FRÉDÉRICK BRAULT

FORCES ET FAIBLESSES DE L'UTILISATION DE TRIGRAMS DANS L'ÉTIQUETAGE AUTOMATIQUE DU FRANÇAIS

Exploration à partir des homographes de type verbe-substantif

Mémoire présenté
à la Faculté des études supérieures de l'Université Laval
dans le cadre du programme de maîtrise en linguistique
pour l'obtention du grade de maître ès arts (M. A.)

FACULTÉ DES LETTRES
UNIVERSITÉ LAVAL
QUÉBEC

OCTOBRE 2004

Résumé

Ce mémoire porte sur l'étiquetage automatique de texte français, c'est-à-dire l'attribution, par un programme informatique appelé *étiqueteur*, de la nature grammaticale des mots d'un texte français. En particulier, ce mémoire explore les forces et les faiblesses de l'utilisation du modèle mathématique des trigrams pour cette tâche. L'efficacité du modèle des trigrams est évaluée à l'aide d'observations sur la désambiguïsation des homographes de type verbe/substantif en français, c'est-à-dire, des mots dont la graphie est la même selon qu'ils soient verbe ou substantif (ex. : *ferme*).

Ce mémoire tente de répondre à trois questions :

1. Pourquoi les étiqueteurs à modèle mathématique comme les trigrams réussissent-ils à 95%?
2. Qu'est-ce qui empêche d'améliorer ces performances?
3. Comment des connaissances linguistiques peuvent-elles permettre d'améliorer ces performances?

En rapport à ces questions, les résultats obtenus lors de ces travaux montrent que :

1. les structures syntaxiques sont suffisamment récurrentes pour permettre aux trigrams de saisir en grande partie les règles syntaxiques nécessaires à la désambiguïsation;
2. le calcul d'un taux de succès général dissimule, d'un point de vue linguistique, des décisions incohérentes du modèle des trigrams qu'ils seraient difficile de rectifier simplement en modifiant les trigrams;
3. la connaissance de contraintes syntaxiques permet d'analyser plus en détail le comportement du modèle des trigrams et de suggérer, en conséquence, des solutions pour améliorer le taux de succès d'un étiqueteur.

Remerciements

Merci à mon directeur, M. Pierre Auger, d'abord de m'avoir fait confiance, ensuite de m'avoir laissé la liberté et le temps nécessaire d'enquêter sur des questions qui m'étaient apparues intéressantes.

Merci à Mme Anne Abeillé, professeure à l'Université Paris VII, de m'avoir prêté une partie du corpus étiqueté *Paris VII – Le Monde*.

Merci au *Fonds Québécois de la Recherche sur la Nature et les Technologies* (anciennement le FCAR), d'avoir soutenu mes études de deuxième cycle par une bourse de maîtrise en recherche.

Merci à Isabelle, pour tout.

TABLE DES MATIÈRES

LISTE DES ILLUSTRATIONS

INTRODUCTION

Parmi les tâches de traitement automatique de la langue écrite, il en est une qui, depuis le début, pose des problèmes dont certains ne sont pas encore résolus : reconnaître la nature des mots. En effet, rendre un ordinateur capable de reconnaître automatiquement la nature lexicale d'un mot pour savoir s'il est, entre autres, un verbe, un substantif ou un adjectif, est tout un défi. Pour un ordinateur, un mot n'est rien d'autre qu'une suite de caractères. Qui plus est, cette suite peut, selon le cas, faire partie de plus d'une catégorie syntaxique. Par exemple, le mot «ferme» est, selon le contexte, un substantif, s'il réfère à un bâtiment agricole, un adjectif, s'il réfère à la qualité de ce qui est ni trop mou ni trop dur, ou une forme fléchie du verbe «fermer» comme dans «Il *ferme* la porte». «Ferme» peut même être un adverbe comme dans la phrase «Il négocie *ferme*».

Dans le cas où l'on souhaite reconnaître automatiquement la nature lexicale des mots d'un texte, la consultation d'un dictionnaire conçu à cette fin ne permet pas de le faire adéquatement. En effet, dans plusieurs cas tel que l'exemple *ferme* apporté plus haut, le dictionnaire ne peut que témoigner de l'ambiguïté en constatant que plusieurs catégories peuvent être associées à une même suite graphique. Le problème est de taille non seulement parce qu'il est complexe à résoudre, mais aussi en raison de sa fréquence. Anderson (1987 : 92-93) a observé que dans ses corpus, la quantité d'homographes est constante. Le pourcentage oscille entre 27% et 29% et ce pourcentage est indépendant de la longueur des textes.

Même lorsque l'on cible une catégorie en particulier, comme le montrent les travaux de Maegaard et Spang-Hanssen (1978), la difficulté n'est pas réduite. Ces derniers ont évalué qu'un peu plus de 30% des formes verbales du français sont homographes. La difficulté de la reconnaissance automatique de la nature des mots d'un texte se trouve donc clairement dans la désambiguïsation, c'est-à-dire dans le bon choix parmi toutes les parties du discours possibles pour un mot.

Les étiqueteurs sont des programmes qui ont comme rôle d'identifier la nature grammaticale des mots. Ils jouent un rôle important car la reconnaissance de la nature des mots est une tâche qui fait partie de plusieurs applications. En repérage d'information, la

nature des mots peut faire la distinction entre certains concepts dont l'orthographe est le même. Enfin, tous les autres traitements automatiques de la langue tels que l'analyse syntaxique, le repérage des unités complexes, le résumé automatique ont recours, d'une façon où d'une autre, à la manipulation de la nature des mots.

Pour sa part, le présent mémoire porte sur cette tâche linguistique qui continue de faire l'objet de nombreuses recherches : l'étiquetage automatique. Il se divise en six parties. La première discute de ce qu'est la linguistique informatique et situe le présent travail à l'intérieur de cette discipline. La deuxième partie expose les objectifs du travail ainsi que les motivations derrière ces objectifs. La troisième partie fait office d'état de la question et expose les problèmes, autant d'ordre pratique que théorique, auxquels il faut faire face pour automatiser une telle tâche. Trois types d'approches y sont présentés en ordre chronologique, soit l'approche linguistique, l'approche statistique et l'apprentissage automatique de règles grammaticales. Le quatrième chapitre décrit les étapes qui ont mené à la création d'un prototype d'étiqueteur et la méthodologie utilisée pour amasser les données. La cinquième partie présente et discute des résultats obtenus. Dans cette partie, l'efficacité du modèle mathématique des n-grams est évaluée à l'aide de l'observation de la désambiguïsation des homographes de type verbe/substantif en français. Nous tentons de répondre à trois questions :

1. Pourquoi les étiqueteurs à modèle mathématique réussissent-ils à 95%?
2. Qu'est-ce qui empêche d'améliorer ces performances?
3. Comment des connaissances linguistiques peuvent-elles permettre d'améliorer ces performances?

Enfin, la sixième et dernière partie présente nos conclusions sur ce travail.

