre architecture. Cependant, les résultats préliminaires nous paraissent particulièrement intéressants. Ainsi, un nombre important d'associations morphologiques qui peuvent être très utiles dans le cadre des analyseurs lexicaux ont été extraites. Entre autres, ces résultats peuvent permettre de confronter les formalismes des règles de formation des mots avec la réalité des corpora ou même des dictionnaires électroniques[1]. Parallèlement, certains des résultats mettent en évidence des caractéristiques intéressantes dans le cadre de la synthèse de la voix. Par exemple, l'association *s#anos* définit le phonème de liaison entre le pluriel du déterminant et son nom associé.

[1]En effet, un dictionnaire peut être considéré comme un énoncé formé par l'ensemble des entrées qu'il contient.

| ME | Fréquence | Association Morphologique | Occurrences |
|---|---|---|---|
| 0.00012 | 4663 | #in | #independente<br>#indesejados |
| 1.6e-05 | 64 | goz | #gozam<br>regoziga-se |
| 0.00026 | 732 | vel# | aplicável#<br>favorável# |
| 0.00022 | 465 | posiç | #posição<br>disposição |
| 5.6e-05 | 199 | s#anos | três#anos<br>muitos#anos |

TAB. 10.2 – Associations Morphologiques du Portugais

Dans le même ordre d'idée, nous avons testé notre architecture à partir d'un énoncé d'étiquettes morpho-syntaxiques. Ainsi, nous proposons d'extraire des séquences d'étiquettes pertinentes à partir des hypothèses de cohésion énoncées initialement pour les formes graphiques. Le but principal de cette étude est d'extraire automatiquement des patrons syntaxiques pertinents pour la construction de grammaires locales ou non. Dans ce contexte, nous avons testé notre architecture à partir du Corpus Brown pour lequel seules les étiquettes morpho-syntaxiques ont été retenues. Les résultats sont présentés dans le tableau 10.3 auquel nous avons associé le tableau de correspondance 10.4.

| ME | Fréquence | Associations Morpho-syntaxiques |
|---|---|---|
| 0.00629 | 12786 | AT JJ NN |
| 3.5e-06 | 7 | JJ NP CC JJ NP |
| 0.00011 | 228 | NP $ JJ NN |
| 1.0e-05 | 26 | AT JJ CC AT JJ |
| 1.0e-05 | 147 | HV RB BEN |

TAB. 10.3 – Associations Morpho-syntaxiques du Corpus Brown

Les résultats ont permis de mettre en évidence un nombre important de patrons

| Etiquette Morpho-syntaxique | Correspondance |
|---|---|
| $ | Possessif |
| AT | Article |
| BEN | Participe passé de l'auxiliaire *be* |
| CC | Conjonction de Coordination |
| HV | Auxiliaire *have* |
| JJ | Adjectif |
| NN | Nom commum |
| NP | Nom propre |
| RB | Adverbe |

TAB. 10.4 – Tableau de Correspondance pour le Corpus Brown

syntaxiques tels que les syntagmes nominaux, les syntagmes verbaux et les associations Sujet-Verbe et Verbe-Objet. Parallèlement, notre extracteur a démontré des capacités notables d'extraction de syntagmes contenant des conjonctions de coordination qui posent souvent des problèmes d'analyse. Là encore, notre étude est restée très superficielle. Nous souhaitons en effet élargir ces expériences à des corpora du Français et du Portugais et vérifier si des résultats similaires sont possibles.

Finalement, nous avons réalisé une dernière expérience dans le domaine de l'alignement de textes avec A. Ribeiro [112]. Dans ce contexte, de nombreuses études ont démontré l'intérêt de l'utilisation de cognats i.e. de mots dont la forme est similaire dans les deux langues à analyser. Par exemple, *Parlement* et *Parlamento* peuvent être considérés comme étant deux cognats. Cependant, les heuristiques proposées pour leur identification se sont souvent limitées à la comparaison de suites de caractères à partir de mesures de similitude. Or, ces méthodes sont peu fiables du fait de leur faible base de travail au niveau du matériel textuel. En effet, la similitude entre les suites de caractères est souvent basée sur une faible portion du texte et ne prend donc pas en compte l'ensemble de l'information disponible dans les corpora. Nous avons donc proposé une nouvelle approche à ce problème. Ainsi, nous avons défini un cognat potentiel comme étant une suite, contiguë ou non de caractères, pertinente commune aux deux[2] langues considérées pour l'alignement. Un cognat potentiel sera donc un $N$-gram positionnel commun aux

[2]Nous pouvons également travailler avec un nombre théoriquement infini de langues.

deux langues et formant un maximum local[3].

Le principe de reconnaissance de cognats se divise en quatre étapes. Dans un premier temps, les deux textes à aligner sont transformés en suites de caractères, l'unité textuelle de base étant évidemment le caractère. Ensuite, les deux textes transformés sont concaténés pour former un seul corpus bilingue. Dans une troisième étape, notre architecture est appliquée sur ce nouveau corpus et produit un ensemble de $N$-grams positionnels pertinents. Finalement, les cognats potentiels sont les $N$-grams positionnels extraits qui sont communs aux deux langues. Le processus d'extraction est illustré à partir de la figure 10.1.