Linguistique et informatique

La linguistique est la discipline qui se concentre sur l'étude du langage. Plusieurs branches composent la linguistique et ensemble, elles tentent de cerner les phénomènes par lesquels les êtres humains peuvent communiquer entre eux. Le langage humain a ceci qui le distingue des autres modes de communication (langage de programmation, signaux routiers, etc.) et du langage qu'utilisent les animaux: il est doublement articulé. D'abord, des signes élémentaires, indivisibles et dépourvus de sens, se combinent en morphèmes qui eux, sont pourvus de sens. À l'oral, il s'agit des phonèmes et à l'écrit, s'il y a un parallèle à faire, il s'agit des graphèmes (lettres et ponctuation). C'est la première articulation. Les morphèmes, pourvus de sens, se combinent alors à leur tour pour former des syntagmes et des phrases dont le sens est généralement construit par l'addition du sens des morphèmes et des mots. C'est la deuxième articulation du langage. Cette propriété de combiner des éléments de premier ordre et de second ordre confère au langage humain un aspect qu'aucun autre mode de communication ne possède: la possibilité d’encoder une infinité de messages. En effet, il est possible de générer une infinité de phrases à partir d'éléments de base en nombre fini. En généralisant, l'objectif de la linguistique est de découvrir les lois et phénomènes qui régissent les deux axes d'articulation du langage (Carré et al. 1991 : 35). Cet objectif sous-entend que le langage est un système car seul un système est pourvu de règles. Autrement, sans système sous-jacent, on peut supposer que les phénomènes sont le fruit du hasard. La physique a découvert de nombreuses lois de la nature et continue d'explorer certaines sphères de notre environnement car le monde dans lequel nous vivons est un système. Les mêmes lois s'appliquent à tous. La chimie de son côté, recherche les règles de composition des éléments qui nous entourent. Elle en a découvertes plusieurs car là encore, il s'agit d'un système. Les éléments chimiques obéissent tous aux mêmes lois. La linguistique poursuit le même but sur son objet qu'est la langue. Elle réussira à condition que la langue soit un système. Heureusement, les découvertes linguistiques tendent à montrer que la langue se comporte comme un système. Cependant, c'est un système difficile à cerner, quelquefois flou (jugements grammaticaux et d'acceptabilité de Chomsky), mais il faut admettre qu'il en est un, car, puisque les êtres humains peuvent encoder et décoder un même message entre eux, il faut un système sous-jacent pour permettre un encodage et un décodage uniforme chez des locuteurs d'une même langue.

Tout comme la linguistique, l'informatique se subdivise également en plusieurs champs d'étude. Les concepts qui ont donné naissance à l'informatique remontent à près d'un siècle alors que des mathématiciens ont élaboré une théorie de l'information. Les premiers modèles de cette théorie, très formels, ont été appliqués à la conception de machines pour traiter l'information. Ces machines portent le nom d'*ordinateur* depuis. À ce moment très théorique, l'informatique s'est ensuite diversifiée au fur et à mesure que les machines se sont complexifiées pour aujourd'hui constituer un ensemble de domaines de plus en plus pratiques. En effet, de nos jours, l'informatique inclut des aspects matériels, logiciels en plus de l'aspect théorique toujours présent. Le côté matériel, ce que la langue anglaise appelle "hardware", se concrétise par la conception et la confection des composants électroniques internes ainsi que la fabrication des appareils périphériques. Cet aspect est essentiel au domaine car il fournit le matériel qui permet de réaliser les recherches dans les autres sphères de l'informatique. Bien sûr, tous les domaines de recherche comportent des aspects théoriques, mais la conception matérielle a naturellement des préoccupations plus pratiques que théoriques dans ses objectifs. Ce côté matériel est côtoyé par l'aspect logiciel, en anglais "software", qui, un peu moins près de la conception des machines, comprend non seulement l'architecture des ordinateurs, l'architecture des systèmes d'exploitation et l'organisation des réseaux mais aussi tout ce qui entoure la programmation: langages, compilation, analyse, techniques de programmation, algorithmique, bref, tout ce qui est lié à l'utilisation de l'ordinateur plutôt qu'à ses mécanismes et composants de base. Finalement, toujours présente, l'informatique théorique, moins proche de l'aspect matériel, se préoccupe encore des théories de l'information en élaborant les grammaires formelles et les automates. En résumé, l'informatique, dans son ensemble, poursuit comme objectif de traiter automatiquement de l'information (Fuchs et al. 1993 : 23) et ce, en fournissant un support matériel pour la stocker et des moyens de l'exploiter.

La linguistique et l'informatique sont, à prime abord, des disciplines qui n'ont rien à voir l'une avec l'autre. Pourtant, elles unissent leurs efforts pour doter une machine, l'ordinateur, de réelles capacités de communication. Le terme *traitement automatique du langage* (T.A.L.) désigne toutes les activités de conception d'outils informatiques de manipulation automatique de données textuelles (Fuchs et al. 1993, Wehrli 1997). Dans ce terme, « automatique » réfère au fait que les processus de traitement et les manipulations se

font sans l'intervention de l'être humain mais plutôt, uniquement par un ordinateur. Ce domaine de recherche est facile à concevoir mais pourtant difficile à circonscrire. En effet, il est aisé d'imaginer la combinaison de la technologie apportée par l'informatique aux connaissances des spécialistes de la linguistique. Cependant, il est difficile d'en obtenir un portrait clair. En français, « linguistique informatique » et « informatique linguistique » sont des termes que l'on rencontre dans la littérature pour désigner les champs de recherche qui ont comme objet le traitement automatique de données langagières. Comme leur nom le suggère, elles sont un mélange des deux disciplines. À première vue, par simple déduction, il semble qu'attribuer des préoccupations plus linguistiques à la première et des intérêts plus informatiques à la deuxième va de soi. Pourtant, la réalité n'est pas aussi claire. Les termes "linguistique informatique" et "informatique linguistique" sont souvent utilisés sans distinction significative d'un auteur à l'autre et, par conséquent, il est difficile de départager clairement les deux disciplines. À notre connaissance, il n'existe pas de consensus sur les frontières qui délimitent ces deux champs d'expertise. Les définitions et les termes utilisés se concurrencent et parfois même, se contredisent. Selon Fuchs et al. (1993:22), l'informatique linguistique est une branche de l'informatique qui a comme type de données, des données linguistiques. Les méthodes et techniques de traitement sont les mêmes que celles utilisées dans d'autre domaines, mais dans ce cas, appliquées à des données linguistiques. Selon Fuchs et al., le terme « informatique linguistique » renvoie à « l'utilisation de logiciels […] pour opérer des calculs sur les mots ou suites de mots contenus dans un texte […] ». Autrement dit, cela désigne « l'ensemble des traitements automatiques de données linguistiques » ce qui s'approche de la signification générale de T.A.L. La « linguistique informatique » quant à elle, est décrite par Fuchs et al. comme une branche de la linguistique qui utilise les outils développés par l'informatique pour valider des hypothèses théoriques sur le fonctionnement de la langue. Une idée partagée également par Carré et al. (1991 : 30). Cependant, Fuchs et al. observent que la plupart du temps, au sein des équipes de recherche, c'est plutôt une autre conception qui prévaut : « y sont dits relever de la "linguistique informatique" tous les travaux en traitement automatique des langues qui , d'une manière ou d'une autre, s'appuient sur des éléments d'analyse linguistique […] ». Bref, ce qui revient à dire que les termes « linguistique informatique » et « informatique linguistique » sont souvent utilisés sans distinction de sens pour désigner

tout type de traitement de la langue. Rastier et al. (1994 :2) appuient la thèse de la linguistique informatique comme branche de la linguistique théorique et appliquée et relèguent, quant à eux, l'informatique linguistique au niveau de technologie au service des découvertes linguistiques, tandis que la linguistique est une science. Wehrli (1997 : 1) parle de linguistique computationnelle comme synonyme de linguistique informatique et reconnaît la double signification de ces termes : parfois très vagues, ils désignent l'ensemble des travaux qui impliquent du traitement automatique de données linguistiques alors qu'ils devraient, dans un sens plus restreint, ne s'appliquer qu'« aux applications informatiques qui font réellement appel à des connaissances linguistiques […] ». En anglais, la situation est plus simple car la distinction entre « linguistique informatique » et « informatique linguistique » n'existe pas. En effet, seul le terme « computational linguistics » est utilisé. Morphologiquement, « computational » est un adjectif qui qualifie le nom « linguistics » ce qui veut dire qu'en anglais, il s'agit bel et bien de linguistique informatique (Fuchs et al. 1993, Rosner et Johnson 1992). Il ne s'agit pas de la seule différence entre l'anglais et le français. En effet, en anglais, on prend soin de préciser que les traitements automatiques s'opèrent sur des langages naturels pour ne par confondre avec d'autres types de langages ceux utilisés en programmation (C++, Prolog, Perl, Lisp, etc.). Cette précision est due au fait qu'en anglais, « language » signifie à la fois « langue » et « langage ». En français, il est inutile d'insister sur le fait que les langages en question sont « naturels » car le mot « langues » réfère aux modes de communication de l'être humain alors que « langages » s'emploie pour les autres modes de communication (langages de programmation, langage gestuel, langage corporel, langage des signes, etc.). Ainsi, on retrouve parfois le calque de l'anglais « Traitement Automatique des Langues Naturelles (T.A.L.N.)» alors que, de l'avis de Fuchs et al. et de Abeillé et al., on devrait tout simplement parler, en français, de traitement automatique de la *langue*.