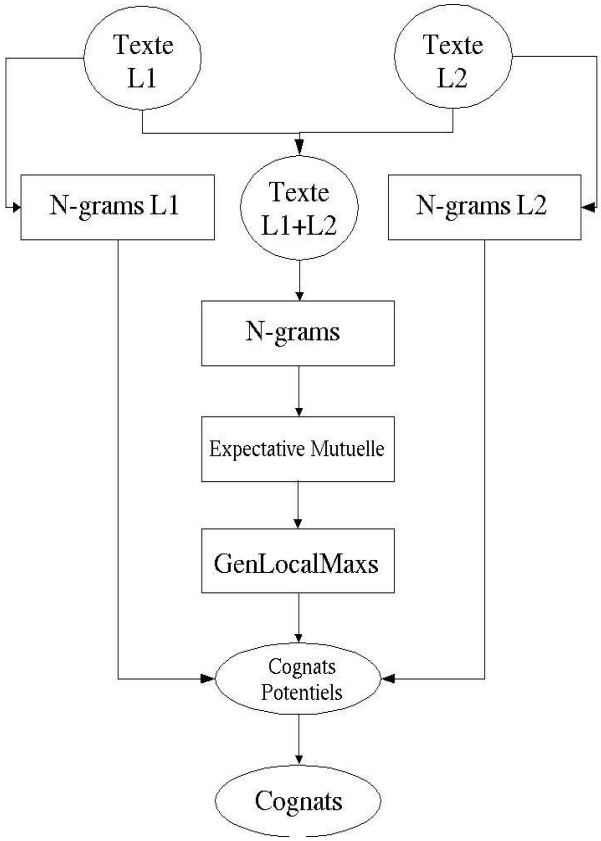


FIG. 10.1 – Extraction de Cognats

Nous avons réalisé plusieurs expériences entre différentes paires de langues : Anglais-Slovène, Anglais-Portugais, Portugais-Français, Espagnol-Portugais entre autres. Dans tous les cas de figures, les résultats se sont montrés à la hauteur de nos expectatives. Nous

[3]Dans l'expérience présentée dans [112], seuls les cognats potentiels dont la fréquence est égale dans les deux langues ont été utilisés.

illustrons, dans le tableau 10.5, quelques exemples extraits d'une expérience faite à partir de deux textes d'environ 50000 formes graphiques, l'un Portugais et l'autre Français [112].

| ME | Fréquence | Cognat | Occurrences |
|---|---|---|---|
| 0.00019 | 33 | migra | immigration<br>imigração |
| 0.00016 | 28 | audi ____ nc | audience<br>audiência |
| 0.00019 | 32 | #l ____ gisl | #législatif<br>#legislação |
| 4.18e-05 | 7 | #exclu | #exclure<br>#exclusão |
| 0.00019 | 36 | #Conse | #Conseil<br>#Conselho |

TAB. 10.5 – Cognats du Français et du Portugais

L'utilisation des modèles $N$-gram positionnels s'est avérée être un atout important pour la reconnaissance de vrais cognats. En effet, comme le montrent les résultats obtenus, de nombreuses formes graphiques diffèrent d'une langue à l'autre à cause de l'échange d'un caractère comme par exemple *Parlement* et *Parlamento* qui forment ainsi le cognat *Parl* ____ *ment*.

Cependant, l'utilisation de positions fixes peut s'avérer problématique. En effet, supposons les deux unités lexicales *graphiquement* et *graficamente*. Notre extracteur ne serait capable d'identifier que les trois premiers caractères i.e. *gra* bien que les deux formes graphiques partagent un plus grand nombre de lettres ceci à cause de l'utilisation des caractères "*ph*" en Français comparativement au seul caractère "*f*" pour le Portugais. Ainsi, tous les caractères suivants se trouvent décalés d'une langue à l'autre. Nous avons déjà réfléchi au problème et l'utilisation de positions moins rigides serait une solution à tester. Ainsi, nous garderions simplement l'ordre des caractères et non pas leur position. Nous pourrions également laisser à l'aligneur le soin de résoudre ce problème à partir de l'identification de séquences plus petites comme nous le proposons dans [112].

Nous avons donc vu dans cette section que la flexibilité de notre extracteur permet son utilisation dans un nombre important de tâches du traitement automatique du langage naturel. Ainsi, il suffit d'adapter le matériel textuel aux phénomènes recherchés. Là encore, notre démarche initiale prônant une totale versatilité s'est avérée être un succès permettant en autre la construction d'un système facilement adaptable. Il est clair cependant que de nombreuses études sont nécessaires pour déterminer la véritable utilité de notre système d'extraction.

Finalement, il convient de revenir sur les problèmes précédemment illustrés de notre architecture afin de proposer certaines améliorations dont en particulier l'introduction de connaissances linguistiques.

## 10.3 Améliorations

L'une des caractéristiques de notre extracteur est d'élire préférablement les suites récurrentes les plus longues présentes dans le corpus. Or, ceci n'est pas sans poser quelques problèmes. En effet, comme nous l'avons déjà remarqué, ceci implique souvent l'extraction de marqueurs au détriment de véritables associations lexicales. Pour reprendre un exemple déjà cité, notre architecture devrait préférer le nom composé *petites morues* au marqueur *de petites morues* réellement extrait. Ceci est dû au fait que chaque fois que la suite *petites morues* apparaît dans le corpus, celle-ci est précédée de la préposition *de*. Cette caractéristique est également responsable de l'extraction de suites non contiguës incorrectes comme c'est le cas pour la séquence *période transitoire* ____ ____ *à* où l'occurrence de la préposition *à* est fortuite et ne devrait pas être repérée.

Malheureusement, aucune autre mesure d'association ni aucun autre processus d'extraction ne propose une solution à ce problème. L'unique solution réside donc dans l'introduction de connaisssances linguistiques. Dans ce cadre, de nombreuses études [21] [49] [22] [98] ont proposé de réduire l'espace de recherche des unités polylexicales en définissant *a priori* des patrons syntaxiques pertinents pour leur extraction. Cependant, ces systèmes définissent deux étapes distinctes qui ne prennent pas en compte l'inter-dépendance entre les étapes de filtrage et d'acquisition. De plus, de nombreuses réserves doivent être posées sur la définition des patrons syntaxiques pertinents. En effet, ces patrons sont la plupart du temps de type nominal et ne couvrent ainsi qu'un

sous-ensemble des unités polylexicales existantes. Parallèlement, leur définition dépend forcément de la langue utilisée et ne permet donc pas l'élaboration de systèmes facilement portables.