On se rend donc compte que même si le domaine du T.A.L. est facile à concevoir, il est plus difficile de s'entendre sur ces sous-spécialisations. Pour notre part, nous nous rangeons du côté de la conception de Fuchs et al. (1993), car pour nous, la linguistique informatique a comme intérêt principal, la découverte de faits linguistiques à l'aide d'appareils et de processus informatiques. En conséquence, on pourrait représenter les recherches sur un

continuum allant de la linguistique à l'informatique, en supposant toujours un mélange des deux, comme le montre l'illustration ci-dessous.

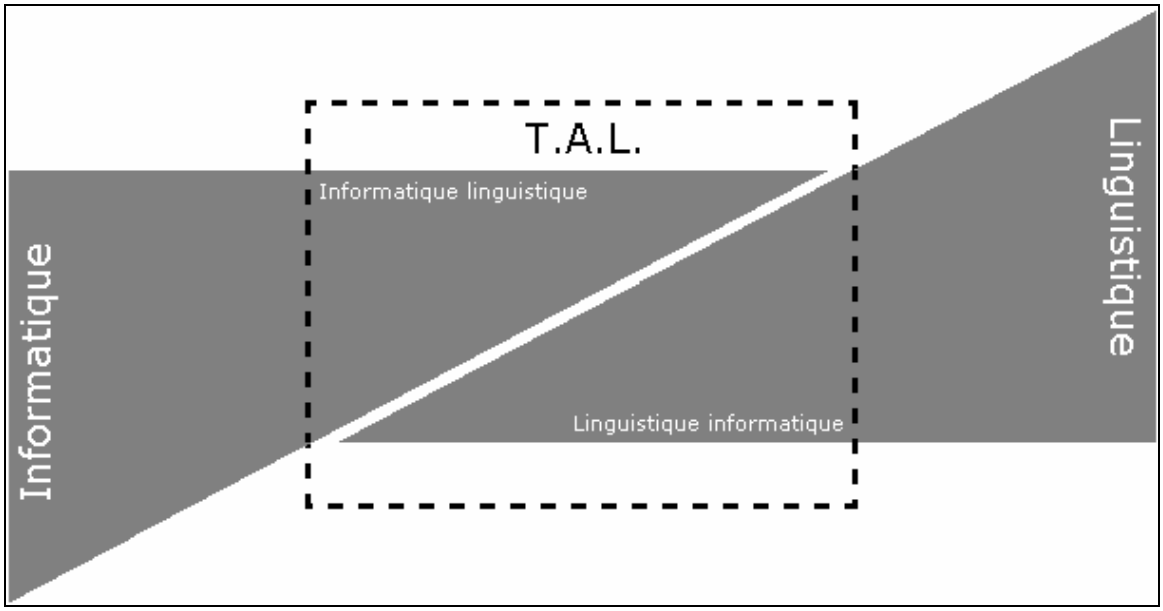


Figure 2-1 L'informatique et la linguistique s'associent pour le traitement automatique de la langue.

Aux extrémités de la portion du T.A.L., on retrouve les buts suivants: pour la linguistique informatique, explorer le langage à l'aide d'outils informatiques et pour l'informatique linguistique, manipuler l'information linguistique et automatiser des tâches linguistiques.

Le présent mémoire se trouve à mi-chemin entre les deux car il vise l'exploration d'une méthode issue de l'informatique linguistique, l'utilisation des ngrams pour étiqueter un texte, à l'aide de connaissances issues de la linguistique.

OBJECTIFS

The focus in computational linguistics has admittedly been on technology. But the same techniques promise progress on issues concerning the nature of language that have remained mysterious for so long. The time is ripe to apply them.
Abney (1996)

L'étiquetage de texte est une tâche à laquelle plusieurs chercheurs et projets de recherche ont déjà consacré de nombreuses années de travail. Ce travail s'échelonne maintenant sur plus de quatre décennies et la documentation sur le problème est abondante. L'état de la question du chapitre suivant tentera de résumer le mieux possible les points importants de la problématique qui concerne l'étiquetage automatique de texte. L'état de la question fera ressortir plus particulièrement les points qui motivent ce mémoire.

Ce qui frappe le linguiste à la lecture des articles sur la désambiguïsation automatique, c'est le faible recours aux théories et connaissances linguistiques dans les solutions proposées. En effet, dépendamment de ce que l'on considère relever de la linguistique, peu ou pas de systèmes sont entièrement linguistiques. Plus encore, la plupart des étiqueteurs utilisent peu ou pas d'informations syntaxiques ou sémantiques. Les systèmes ayant recours à des modèles mathématiques ont affiché dès le départ des taux de succès fort satisfaisant. Cependant, un premier constat s'impose : peu importe la complexité du modèle mathématique utilisé, ces systèmes réussissent, ou plafonnent, selon le point de vue, tous à environ 95%.

Un deuxième constat dérive du premier : les taux de succès, affichés par les chercheurs qui ont mis au point ces systèmes, qui sont, à notre connaissance, peu analysés en détails. Les publications ne fournissent pas de portrait clair de l'ambiguïté résiduelle. On connaît peu ce en quoi consiste le 5% qui échappe encore à la désambiguïsation. Également, on ne sait pas, par exemple, si certains types d'ambiguïté demeurent problématiques et si d'autres ne causent plus ou presque plus de problèmes pour la désambiguïsation. Il se pourrait que l'on arrive à parfaitement désambiguïser les homographes *adjectifs/noms* alors que les techniques utilisées aient moins de succès pour la désambiguïsation des homographes *noms/verbes*.

Le deuxième constat en amène un troisième. Les étiqueteurs sont d'abord et avant tout réalisés pour étiqueter de grands corpus qui eux, serviront aux recherches qui auront pour but de connaître davantage la langue. Pourtant, les étiqueteurs eux-mêmes pourraient être utilisés pour connaître mieux le fonctionnement des langues.

Ces constats suscitent naturellement des questions. Pourquoi les étiqueteurs plafonnent-ils tous à environ 95%? Cette question fait écho au premier constat. En effet, il serait nécessaire de connaître les raisons qui empêchent les étiqueteurs actuels d'atteindre un taux de succès de 100% pour en déduire les pistes de recherche qui pourraient mener à des solutions pour atteindre un tel taux ou s'en approcher. Cette question en amène une seconde, en rapport avec le deuxième constat : Quelle est la nature de cette limite? En d'autres mots, il s'agit de regarder de près non seulement les cas de désambiguïsation réussis, mais aussi ceux qui ont échoué. Ces derniers peuvent nous en apprendre autant, sinon plus sur cette limite. Par exemple, nous ne savons pas jusqu'à quel point certaines constructions syntaxiques résistent plus que d'autres à la désambiguïsation. Nous ne savons pas non plus si et dans quelle mesure la sémantique est responsable de certaines ambiguïtés mal résolues. Il semble donc nécessaire d'en connaître plus sur ce qui ne fonctionne pas.