Afin de pallier toutes ces difficultés, nous proposons une solution originale basée sur l'acquisition automatique de patrons syntaxiques combinée à l'analyse des cohésions entre unités lexicales. Pour reprendre la terminologie employée par D. Bourigault [15], nous parlerons d'apprentissage endogène. Ainsi, nous basons notre proposition à partir de deux hypothèses. Premièrement, un grand nombre d'études lexicographiques et terminologiques défendent que les associations lexicales mettent en évidence des structures syntaxiques bien connues : [51] et [111]. Deuxièmement, les associations lexicales couvrent une grande surface des énoncés. Ainsi, d'après B. Habert et C. Jacquemin [43], les unités polylexicales peuvent représenter jusqu'à un cinquième de la surface totale d'un énoncé. Par conséquent, à partir de ces deux hypothèses, il est raisonnable de penser que les patrons syntaxiques des unités polylexicales peuvent être appris à partir de notre architecture en l'appliquant sur des textes d'étiquettes morpho-syntaxiques. Ainsi, il suffira de conjuguer le degré de cohésion lexicale de chaque $N$-gram positionnel et le degré de cohésion de son $N$-gram positionnel d'étiquettes morpho-syntaxiques associé, pour définir un degré de cohésion global.

L'idée sous-jacente de cette analyse est la suivante. Reprenons le cas de la suite *de petites morues* qui est repérée comme étant pertinente par l'analyse exclusive des unités lexicales. Il est fort probable que sa structure syntaxique $PREP \quad ADJ \quad N$ ne soit pas des plus pertinentes comparativement à celle de la suite *petites morues* qui peut être schématisée de la forme suivante : $ADJ \quad N$. Ainsi, il faut espérer que la conjugaison de l'information morpho-syntaxique et de l'information lexicale aide dans ce cas de figure, la suite de formes graphiques la plus courte c'est-à-dire *petites morues*. Afin d'éclaircir notre proposition, nous illustrons ce nouveau processus d'extraction dans la figure 10.2.

Comme le montre la figure 10.2, le texte d'entrée doit être préalablement étiqueté morpho-syntaxiquement. Dans ce contexte, n'importe quel ensemble d'étiquettes peut être utilisé. L'énoncé est ensuite divisé en deux sous-corpora : l'un de formes graphiques et l'autre d'étiquettes morpho-syntaxiques. Dans un deuxième temps, chaque sous-corpus est segmenté en $N$-grams positionnels. Pour chaque $N$-gram calculé, une mesure de

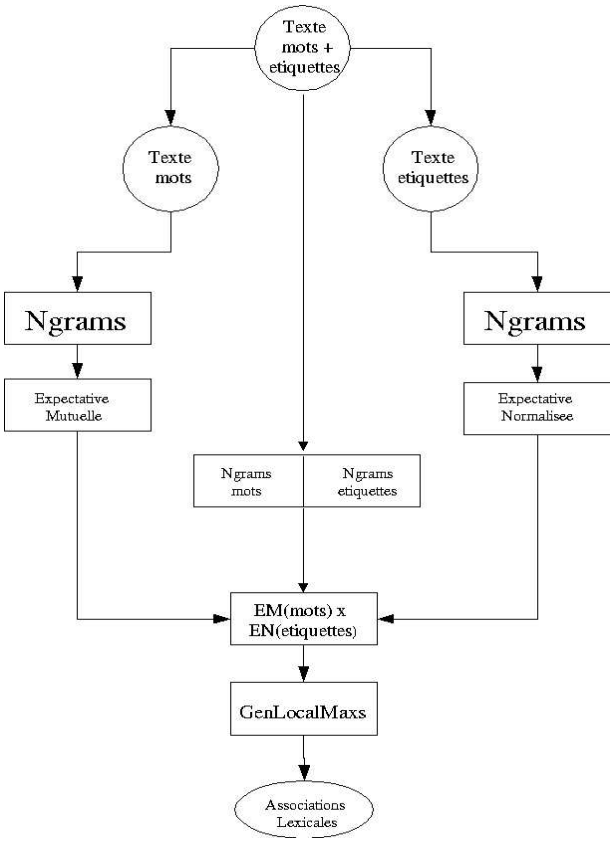


FIG. 10.2 – Apprentissage Endogène

cohésion lui est finalement associée. Dans le cadre des $N$-grams positionnels de formes graphiques, l'Expectative Mutuelle a logiquement été choisie comme mesure de référence. En ce qui concerne les $N$-grams positionnels d'étiquettes, nous avons choisi d'utiliser l'Expectative Normalisée. En effet, à partir des expériences précédentes sur les corpora d'étiquettes, nous nous sommes rendu compte que la pondération par la fréquence relative n'est pas souhaitable, celle-ci donnant effectivement trop d'importance aux associations fréquentes.

Ainsi, à cette étape du processus, nous avons calculé la valeur de cohésion de chaque $N$-gram positionnel de formes graphiques et d'étiquettes. Notre objectif est donc maintenant de conjuguer l'information obtenue à partir des formes graphiques avec celle des étiquettes. A cette fin, le texte initial est segmenté en $N$-grams positionnels de formes graphiques auxquels sont associés les $N$-grams positionnels d'étiquettes correspondants. Il suffit alors pour un $N$-gram positionnel donné de combiner son Expectative Mutuelle — information venue des formes graphiques — avec son Expectative Normalisée —

information venue des étiquettes morpho-syntaxiques. Pour y parvenir, nous avons choisi de multiplier les deux valeurs de cohésion. Ainsi, plus une séquence démontre une forte valeur de cohésion lexicale et une forte association entre ses constituants morpho-syntaxiques, plus il est probable que cette suite soit une association lexicale. Finalement, le GenLocalMaxs est appliqué naturellement pour extraire les associations lexicales les plus pertinentes, de telle forme que les sous-groupes et super-groupes sont calculés en fonction des constituants morpho-syntaxiques.

Les premières expériences mises en évidence dans [114] ont été réalisées à partir du corpus Brown et démontrent des résultats particulièrement intéressants. Ainsi, comparativement à l'utilisation exclusive des formes graphiques pour le processus d'extraction, l'introduction de connaissances morpho-syntaxiques permet d'atteindre des niveaux de performance supérieurs pour le cas des modèles trigrams et tetragrams. En effet, les effets de bord impliqués par l'occurrence fortuite des prépositions sont moins fréquents et permettent ainsi de meilleurs résultats d'extraction. Au contraire, les résultats obtenus pour les modèles digrams positionnels sont franchement inférieurs. Ceci est fondamentalement dû aux fortes fréquences des digrams positionnels d'étiquettes qui sous-estiment par conséquent les forces de cohésion lexicales.

Ces résultats nous ont donc mené à penser à une pondération des valeurs de cohésion. Ainsi, il semblerait nécessaire d'équilibrer de forme compensatoire chacune des mesures de cohésion. Dans ce contexte, pondérer favorablement les valeurs de cohésion lexicales devrait impliquer nécessairement une pondération négative des valeurs de cohésion morpho-syntaxiques. Cette situation peut facilement être résumée par l'équation 10.1 où $ASS$ serait la nouvelle mesure d'association d'un $N$-gram positionnel enrichi par l'ensemble des étiquettes de ses constituants i.e. $[p_{11}u_1t_1p_{12}u_2t_2...p_{1n}u_nt_n]$ où $\forall i = 1..n, t_i$ correspond à l'étiquette morpho-syntaxique de l'unité textuelle $u_i$.

$$\begin{aligned} ASS([p_{11}u_1t_1p_{12}u_2t_2...p_{1n}u_nt_n]) = \; & EM([p_{11}u_1p_{12}u_2...p_{1n}u_n])^{\alpha} * \\ & EN([p_{11}t_1p_{12}t_2...p_{1n}t_n])^{1-\alpha} \end{aligned} \tag{10.1}$$

Dans ces conditions, il suffirait alors de déterminer le coefficient $\alpha$ idéal. Cette opération

peut être faite à partir de techniques d'apprentissage automatique telles que les algorithmes génétiques ou encore le recuit simulé[4].

En dehors des améliorations possibles du système d'extraction, il est bien évident qu'un nombre important de travaux dans le domaine de l'organisation des données reste encore à faire. Nous entrerions alors dans le domaine de l'acquisition des connaissances. Toutefois, par souci de modération, nous limiterons ici notre exposé en laissant volontairement cette hypothèse de côté. De nombreuses références utiles pourront être trouvées dans l'article de N. Aussenac-Gilles *et al.* [115].

Tout au long de ce rapport, nous nous sommes efforcés de mettre en valeur les atouts de notre architecture sans pour autant cacher ses lacunes. Nous espérons ainsi avoir contribué positivement au développement de systèmes d'acquisition automatique flexibles et versatiles du traitement automatique du langage naturel. Dans ce cadre, nous avons introduit de nouveaux concepts qui nous semblent originaux et innovants et qui nous laissent à penser qu'un futur riche en perspectives se propose à nous. Nous les mentionnons encore une fois : les modèles $N$-grams positionnels, l'Expectative Mutuelle, la normalisation de mesures d'association binaires et enfin l'algorithme de sélection GenLocalMaxs. Ce futur ne se fera cependant pas sans réflexion, rigueur et ingéniosité. En effet, cette première étape qui présage un bon nombre d'années à venir de dur labeur, n'est qu'une goutte d'eau dans un océan d'informations ■

[4] Traduction de l'Anglais *Simulated Annealing*.

# Annexe A

# Propriété de Récursivité de l'Expectative Normalisée

La définition de l'Expectative Normalisée dans un espace probabilisé bien fondé permet d'étudier ses caractéristiques à partir des propriétés inhérentes à la théorie des probabilités. En particulier, l'Expectative Normalisée démontre la propriété de récursivité qui est un atout fondamental dans le cadre de son implémentation informatique. Avant de démontrer cette propriété, nous définissons d'abord un ensemble de notations pour lesquelles l'accent circonflexe correspond à une convention fréquemment utilisée en Algèbre et qui consiste à définir le terme omis d'une suite donnée. Ces notations sont définies comme suit.

$$X_{i0} = f([p_{11}u_1 \dots p_{1N}u_N])$$

$$X_{i1} = f([p_{11}u_1 \dots p_{1\hat{i}1}u_{\hat{i}1} \dots p_{1N}u_N]), \forall i1 = 1..N$$

$$X_{i1,i2} = f([p_{11}u_1 \dots p_{1\hat{i}1}u_{\hat{i}1} \dots p_{1\hat{i}2}u_{\hat{i}2} \dots p_{1N}u_N]), \begin{cases} \forall i1 = 1..N, \\ i2 = 1..N-1, \\ \wedge \quad i1 \neq i2 \end{cases}$$

Si l'on répète ce procédé jusquà $i(N-1)$, on obtient la dernière notation suivante.

$$X_{i1,i2,\dots,i(N-1)} = f([p_{11}u_1 \dots p_{1\hat{i}1}u_{\hat{i}1} \dots p_{1\hat{i}2}u_{\hat{i}2} \dots p_{1i(\hat{N}-1)}u_{i(\hat{N}-1)} \dots p_{1N}u_N])$$

$$\begin{cases} \forall i1 = 1..N, i2 = 1..N-1, \ldots, i(N-1) = 1..2 \\ \wedge \quad i1 \neq i2 \neq \ldots \neq i(N-1) \end{cases}$$