Enfin, le troisième constat, soulève la question suivante : Est-ce que des connaissances linguistiques permettent de repousser cette limite? Si tel était le cas, il faudrait savoir dans quelle mesure elles le permettent. Si, au contraire, elles ne le permettent pas, il faudrait savoir pourquoi. À ce sujet, les nombreux modèles mathématiques mis à l'essai suggèrent que la complexité du langage ne peut être saisie en totalité par les statistiques et les probabilités; autrement, un taux de succès de près de 100% serait déjà atteint. Par conséquent, l'application des connaissances que nous avons de la langue devrait permettre d'améliorer la performance des étiqueteurs, mais pourtant, aucune publication à propos d'étiqueteur ne détaille suffisamment l'utilisation de règles linguistiques pour en avoir une idée.

Ce mémoire a donc comme but d'enquêter sur les questions soulevées par les constats mentionnés précédemment. En particulier, il a les objectifs suivants :

a) Expliquer une part du succès des méthodes probabilistes

Nous posons comme hypothèse que les structures syntaxiques trouvées en corpus sont en grande partie récurrentes et que ces structures répondent à des modèles simples (ngrams) dont les probabilités d'occurrence peuvent être déduites automatiquement.

b) Identifier une part des limites des approches probabilistes

Nous posons l'hypothèse que certaines constructions syntaxiques sont difficiles, voire impossibles à saisir grâce à des modèles probabilistes basés sur des ngrams.

c) Établir dans quelle mesure la linguistique peut aider à améliorer les performances des étiqueteurs probabilistes.

Nous posons l'hypothèse que des connaissances linguistiques peuvent permettre d'en apprendre davantage sur le comportement des étiqueteurs probabilistes et que cela peut mener à de nouvelles pistes de recherches.

Enfin, au cours de nos observations sur les forces et les faiblesses des méthodes probabilistes, nous avons réfléchi sur la place des connaissances linguistiques dans l'étiquetage automatique de textes au sein des modèles mathématiques qui priment dans les systèmes proposés en ce moment.

ÉTAT DE LA QUESTION

Survol du Traitement Automatique du Langage au cours des cinq dernières décennies.

La recherche en Traitement Automatique du Langage a commencé nécessairement avec l'arrivée des ordinateurs. Comme dans tous les domaines de recherche, les travaux sont influencés et guidés, d'une part, par les courants théoriques, et d'autre part, par les besoins. Les prochains paragraphes présentent donc un bref survol de l'influence des uns et des autres sur la recherche en linguistique informatique, en particulier, sur l'étiquetage automatique.

Les années '50

La décennie des années '50 est marquée par le structuralisme et la traduction automatique. Les deux courants influenceront les recherches en linguistique informatique.

À cette époque, le structuralisme, sous toutes ses formes, était alors le courant dominant en linguistique. Différentes méthodologies ont donné naissance à plusieurs formes de structuralisme : le fonctionnalisme, la glossématique et le distributionnalisme. Ce dernier fut d'un grand secours aux États-Unis au milieu du XXe siècle. En effet, plus de 1000 langues amérindiennes étaient parlées sur le territoire. Le distributionnalisme a donc fournit une méthodologie à de nombreux linguistes et anthropologues pour décrire le système de ces langues. Le principe général du distributionnalisme est d'inférer le système d'une langue en observant la distribution de ses constituants. En effet, on a établi depuis longtemps que les constituants ne peuvent occuper n'importe quelle place les uns par rapport aux autres. En rassemblant ces constituants en classes distributionnelles et en décrivant leurs comportements respectifs, on obtient la structure interne de la langue.

Du côté de l'informatique, l'une des premières tâches que l'on confia à l'ordinateur fut la traduction automatique. La deuxième guerre mondiale puis la guerre froide opposant les États-Unis et l'ex-URSS ont servi de moteurs à cette recherche. L'espionnage qui sévissait alors nécessitait de nombreux traducteurs de part et d'autre des camps opposés. L'ordinateur apparaissait donc comme une machine pouvant imiter le travail des

traducteurs. On tenta alors de résoudre les problèmes de la traduction automatique avec les principes du distributionnalisme.

Les besoins de la traduction automatique étaient grands et multiples : d'abord analyser la langue de départ, ensuite extraire le message, puis l'exprimer dans l'autre langue. Pour y arriver, on utilisa donc les principes du distributionnalisme. Dans cette perspective, la traduction automatique apparaissait comme une tâche relativement simple : il s'agissait de faire l'inventaire des structures possibles des deux langues puis de les faire correspondre les unes aux autres. Klein & Simmons (1963 :336) expriment bien cette pensée : « For instance, some of the rules necessary for translating a sentence such as *The red book is on the table* into another language need be formulated only in terms of general classes, e.g. subject and predicate: noun phrase, verb phrase, verb modifying phrase; adjective, article, noun, verb, and preposition. »

Zelig Harris est cité comme le premier à s'être penché sur la reconnaissance automatique de la nature des unités syntaxiques et lexicales. La lecture de Harris (1964) illustre bien l'utilisation du distributionnalisme pour le traitement automatique du langage. Selon Harris, l'analyse des phrases ne peut être basée sur l'identification des suites de catégorie de mots car la plupart peuvent se suivre et se précéder les unes les autres. Par contre, remarque Harris, les catégories de mots ont un ordre plutôt contraint à l'intérieur des constituants syntaxiques (syntagmes). Voilà pourquoi reconnaître les constituants syntaxiques fut identifié comme la première étape pour pouvoir pousser plus loin l'analyse d'une phrase.

Dans son ouvrage, Harris pose qu'il existe un nombre fini de structures de phrases, ce qui, aujourd'hui, peut paraître étonnant : « The intention is that a few classes of strings, with simple rules describing how they occur in relation to each other, will suffice to characterize all sentences of the language (Harris 1964:10). »

Harris est tout de même conscient des limites de la méthode mais il insiste tout de même pour la justifier: « No claim is made here that any list of strings can be complete for a language, or that all properties of a sentence can be given by its string decomposition. However, a great amount of information about the sentences of a language can be obtained by decomposing them in respect to a reasonably adequate string list (Harris 1964:43) ».

Harris prétend ainsi arriver à établir le modèle de la phrase type de l'anglais et à faire l'inventaire de tous les patrons de phrase et de syntagmes de l'anglais. Cependant, il note au passage quelques problèmes reliés à la méthode. Par exemple, certains verbes n'apparaissent qu'avec des sujets animés ou encore, certains phénomènes comme la négation (not) modifient une séquence de mots et la morphologie (He walked / He did not walk.). Il existe donc des propriétés de constituants dont il faut tenir compte dans certaines exceptions.

L'ouvrage de Harris représente une masse énorme, voire gigantesque, d'observations. Cependant, le formalisme du distributionnalisme n'a pas le pouvoir explicatif que la grammaire générative de Chomsky apportera plus tard. Par ailleurs, avec le recul, le fait de considérer la langue comme un ensemble fini de propositions apparaît nettement irréaliste. C'est ce que le courant générativiste montrera en mettant de l'avant la créativité du locuteur.

Les années '60

C'est ainsi qu'au cours des années 1960, on assiste à l'essor de la grammaire générative, à la baisse d'intérêt pour la traduction automatique et à l'apparition du premier grand corpus.

En 1957, Chomsky publie *Syntactic Structures*, et jette les bases de la grammaire générative. En réaction contre le distributionnalisme, il introduit plusieurs aspects novateurs qui vont réorienter l'étude des langues. Son modèle transformationnel est plus simple tout en étant doté d'une puissance descriptive et explicative supérieure. D'abord, la prise en considération de la créativité du locuteur tranche nettement avec les approches antérieures. La notion de grammaticalité permet aussi de mieux cerner ce que doit décrire une grammaire et ce qu'elle ne doit pas décrire. Par ailleurs, les concepts de performance et de compétence amènent une réflexion nouvelle sur les objets à étudier et laissent place à une certaine introspection de la part du linguiste, ce qui avait été éliminé totalement par les courants structuralistes précédents.

En résumé, la théorie de Chomsky pose que l'étude du langage ne peut se baser sur l'étude exclusive de corpus car le sujet parlant possède des connaissances qui lui permettent de comprendre et d'énoncer des phrases qu'il n'a jamais entendues. Ce sont ces connaissances

qu'il faut découvrir pour être en mesure de comprendre le fonctionnement du langage. Chomsky émet l'hypothèse que, pour s'exprimer, un locuteur puise dans ses connaissances et élabore une structure profonde du message qu'il veut livrer. Des transformations s'appliquent alors successivement pour modifier cette structure profonde en une structure de surface, celle que l'on peut observer lorsqu'elle est énoncée.

Pendant que la grammaire générative gagne en popularité, les insuccès de la traduction automatique font que peu à peu les recherches dans cette voie cessent. Même le gouvernement américain, qui était alors le grand bailleur de fond de cette recherche, met fin à son financement dans ce domaine. Les résultats de la traduction automatique stagnent et on constate que le problème du traitement automatique du langage est plus complexe qu'il apparaissait au début et on s'attaque alors à des problèmes plus restreints.

Bien que la grammaire générative ait beaucoup d'impact dans les recherches fondamentales et en traitement automatique du langage, certains chercheurs continuent de travailler sur des corpus. C'est ainsi que l'équipe de chercheurs dirigée principalement par William Nelson Francis et Henry Kučera prépare un corpus d'anglais américain d'un million de mots à l'Université Brown. Le corpus Brown avait pour but de dresser un portrait synchronique de l'anglais contemporain des États-Unis. Le corpus a été élaboré de façon très minutieuse (c.f. Francis (1980)). C'est, entre autres, avec un tel corpus qu'est apparu le besoin d'outils linguistiques informatiques autres que dans une perspective de traduction automatique. Comme tous les autres corpus plus restreints réalisés auparavant, le corpus Brown devait être étiqueté pour être utile. Auparavant, l'étiquetage se faisait à la main ou par la consultation automatique de dictionnaires. En ce qui concerne cette dernière méthode, l'ambiguïté résiduelle était grande et il fallait revérifier manuellement le travail de l'ordinateur. Avec le corpus Brown, le besoin d'étiqueter automatiquement un corpus se fait sentir plus que jamais puisqu'il est impossible d'imaginer étiqueter manuellement un tel nombre de mots, d'abord parce que cela prendrait trop de temps, et ensuite, parce qu'il serait difficile de maintenir la cohérence des étiquettes tout au long du corpus, à plus forte raison si plusieurs personnes sont affectées à cette tâche.

Les années ‘70

La décennie 1970 est témoin de l’explosion des recherches génératives. Pendant que l’on s’affaire à étiqueter le corpus Brown, les recherches pour ou contre la grammaire générative vont bon train. Les travaux portent surtout sur la découverte de règles transformationnelles permettant d’expliquer des phénomènes de surface. Pour cette raison, de nombreuses recherches en informatique linguistique ont porté sur les parseurs. En fait, les efforts en linguistique étaient mis pour développer des théories basées sur des jugements de grammaticalité et les parseurs développés l’étaient pour expliquer ces jugements (Marshall 1983 :141). Les chercheurs travaillent alors surtout sur des exemples « inventés », c’est-à-dire, des exemples qu’ils génèrent eux-mêmes plutôt que des phénomènes observés d’une quelconque façon. Chomsky en profite pour remanier sa théorie et la présente alors sous le nom de Théorie Standard Étendue Révisée. L’équipe de Brown, quant à elle, prend plusieurs années pour étiqueter son corpus tandis qu’en Europe francophone, la Théorie du Lexique-grammaire de Maurice Gross s’apprête à prendre toute la place dans la recherche du traitement automatique du langage.

Les années ‘80

Dans les années 1980, les recherches se détachent peu à peu des considérations théoriques et se préoccupent de plus en plus des aspects pratiques. La linguistique informatique révise sa position et rejette explicitement le modèle compétence / performance de Chomsky et se détache graduellement des considérations des courants théoriques. C’est aussi l’utilisation des probabilités qui amènent un regain d’intérêt pour la recherche en automatisation linguistique. Le succès de l’utilisation des statistiques, en particulier pour l’étiquetage automatique, ouvre la voie à de nouvelles recherches pour perfectionner les techniques utilisées et les appliquer à d’autres problèmes tels que le parsage et l’identification automatique des syntagmes prépositionnels.

Plusieurs chercheurs s’aperçoivent que les outils informatiques doivent traiter du texte naturel ce qui n’a rien à voir avec le matériel linguistique artificiel qu’utilisent la plupart des génératives. En effet, même si des phrases telle que « Il regarde l’homme sur la colline avec un télescope » sont de plusieurs façons ambiguës, elles ne se rencontrent pas souvent dans les corpus. Le corpus Brown et le corpus LOB montrent que les tentatives

d'obtenir une grammaire parfaite, capable de générer toutes les phrases acceptables d'une langue, et seulement celles-ci, est utopique. La plupart des règles proposées font l'objet d'exceptions observées dans les corpus. Modifier ou créer de nouvelles règles pour ces exceptions ne ferait que compliquer la grammaire d'une part, et est, d'autre part, contradictoire avec l'universalité supposée des règles générativistes. À ce sujet, Jensen & Heidorn (1982:2) écrivent : « […] trying to write a grammar to describe explicitly all and only the sentences of a natural language is about as practical as trying to find the Holy Grail. »

Sampson (1987:20) va plus loin en affirmant que : « […] the idea of basing automatic language-processing on generative grammars of any category seems to me a dead end. »

Pendant que la grammaire générative s'essouffle, le succès de l'équipe de Brown et l'utilité de leur corpus incitent des chercheurs européens à constituer un corpus semblable d'anglais britannique. Pour ce deuxième corpus, nommé le corpus LOB, les chercheurs disposent de l'expérience acquise par l'équipe de Brown et de leurs données. En réutilisant le matériel, les chercheurs ont l'idée d'utiliser les statistiques tirées du corpus Brown pour évaluer la probabilité de rencontrer telle ou telle étiquette. Mises à part la reconnaissance et la synthèse de la parole, la linguistique informatique voit donc apparaître pour la première fois l'utilisation des probabilités dans l'étiquetage automatique de texte.

En peu de temps, les approches probabilistes deviennent très populaires.

Les années '90

Dans les années 1990, on assiste, d'un côté, à l'augmentation de la puissance de l'ordinateur ainsi qu'à l'arrivée d'Internet et, d'un autre côté, à l'utilisation sans précédent des approches probabilistes.

Au cours de la dernière décennie, les ordinateurs personnels n'ont pas cessé de devenir de plus en plus rapides. En moins de 10 ans, la vitesse des processeurs est passée de moins de 25Mhz à plus de 1Ghz. La capacité de stockage a crû de façon aussi importante. En effet, à la fin des années 1980, certains micro-ordinateurs ne possédaient pas de disque dur comme support-mémoire. Aujourd'hui, tous les ordinateurs en sont dotés et les capacités de ces

disques durs en terme de vitesse et de mémoire ne cessent de croître. Les limites de mémoire ou de vitesse non seulement imposent de moins en moins de contraintes, mais permettent aussi d'aller plus loin dans l'expérimentation de nouvelles approches.

De son côté, Internet a changé le monde des communications en favorisant les échanges de documents. Beaucoup d'information se retrouve sur les sites qui sont de plus en plus nombreux. Les internautes sont de plus en plus confrontés à des langues qu'ils n'ont jamais vues. Le besoin de traduction automatique refait surface. Les besoins linguistiques ne sont plus seulement d'ordre théorique. La recherche doit répondre aux besoins d'outils linguistiques du marché.