A partir de ces nouvelles variables de notation, l'EN d'un $N$-gram positionnel peut s'écrire de la forme suivante. On notera que l'ACM peut être défini par la formule $\frac{1}{N} \times \sum_{i1=1}^{N} f([p_{11}u_1 \ldots p_{\hat{1}i1}u_{\hat{i1}} \ldots p_{1N}u_N])$.

$$\begin{aligned} EN([p_{11}u_1 \ldots p_{1i1}u_{i1} \ldots p_{1N}u_N]) &= \frac{f([p_{11}u_1 \ldots p_{1i1}u_{i1} \ldots p_{1N}u_N])}{\frac{1}{N} \times \sum_{i1=1}^{N} f([p_{11}u_1 \ldots p_{\hat{1}i1}u_{\hat{i1}} \ldots p_{1N}u_N])} \\ &= \frac{X_{i0}}{\frac{1}{N} \times \sum_{i1=1}^{N} X_{i1}} \end{aligned} \tag{A.1}$$

Si l'on applique ce même calcul à un sous-groupe de rang $N-1$ du $N$-gram considéré, on obtient l'équation suivante. On notera $EN(N-1, i1)$ le terme qui correspond à l'Expectative Normalisée du sous-groupe de rang $N-1$ du $N$-gram positionnel considéré dans lequel l'UT d'indice $i1$ a été extraite.

$$EN([p_{11}u1 \ldots p_{\hat{1}i1}u_{\hat{i1}} \ldots p_{1N}u_N]) = \frac{X_{i1}}{\frac{1}{N-1} \times \sum_{i2=1}^{N-1} X_{i1,i2}} \equiv EN(N-1, i1) \tag{A.2}$$

Si l'on remplace le résultat de l'équation A.2 dans l'équation A.1, on obtient l'égalité suivante.

$$EN([p_{11}u_1 \ldots p_{1i1}u_{i1} \ldots p_{1N}u_N]) = \frac{X_{i0}}{\frac{1}{N} \times \sum_{i1=1}^{N} \left( EN(N-1, i1) \times \frac{1}{N-1} \times \sum_{i2=1}^{N-1} X_{i1,i2} \right)} \tag{A.3}$$

Si l'on répète ce processus récursivement, nous pouvons conclure le résultat de l'équation A.4 où les indices $ij$, tel que $\forall j, j = 1..N-1$ ne sont jamais égaux entre eux.

$$EN([p_{11}u_1 \dots p_{1i1}u_{i1} \dots p_{1N}u_N]) = \frac{N! \times X_{i0}}{\sum_{i1=1}^{N}\sum_{i2=1}^{N-1}\cdots\sum_{i(N-1)=1}^{2}\begin{pmatrix} EN(N-1,i1)EN(N-2,i1i2)\dots \\ EN(2,i1..i(N-2))EN(1,i1..i(N-1)) \end{pmatrix}} \quad \text{(A.4)}$$

L'EN d'un $N$-gram positionnel peut donc être calculée à partir des EN de l'ensemble de tous ses sous-groupes montrant ainsi son caractère récursif. Cette propriété est particulièrement intéressante pour son implémentation bénéficiant du développement d'algorithmes performants.

# Bibliographie


[1] L. Lebart and A. Salem, *Statistique textuelle.* Paris : Dunod, 1994.

[2] J. Lyons, *Introduction to theoretical linguistics.* Cambridge : Cambridge University Press, 1968.

[3] S. Jones and J. Sinclair, "English lexical collocations : A study in computational linguistics," *Cahiers de Lexicologie*, vol. 23, no. 2, pp. 15–61, 1974.

[4] A. Pawley and H. Sider, "Two puzzles for linguistic theory : Native-like selection and native-like fluency," *Language and Communication*, pp. 191–226, 1983.

[5] O. Jespersen, *La Philosophie de la Grammaire.* Paris : Les Editions de Minuit, 1971.

[6] R. Betts and D. Marrable, "Free text vs controlled vocabulary, retrieval precision and recall over large databases," *15th Online Information*, pp. 153–165, 1991.

[7] G. Gréfenstette, *Explorations In Automatic Thesaurus Discovery.* Boston Dordrecht London : Kluwer Academic Publishers, 1994.

[8] P. Gamallo, "Bases lexicales et systèmes d'héritage conduits par la relation de méréonymie," *Revue Française de Lingusitique Appliquée*, vol. 5, no. 2, pp. 45–56, 2000.

[9] D. Evans and R. Lefferts, "Design and evaluation of the clarit-trec-2 system," *TREC93*, pp. 137–150, 1993.

[10] P. Quaresma, I. Rodrigues, and G. Lopes, "Pgr project : the portuguese attorney general decisions on the web," *The Law in the Information Society*, 1998.

[11] P. Dowing, "On the creation and use of english compound nouns," *Language*, vol. 53, no. 4, pp. 810–842, 1977.

[12] D. Corbin, *La formation des mots : structures et interprétation.* Villeneuve d'Ascq : Presses Universitaires de Lille, 1991.

[13] I. Mel'cuk, *Dependency syntax : theory and practice.* New York : SUNY, 1988.

[14] M. Silberztein, *Dictionnaires électroniques et analyse automatique de textes. Le système INTEX.* Paris : Masson, 1993.

[15] D. Bourigault, "Analyse syntaxique locale pour le repérage de termes complexes dans un texte," *Traitement Automatique des Langues*, vol. 34, no. 2, pp. 105–117, 1993.