Au cours des années 1990, l'approche probabiliste a supplanté presque toutes les autres approches. Elle est non seulement utilisée pour étiqueter de grands corpus de textes, mais aussi pour parser et régler les problèmes que causent certains phénomènes linguistiques tel que l'attachement des syntagmes prépositionnels. Cette décennie est également marquée par les méthodes d'apprentissage automatique qui ont été appliquées à l'étiquetage automatique de textes. Entre autres, Eric Brill a présenté en 1995 une méthode d'extraction de règles grammaticales qui s'appliquent à l'étiquetage de texte.

Problématique de l'étiquetage

L'étiquetage d'un texte ou d'un corpus se décompose en trois étapes. D'abord, il faut identifier les unités lexicales. Le texte se présente comme une série de caractères parmi lesquels il faut identifier ceux qui correspondent à des mots. Ensuite, il s'agit d'attribuer à chacune de ces unités l'ensemble des étiquettes qui peuvent s'appliquer. Certaines unités ne sont, au départ, pas ambiguës. Elles recevront une seule étiquette. Par contre, d'autres recevront plusieurs étiquettes. La dernière étape est justement celle qui déterminera laquelle, parmi les étiquettes possibles, est la bonne. La désambiguïsation désigne cette dernière étape.

Au-delà de l'algorithme qui attribue des étiquettes aux mots d'un corpus, il faut aussi tenir compte d'autres éléments pour obtenir de bons résultats. En voici un aperçu.

Le corpus d'étiquettes

Le jeu d'étiquettes dont le programme de désambiguïsation dispose est important. En effet, il faut déterminer quelle est la portée de chacune des étiquettes, c'est-à-dire, quels seront les mots qui seront identifiés par une même étiquette. En posant que *1 étiquette = 1 catégorie grammaticale*, cette tâche semble aller de soi, mais ce n'est pas le cas. C'est surtout dans l'évaluation d'un étiqueteur que ce choix révèle son importance car un haut taux de succès peut dissimuler des étiquettes ambiguës alors qu'un taux de réussite plus faible peut, en réalité, être plus élevé que le premier. Ce fait est plus facilement mis en évidence par l'extrême. Il s'agit d'imaginer que chaque séquence de caractères située entre espaces blancs ou signes de ponctuation soit identifiée par une seule étiquette : « MOT ». Dans cette situation farfelue, le système n'attribue qu'une seule étiquette : MOT. Du point de vue du corpus d'étiquettes, le système a correctement attribué 100% des étiquettes, c'est-à-dire que le taux de réussite est parfait. Par contre, du point de vue de l'ambiguité, aucun problème n'a été réglé puisque l'étiquette « MOT » est elle-même ambiguë. En fait, elle n'identifie aucune classe grammaticale. De ce point de vue, le système n'a reconnu aucune catégorie grammaticale ce qui équivaut à un taux de succès de 0%. Dans la réalité, les situations sont moins caricaturales, mais il faut tout de même prendre au sérieux la désignation des étiquettes. Par exemple, le premier système d'étiquetage, Proto-Synthex (Simmons et al. 1962), regroupait sous une même étiquette, les verbes et les auxiliaires. Il étiquetait également tous les mots se terminant par *–ing* avec la même étiquette. Pourtant, on sait qu'en anglais plusieurs types de mots peuvent se terminer par *ing.*

Le nombre d'étiquettes

Le nombre d'étiquettes utilisées a des incidences directes sur le succès d'un étiqueteur. Certains n'utilisent qu'une dizaine d'étiquettes. D'autres en ont plus de trois cents. En anglais, il semble qu'un standard de trente-quatre étiquettes se soit installé, surtout pour faciliter la comparaison entre les étiqueteurs. Chose certaine, il est évident que plus le jeu d'étiquettes est important, plus il est difficile d'attribuer la bonne étiquette car le choix est vaste. Autrement dit, plus le nombre d'étiquettes est grand, plus il y a de risques d'erreurs.

Les types d'étiquettes

Le nombre d'étiquettes varie d'un système à l'autre. Par contre, la plupart utilisent des étiquettes de trois types : les étiquettes grammaticales, celles de ponctuation et les étiquettes spéciales. Les étiquettes grammaticales sont les plus nombreuses. Elles correspondent aux catégories grammaticales associées aux mots. Par exemple, le nom « chat » est associé à l'étiquette NN car c'est un substantif. De son côté, « rouge » sera identifié par l'étiquette JJ qui signifie « adjectif ». Les étiquettes grammaticales sont pour la plupart « décomposables », c'est-à-dire qu'elles sont construites d'une étiquette de base et d'autres étiquettes qui viennent s'y concaténer pour ajouter de l'information à la catégorie. Par exemple, dans la plupart des étiqueteurs pour l'anglais, un verbe se voit attribué l'étiquette VB et l'étiquette Z pour signifier respectivement qu'il s'agit d'un verbe et qu'il est à la troisième personne du singulier ce qui donne l'étiquette VBZ.

Les étiquettes de ponctuation sont les étiquettes attribuées aux signes de ponctuation. Ces signes sont considérés comme des « mots » non ambigus et sont très précieux pour la désambiguïsation. Par exemple, le point « . », la virgule « , » et le point-virgule « ; » se verront attribuer chacun une étiquette. Dans ces cas et dans quelques cas d'étiquettes grammaticales non ambiguës, une étiquette correspond à une catégorie grammaticale.

Enfin, les étiquettes spéciales correspondent, comme le nom l'indique, à des situations spéciales que les concepteurs des systèmes ont jugées bon d'identifier. En général, il s'agit d'étiquettes qui identifient des locutions pour éliminer une certaine ambiguïté qui demanderait un traitement supplémentaire. Dans ce cas, plusieurs unités graphiques discrètes se font attribuer une seule étiquette. Par exemple, en anglais, l'expression *such as* peut être identifiée par une étiquette particulière. Certains étiqueteurs utilisent aussi des étiquettes spéciales pour identifier les mots étrangers, les formules de tout genre (mathématiques, chimiques, statistiques, etc.), les lettres de l'alphabet, les frontières de phrases et de paragraphes, et d'autres unités jugées utiles par ces mêmes étiqueteurs.

L'étiquetage des formes complexes, contractées et discontinues

Théoriquement, l'étiqueteur devrait disposer d'un algorithme et d'un jeu d'étiquettes pour étiqueter les formes complexes telle que *pomme de terre*, les formes contractées telle que

du = de + le et les formes discontinue comme *ne… pas*. Cependant, la difficulté d'étiqueter ces formes ne réside pas seulement dans l'attribution d'une étiquette, mais également dans la définition même de ces formes. En particulier, les formes complexes. En effet, il n'existe pas de critères acceptés universellement pour déterminer si une suite de mots est un mot complexe ou une simple suite fortuite. Par ailleurs, il existe certaines suites de mots qui peuvent, dépendant du contexte, être un mot complexe ou non. Par exemple, *cordon bleu* dans *Le chef de ce restaurant est un cordon bleu* et dans *Le chef de ce restaurant attache un cordon bleu* est respectivement un mot complexe et une suite fortuite. Ainsi, puisqu'il est difficile de s'entendre sur la nature de ces formes, il est également difficile de les étiqueter. Peu d'étiqueteurs traitent ce problème mais certaines approches ont été proposées (Paroubek & Rajman (2000 : 134-135)).

Les mots peu fréquents et les hapax

Les mots peu fréquents sont une des difficultés majeures de l'approche probabiliste. Il est effectivement difficile d'estimer la probabilité d'occurrence d'un mot qui n'apparaît que quelques fois dans un corpus d'un million de mots.

Les mots peu fréquents causent un problème à deux niveaux : d'abord dans l'évaluation de la probabilité de chaque étiquette pour un mot hors contexte (probabilité à priori), ensuite dans l'évaluation de la probabilité des séquences (probabilité des ngrams). Supposons par exemple, que le mot « ferme » apparaisse 100 fois dans un corpus. Il apparaît 41 fois en tant que verbe, 30 fois en tant que substantif et 29 fois comme adjectif. Les étiquettes probables pour « ferme » sont donc VB (41%), NN (30%) et JJ (29%). Pourtant, il est également possible que « ferme » soit un adverbe comme dans « il travaille ferme ». Malheureusement, puisque que cette occurrence n'apparaît pas dans le corpus, la probabilité que « ferme » soit un adverbe sera nulle. Par conséquent, lorsque la probabilité de la séquence « travaille ferme » sera calculée, la probabilité de la séquence d'étiquettes VB+ADV sera évaluée à 0%. Il s'agit pourtant des bonnes étiquettes, mais l'étiqueteur sera dans l'impossibilité d'accorder les étiquettes correctes. Cette difficulté est courante. En effet, dans le corpus Brown, environ 40 000 formes lexicales ont une fréquence inférieure ou égale à cinq (Church (1988)).

Les hapax sont des mots qui n'apparaissent qu'une seule fois. Dans leur cas, le problème de la fréquence est encore plus évident. Que faire si un mot n'apparaît qu'une seule fois en tant que substantif dans le corpus si l'on sait qu'il peut être aussi un adjectif? La situation à éviter est d'attribuer une probabilité nulle à une étiquette potentielle. La solution proposée par Church (1988) consiste à augmenter de 1, toutes les occurrences des étiquettes et d'attribuer la valeur 1 à l'étiquette potentielle qui n'occurre pas dans le corpus. Dans notre exemple de tout à l'heure, il serait supposé que « ferme » apparaisse 42 fois comme verbe, 31 fois comme substantif, 30 fois comme adjectif et 1 fois comme adverbe. La probabilité que « ferme » soit un adverbe serait alors très faible, mais elle éviterait un résultat nul dans le calcul des ngrams.

Un problème similaire se produit aussi avec les noms propres et les mots commençant par une lettre majuscule. Church (1988) donne l'exemple de « Fall » qui peut être un surnom. Pourtant, dans le corpus Brown, cette possibilité ne se présente pas. Il faut pourtant la prévoir. Dans ce cas aussi, le nombre de toutes les occurrences est augmenté de 1 et Fall/NOM PROPRE se voit accorder une occurrence de 1.

Utiliser de plus gros corpus ne règlerait pas le problèmes des hapax et des mots peu fréquents car comme le prédit la loi de Zipf, même un énorme corpus comporterait de toute façon, des mots qui apparaissent peu souvent, à plus forte raison en français (Tzoukermann et al. (1997 :6)).

L'évaluation

Un problème qui peut sembler surprenant à prime abord est sans doute l'évaluation de l'efficacité d'un étiqueteur. En effet, il faut être conservateur. Par exemple, le taux de désambiguïsation des étiqueteurs actuels est d'environ 95% et parfois plus. Cela correspond certes à de bonnes performances, mais il faut revoir ces chiffres avec discernement. Une grande part de l'ambiguïté est due à un petit nombre de formes très fréquentes. En effet, environ 60% d'un texte n'est pas ambigu. Pour cette portion, la simple consultation d'un dictionnaire suffit. La difficulté de l'étiquetage ne concerne donc que 40% d'un texte. Par ailleurs, en incluant les mots qui ne sont pas ambigus, 90% d'un texte peut être correctement étiqueté en attribuant aux mots l'étiquette qui leur est la plus fréquente. Selon

ce point de vue, il ne reste que 10% du texte qui cause un problème. Par conséquent, malgré les nombreuses techniques utilisées, seulement la moitié de ce 10% est résolu lorsqu'un système affiche un taux de réussite de 95%. C'est dire le défi que constitue la résolution de l'ambiguïté. Pour confirmer ces propos, le projet Grace (c.f. Paroubek & Rajman (2000) a calculé qu'en moyenne, 89% des mots français pouvaient être correctement identifiés avec un peu plus d'une dizaine d'étiquettes en leur attribuant l'étiquette la plus fréquente ou l'étiquette NOM_COMMUN par défaut. Ces observations s'appliquent cependant aux catégories syntaxiques majeures. Dans le cas d'un jeu d'étiquettes plus étendu (plus de 300 étiquettes), le taux de réussite chute à 50% mais peut tout de même atteindre 60% avec seulement quatre règles de désambiguïsation (Paroubek & Rajman (2000 : 136)). Également, Tzoukermann et al.(1997 :7) rapportent qu'environ 57% des mots de leurs corpus français n'ont qu'une seule étiquette avec un jeu de 253 étiquettes.

Malgré ce qui vient d'être exposé, les étiqueteurs font plus que s'attarder à 10% d'un texte. En effet, même si l'attribution de la catégorie la plus fréquente réussit dans bien des cas, l'étiqueteur doit procéder à l'identification de l'étiquette pour chaque mot ambigu car il ne sait pas s'il lui suffit ou non d'attribuer l'étiquette la plus fréquente.

Il semble intuitif d'évaluer le taux de réussite d'un étiqueteur par le nombre de mots correctement étiquetés et le taux d'échec par le nombre de mots incorrectement étiquetés. Malheureusement, la notion de ce qui est « correct » laisse place à l'interprétation. Cependant, dans sa forme la plus simple, les mots correctement étiquetés sont ceux dont l'étiquette attribuée par l'étiqueteur correspond à l'étiquette du corpus de référence, généralement étiqueté à la main par un expert. L'inconvénient de ce type d'évaluation est que cela ne permet pas d'identifier la nature des erreurs. Par conséquent, des étiqueteurs affichant des taux de réussite semblables ne réussissent peut-être pas à désambiguïser le même matériel. Par exemple, un étiqueteur peut correctement identifier tous les noms communs sans exception et faire des erreurs dans la reconnaissance des verbes, alors qu'un autre étiqueteur peut se comporter dans le sens opposé : correctement identifier tous les verbes et se tromper dans le cas des noms communs. Dans ce cas, un taux de réussite semblable ne permet pas de mettre en évidence les avantages et les inconvénients de ces étiqueteurs.

On peut également utiliser d'autres unités que le mot pour évaluer les étiqueteurs. Selon Paroubek et Rajman (2000), les phrases, les paragraphes et les documents peuvent servir de référence. Dans le cas des phrases, un taux d'étiquetage de 96% correspond à un taux de réussite de 54,2% au niveau des phrases, c'est-à-dire qu'un peu moins de la moitié des phrases d'un texte ne sont pas totalement étiquetées correctement.

En définitive, il est important de se rappeler que le taux de réussite affiché d'un étiqueteur est fonction de plusieurs éléments : le jeu d'étiquettes, le corpus de référence et l'algorithme de segmentation. Par conséquent, les chiffres doivent être considérés sous plusieurs aspects.

Comparaison des étiqueteurs

Un autre aspect de l'étiquetage est la difficulté de comparer les étiqueteurs entre eux. S'ils n'utilisent pas un jeu d'étiquettes identique, et c'est souvent le cas, il faut construire des tables d'équivalence entre les étiquettes. Par exemple, dans un cas simple, l'étiquette NOM_MASCULIN d'un étiqueteur correspond à NOM_MASC d'un autre. Il s'agit alors de faire en sorte que NOM_MASCULIN = NOM_MASC. Ce cas est simple car il s'agit uniquement d'un problème de nomenclature. La catégorie est la même, mais l'étiquette qui l'identifie n'est pas identique. Cependant, on peut imaginer la difficulté de trouver des équivalences si les étiqueteurs ne segmentent pas les mots de la même façon (mots complexes, unités discontinues) ou s'ils n'ont pas les mêmes catégories, c'est-à-dire s'ils ne classent pas les mots de la même façon. Par exemple, si un étiqueteur ne fait pas la distinction de genre pour les noms et les adjectifs (c.f. Chanod & Tapanainen (1995)), bien des problèmes surviennent pour faire correspondre ses catégories avec celles d'un autre étiqueteur qui lui, distingue le genre. Comme on peut s'en douter, le problème s'amplifie si l'on veut comparer plus de deux étiqueteurs à la fois.