[16] S. David and P. Plante, "Termino version 1.0," tech. rep., Centre d'Analyse de Textes par Ordinateur, Université du Québec, Canada, 1990.

[17] B. Habert, A. Nazarenko, and A. Salem, *Les Linguistiques du Corpus.* Paris : Armand Collin, 1997.

[18] L. Ramshaw and M. Marcus, "Text chunking using transformation-based learning," *3rd ACL Workshop on Very Large Corpora*, pp. 82–94, 1995.

[19] S. Abney, "Rapid incremental parsing with repair," *Waterloo Conference on Electronic Text Research*, pp. 1–9, 1990.

[20] N. Marques, A. Coelho, and J. Lopes, "Mining subcategorization information by using multiple feature loglinear models," *10th Computational Liguistics in the Netherlands*, 1999.

[21] B. Daille, "Study and implementation of combined techniques for automatic extraction of terminology," *The balancing act combining symbolic and statistical approaches to language*, pp. 49–66, 1996.

[22] J. Justeson and S. Katz, "Technical terminology : Some linguistic properties and an algorithm for identification in text," tech. rep., IBM, 1993.

[23] C. Enguehard, "Acquisition de terminologie à partir de gros corpus," *Informatique et Langue Naturelle*, pp. 373–384, 1993.

[24] R. Feldman, M. Fresko, Y. Kinar, Y. Lindell, O. Liphstat, M. Rajman, Y. Schler, and O. Zamir, "Text mining at the term level," *PKDD'98*, vol. Lecture Notes in AI 1510, pp. 65–73, 1998.

[25] M. H. Picard, "Construction de terminologies : une chaîne de traitement supportée par un atelier intégrant outils linguistiques et statistiques," tech. rep., EDF-GDR, France, 1996.

[26] I. Dagan and K. Church, "Termight : Identifying and translating technical terminology," *4th Conference on Applied Natural Language Processing*, pp. 34–40, 1994.

[27] K. Church and P. Hanks, "Word association norms mutual information and lexicography," *Computational Linguistics*, vol. 16, no. 1, pp. 22–29, 1990.

[28] T. Dunning, "Accurate methods for the statistics of surprise and coincidence," *Computational Linguistics*, vol. 19, no. 1, pp. 61–74, 1993.

[29] W. Gale and K. Church, "Concordances for parallel texts," *7th Annual Conference of the UW Center for the New OED and Text Research, Using Corpora*, pp. 40–62, 1991.

[30] J. Silva, G. Dias, S. Guilloré, and G. Lopes, "Using localmaxs algorithm for the extraction of contiguous and non-contiguous multiword lexical units," *9th Portuguese Conference on Artificial Intelligence*, vol. 1695, pp. 113–132, September 1999.

[31] F. Smadja, "Retrieving collocations from text : Xtract," *Computational Linguistics*, vol. 19, no. 1, pp. 143–177, 1993.

[32] S. Shimohata, T. Sugio, and J. Nagata, "Retrieving collocations by co-occurrences and word order constraints," *35th annual meeting of the Association for Computational Linguistics*, pp. 476–481, 1997.

[33] K. Frantzi and S. Ananiadou, "Extracting nested collocations," *International Conference on Computational Linguistics (COLING)*, pp. 41–46, 1996.

[34] A. Salem, *La pratique des segments répétés*. Paris : Klincksieck, 1987.

[35] K. W. Church, "One term or two ?," *SIGIR*, pp. 310–318, 1995.

[36] E. Gaussier, G. Grefenstette, and M. Schulze, "Traitements du langage naturel et recherche d'information : Quelques expériences sur le français," *FRANCIL'97*, pp. 9–14, 1997.

[37] G. Gréfenstette, *Cross-Language Information Retrieval*. Kluwer Editions, 1998.

[38] M. Sussna, "Word sense disambiguation for rree-text indexing using a massive semantic network," *Second International Conference on Information and Knowledge Management*, pp. 67–74, 1993.

[39] O. Mason, "The weight of words : an investigation of lexical gravity," *PALC'97*, pp. 361–375, 1997.

[40] J. Firth, "A synopsis of linguistic theory 1930-1955," *Studies in Linguistic Analysis*, pp. 1–32, 1957.

[41] M. Halliday, "Lexis as a linguistic level," *Memory of J.R. Firth*, pp. 148–162, 1966.

[42] C. D. Manning and H. Shutze, *Foundations of Statistical Natural Language Processing.* London : MIT Press, 1999.

[43] B. Habert and C. Jacquemin, "Noms composés, termes, dénominations complexes : Problématiques linguistiques et traitements automatiques," *Traitement Automatique des Langues*, vol. 34, no. 2, pp. 5–41, 1993.

[44] R. Agrawal, H. Mannila, R. Srikant, H. Toivonen, and A. Verkamo, "Fast discovery of association rules," *Advances in Knowledge Discovery and Data Mining*, 1996.

[45] F. Smadja, K. McKeown, and V. Hatzivassiloglou, "Translating collocations for bilingual lexicons : A statistical approach," *Computational Linguistics*, vol. 22, no. 1, 1996.

[46] E. Naulleau and M. Monteil, "Mise en oeuvre d'un outil d'extraction de groupes nominaux à vocation terminologique : approches linguistiques et informatiques," tech. rep., EDF-DER, France, 1992.

[47] D. A. Evans and C. Zhai, "Noun-phrase analysis in unrestricted text for information retrieval," *34th Annual Meeting of the Association for Computational Linguistics*, pp. 17–24, 1996.

[48] D. A. Evans, "Concept management in text via natural language processing : the clarit approach," *Working Notes of the 1990 AAAI Symposium on Text-Based Intelligent Systems*, pp. 93–95, March 1990.

[49] U. Heid, "Extracting terminologically relevant collocations from german technical texts," *TKE'99 International Congress on Terminology and Knowledge Engineering*, pp. 241–255, 1999.

[50] B. Daille, "Identification des adjectifs relationnels en corpus," *Traitement Automatique des Langues Naturelles*, pp. 105–114, July 1999.

[51] G. Gross, *Les expressions figées en français.* Paris : Ophrys, 1996.

[52] J. Lyons, *Introduction to theoretical linguistics.* Cambridge : Cambridge University Press, 1968.

[53] J. Sinclair, "Preliminary recommendations on corpus typology," tech. rep., EAGLES (Expert Advisory Group on Language Engineering Standards), 1996.

[54] N. Ménard, *Mesure de la recherche lexicale, théorie et vérifications expérimentales.* Paris : Slatkine-Champion, 1983.

[55] H. K. cera and N. Francis, *Computational analysis of present day american English.* Providence : Brown University Press, 1967.

[56] D. Labbé, *Normes de dépouillement et procédures d'analyse des textes politiques.* Grenoble : CERAT, 1990.

[57] M. Demonet, A. Geoffroy, J. Gouaze, P. Lafon, M. Mouillaud, and M. Tournier, *Des tracts en Mai 68. Mesures de Vocabulaire et de Contenu.* Paris : Armand Colin et Presses de la Fondation Nationale des Sciences Politiques, 1975.

[58] D. Meyer, R. Schvaneveldt, and M. Ruddy, "Loci of contextual effects on visual word recognition," *Attention and Performance V*, 1975.

[59] C. Shannon, "A mathematical theory of communication," *Bell System Technical Journal*, vol. 27, pp. 379–423, 1948.

[60] R. Schneider and I. Renz, "The relevance of frequency lists for error correction and robust lemmatization," *5emes Journées Internationales d'Analyse de Données Textuelles (JADT)*, 2000.

[61] G. Chartron, *Analyse des corpus de données textuelles, sondage de flux d'information.* PhD thesis, Université. Paris 7, Paris, 1988.

[62] P. Kim and Y. Cho, "Indexing compound words from korean texts using mutual information," *Natural Language Processing Rim Symposium*, pp. 85–92, 1993.

[63] D. Maynard and S. Ananiadou, "Identifying contextual information for multiword term extraction," *TKE'99 International Congress on Terminology and Knowledge Engineering*, pp. 212–221, 1998.

[64] P. Downing, "On the creation and use of english compound nouns," *Language*, vol. 53, no. 4, pp. 810–842, 1977.

[65] P. Sébillot and P. Bouher, "Interprétation et génération automatique de noms composés anglais à l'aide de formes logiques," *Traitement Automatique des Langues*, vol. 34, no. 2, pp. 89–104, 1993.

[66] Y. Choueka, T. Klein, and E. Neuwitz, "Automatic retrieval of frequent idiomatic and collocation expressions in a large corpus," *Journal for Literary and Linguistic Computing*, vol. 4, pp. 34–38, 1983.

[67] M. Benson, "The structure of the collocational dictionary," *International Journal of Lexicography*, vol. 2, no. 1, pp. 1–14, 1989.

[68] C. Jacquemin, "Quelques exemples d'application du traitement automatique des langues en accès à l'information," *5emes Journées Internationales d'Analyse de Données Textuelles (JADT)*, vol. 1, 2000.

[69] E. Planas and O. Furuse, "Formalizing translation memories," *MT Summit VII*, pp. 331–339, 1999.

[70] E. Brill, *Lexical Disambiguation*. Marcel Dekker, 1998.

[71] C. Shannon, "Prediction and entropy of printed english," *Bell System Technical Journal*, vol. 30, pp. 50–64, 1951.

[72] A. V. den Bosch, "Instance families in memory-based language learning," *Computational Linguistics in the Netherlands*, pp. 3–17, 1998.

[73] H. Barkema, "Determining the syntactic flexibility of idioms," *Creating and Using English Language Corpora*, pp. 39–52, 1994.

[74] G. K. Zipf, *The psychobiology of language, an introduction to dynamic philology*. Boston : Houghton-Mifflin, 1935.

[75] B. Mandelbrot, "Les constantes chiffrées du discours," *Le Langage*, vol. 25, 1968.

[76] P. T. D. Labbé and D. Serant, *Etude sur la recherche et la structure lexicales*. Paris-Genève : Slatkine-Champion, 1988.

[77] J. D. E. Rivals, O. Delagrange and M. Dauchet, "A first step towards chromosome analysis by compression algorithms," *International IEEE Symposium on Intelligence in Neural and Biological Systems*, 1995.

[78] M. Bécue and R. Peiro, "Les quasi-segments pour une classification automatique des réponses ouvertes," *2emes Journées Internationales d'Analyse des Données Textuelles (JADT)*, pp. 310–325, 1993.

[79] R. Fano, *Transmission of Information : A Statistical Theory of Communications*. MA : MIT Press, 1961.

[80] P. Resnik, "Selectional constraints : an information-theoric model and its computational realization," *Cognition*, vol. 61, pp. 127–159, 1996.

[81] S. Katz, N. Johnson, and C. Read, *Encyclopedia of Statistical Sciences*. New York, Chichester, Brisbane, Toronto, Singapoure : John Wiley and Sons, 1982.

[82] G. Bhattacharyya and R. A. Johnson, *Statistical Concepts and Methods*. New York, Chichester, Brisbane, Toronto, Singapoure : John Wiley and Sons, 1977.

[83] I. Dagan, "Simple methods can do a lot : The generality of vector models in language processing," *Traitement Automatique des Langues Naturelles*, July 1999.

[84] I. D. S. Argamon-Engelson and Y. Krymolowski, "A memory-based approach to learning shallow natural language patterns," *Experimental and Theoretical Artificial Intelligence*, vol. 11, no. 3, pp. 67–73, 1998.

[85] E. Frank, G. Paynter, I. Witten, C. Gutwin, and C. Nevill-Manning, "Domain-specific keyphrase extraction," *Sixteenth International Joint Conference on Artificial Intelligence*, pp. 668–673, 1999.

[86] P. Turney, "Learning to extract keyphrases from texts," tech. rep., National Research Council - Institute for Information Technology, Canada, 1992.

[87] H. Thompson, "The strategic role of evaluation in natural language processing and speech technology," tech. rep., Human Communication Research Centre - University of Edinburgh, Edimburg (Écosse), 1992.

[88] H. Nomura and H. Isahara, "Jeida's criteria on machine translation evaluation," *International Symposium on Natural Language Understanding and AI*, 1992.

[89] D. Bourigault and B. Habert, "Evaluation of terminology extractors : Principles and experiments," *First LREC*, pp. 28–31, 1998.

[90] J. Galliers and K. S. Jones, "Evaluating natural language processing systems," tech. rep., University of Cambridge, Angleterre, 1993.

[91] ISO, "Information technology-software product evaluation, quality characteristics and guidelines for their use," tech. rep., International Organization for Standardization, 1991.

[92] L. Balkan, "Test suites for natural language processing," *Translating and the computer*, vol. 16, pp. 51–58, 1994.

[93] M. King and K. Falkedal, "Using test suites in evaluation of mt systems," *28th Annual Meeting of the Association for Computational Linguistics*, vol. 2, pp. 211–216, 1990.

[94] J. Nerbonne, K. Netter, A. Klein, and L. Dickmann, "A diagnostic tool for german syntax," *Machine Translation*, vol. 8, pp. 85–107, 1993.

[95] I. Blank, "Computer-aided analysis of multilingual patent documentation," *First LREC*, pp. 765–771, 1998.

[96] R. Baesa-Yates and B. ribeiro Neto, *Modern Information Retrieval.* Addison Wesley Longman, 1999.

[97] D. N. Lapedes, *Dictionary of Physics and Mathematics.* McGraw-Hill Editions, 1978.

[98] K. Frantzi and S. Ananiadou, "A hybrid approach to term recognition," *NLP and Industrial Applications*, pp. 93–98, 1996.

[99] C. Zhai, "Exploiting context to identify lexical atoms : a statistical view of linguistic context," *International and Interdisciplinary Conference on Modeling and Using Context*, pp. 119–129, Février 1997.

[100] G. Dias, S. Vintar, S. Guilloré, and G. Lopes, "Identifying and integrating terminologically relevant multiword units in the ijs-elan slovene-english parallel corpus," *10th Computational Linguistics in the Netherlands*, pp. 29–40, November 1999.

[101] G. Dias, J.-H. Kaalep, and K. Muischnek, "Automatic extraction of multiword units for estonian : a comparison between annotated and non-annotated corpora," *International Futuristic Conference on Language Development*, March 2000.

[102] G. Dias, S. Guilloré, and G. Lopes, "Benefiting from multi-domain corpora for extracting terminologically relevant multiword lexical units," *9th EURALEX International Congress*, pp. 339–350, 2000.

[103] G. Dias, S. Guilloré, and G. Lopes, "Mutual expectation : a measure for multiword lexical unit extraction," *Venezia per il Trattamento Automatico delle Lingue*, pp. 22–24, November 1999.

[104] G. Dias, S. Guilloré, and G. Lopes, "Multiword lexical units extraction," *International Symposium on Machine Translation and Computer Language Information Processing*, pp. 26–38, June 1999.

[105] G. Dias, S. Guilloré, and G. Lopes, "Multilingual aspects of multiword lexical units," *Workshop on Language Technologies-Multilingual Aspects*, pp. 8–11, July 1999.

[106] G. Dias, S. Guilloré, and G. Lopes, "Language independent automatic acquisition of rigid multiword units from unrestricted text corpora," *Traitement Automatique des Langues Naturelles*, pp. 12–17, July 1999.

[107] G. Dias, S. Guilloré, and J. Lopes, "Extraction automatique d'associations textuelles à partir de corpora non traités," *5emes Journes Internationales d'Analyse des Données Textuelles (JADT)*, pp. 213–221, March 2000.

[108] G. Gougenheim, "Une catégorie lexico-grammaticale : les locutions verbales," *Etudes de linguistique appliquée*, vol. 2, pp. 56–64, March 1971.

[109] G. Dias, J.-H. Kaalep, and K. Muischnek, "Automatic extraction of verb phrases from annotated corpora : A linguistic evaluation for estonian," *Workshop on Collocation of the joint ACL-EACL meeting*, July 2001.

[110] G. Gross, "Typologie des adjectivaux," *Analyse et Synthèse dans les Langues Romanes et Slaves*, pp. 163–178, 1991.

[111] M. Noally, *Le substantif épithète*. Paris : PUF, 1990.

[112] A. Ribeiro, G. Dias, G. Lopes, and J. Mexia, "Cognates alignment," *Machine Translation Summit VIII (MT Summit VIII)*, September 2001.

[113] G. Dias, S. Guilloré, and G. Lopes, "Mining textual associations in text corpora," *Workshop on Text Mining of the Sixth ACM SIGKDD International Conference on Knowledge Discovery an Data Mining*, pp. 92–95, August 2000.

[114] G. Dias, S. Guilloré, and G. Lopes, "Combining linguistics with statistics for multiword term extraction : a fruitful association ?," *RIAO'2000 Content-Based Multimedia Information Access*, pp. 1–20, April 2000.

[115] N. Aussenac-Gilles, D. Bourigault, A. Condamines, and C. Gros, "How can knowledge acuisition benefit from terminology," *9th Knowledge Acquisition Workshop (KAW)*, 1995.