Les solutions mises de l'avant pour l'étiquetage automatique

Cette section présente, dans l'ordre chronologique, les systèmes qui ont été implémentés pour étiqueter automatiquement les mots d'un texte. Quand cela est possible, les améliorations par rapport aux techniques précédentes et les inconvénients de chacun des systèmes sont discutés. Cette section permet d'avoir une vue d'ensemble du chemin

parcouru depuis le premier système proposé et de mieux comprendre où en est rendu la recherche dans l'étiquetage automatique. Elle se termine sur les projets s'appliquant au français.

Proto-Synthex

Proto-Synthex (Simmons et al. 1962) est un système développé au début des années 1960. L'objectif général de ce système était de permettre l'interrogation d'une base de données en langage naturel. Il s'agissait précisément d'une encyclopédie stockée en mémoire et l'on désirait que l'utilisateur puisse l'interroger comme s'il avait affaire à un être humain. À une question telle que « Quelle est la capitale des États-Unis? », Proto-Synthex devait afficher l'article encyclopédique adéquat dans lequel se trouvait la réponse.

Un des modules de Proto-Synthex, le *computational grammar coder* (CGC), avait comme tâche d'identifier la nature grammaticale des mots de textes écrits en anglais. Avant Proto-Synthex, cette tâche était réalisée par la consultation de dictionnaires électroniques variant de 25 000 à 75 000 mots. C'est cette méthode que Harris (1964), dont il a été question plus tôt, utilisait. Proto-Synthex, par contre, fut le premier à proposer une alternative à la simple consultation de dictionnaires. Le système utilisait tout de même des dictionnaires restreints, mais il se servait aussi d'une batterie de tests pour identifier la nature des mots. L'algorithme procédait en six étapes successives.

Le CGC utilisait deux dictionnaires. Un premier contenant environ 400 entrées auxquelles n'était reliée qu'une seule étiquette. Il s'agissait de mots des classes fermées tels que les articles, les prépositions, les pronoms, les conjonctions, les verbes auxiliaires, les adverbes qui ne finissent pas par *–ly* et certaines formes des verbes *to have* et *to be*. Un second dictionnaire contenait environ 1500 noms, verbes et adjectifs qui étaient des exceptions aux tests des suffixes. Les mots étaient d'abord recherchés dans les dictionnaires. S'ils ne s'y trouvaient pas, alors le système procédait aux tests suivants :

Le test des majuscules. Ce test identifiait les mots qui commençaient par une lettre majuscule sans être le premier mot d'une phrase. Dans ce cas, les mots étaient étiquetés par l'ambiguïté nom/adjectif.

Le test des nombres. Les séquences d'un ou plusieurs chiffres étaient étiquetées ADJECTIF.

Le test des suffixes 1. Ce test était utilisé pour identifier les mots singuliers et pluriels. Il s'agissait en fait de suffixes au sens linguistique, c'est-à-dire de morphèmes. Ce test identifiait, entre autres, les suffixes *-s*, *-es* et *-ies* qui peuvent à la fois indiquer, en anglais, un nom au pluriel ou la troisième personne du singulier. Dans ces cas, le suffixe était modifié et la désambiguïsation était remise au deuxième test des suffixes. Par exemple, dans le cas de *nationalities*, qui peut, à cause de son suffixe, être identifié comme un nom ou un verbe, le suffixe *-ies* sera remplacé par *-y* et le résultat « nationality » sera envoyé au deuxième test des suffixes, dont il est question dans le paragraphe suivant, qui identifiera alors un nom. Le test des suffixes 1 constituait donc, en quelque sorte, une analyse morphologique qui permettait d'éliminer certaines ambiguïtés.

Le test des suffixes 2. Les suffixes utilisés dans ce test ne constituaient pas normalement des suffixes au sens linguistique. Il s'agissait en fait de suites de lettres, à la fin des mots, qui permettaient d'identifier leur nature. L'utilisation de tels suffixes avait pour but de diminuer l'ampleur des dictionnaires requis. Les suffixes avaient jusqu'à cinq lettres de long.

Le test des contextes. Il arrive parfois que, entre deux mots non ambigus, se trouve une suite de mots ambigus. Par exemple, entre un article et un verbe, se trouvent deux mots qui peuvent respectivement être ADJECTIF/VERBE et NOM/ADJECTIF :

Suite de mots			
Mot 1 (non ambigu)	Mot 2 (ambigu)	Mot 3 (ambigu)	Mot 4 (non ambigu)
ARTICLE	ADJECTIF/VERBE	NOM/ADJECTIF	VERBE

En combinant les possibilités, la suite située entre le *Mot 1* et le *Mot 4* peut donc être :

1. adjectif-nom
2. verbe-nom
3. adjectif-adjectif
4. verbe-adjectif

Dans ce cas, le système consultait une base de données qui contenait l'inventaire des suites de deux mots qui peuvent apparaître entre un article et un verbe. Dans cet exemple, l'ordinateur constatait que seule la suite ADJECTIF-NOM est licite et ne retenait donc que cette hypothèse, ce qui est correct. Cette base de données contenait environ 500 suites licites qui ont au plus trois mots. Ces règles linguistiques ont été observées et codées à la main.

L'ambiguïté résiduelle était ensuite traitée par les modules d'analyse syntaxique et sémantique supérieurs.

Les auteurs rapportent que le CGC réussissait à étiqueter correctement 90% des mots. De ce nombre, 45% des mots étaient étiquetés par la simple consultation des dictionnaires alors que l'autre 45% requérait l'application d'un ou plusieurs tests. Bien que ces chiffres soient impressionnants, il faut les remettre en contexte.

D'abord, les trente catégories syntaxiques reconnues par le système ne sont pas celles que l'on connaît aujourd'hui. Premièrement, parce qu'elles sont issues d'observations distributionnelles et ensuite, parce qu'elles sont parfois elles-mêmes ambiguës. Par exemple, les étiquettes /ED/ et /ING/ rassemblaient respectivement les participes passés et adjectifs et les participes présents et adjectifs. Également, l'étiquette V/AUX regroupait les verbes qui fonctionnent tantôt comme des auxiliaires, tantôt comme des verbes. Ces étiquettes, même correctement attribuées, sont intrinsèquement ambiguës.

Ensuite, le corpus utilisé pour les observations était le même que celui qui a servi à évaluer le prototype. De plus, l'évaluation a porté sur seulement quelques pages de l'encyclopédie en question.

Quoi qu'il en soit, comme le montre l'article de Klein & Simmons (1963), on a reconnu très tôt que le langage ne permet pas n'importe quelle combinaison. En utilisant les combinaison licites et en éliminant celles qui sont illicites, on peut déduire la catégorie des mots ambigus. On a aussi reconnu les avantages d'utiliser la finale des mots pour identifier la nature lexicale. Les avantages de cette technique étaient multiples. D'abord, elle réduisait la taille des dictionnaires. Elle améliorait du même coup la rapidité d'exécution de

l'algorithme qui n'avait pas à consulter de dictionnaires qui, à cette époque, étaient sur bandes magnétiques et donc plus longs à consulter. Grâce à cette architecture, Proto-Synthex n'avait besoin que d'environ 16Ko de mémoire et pouvait traiter plus de 20 mots à la seconde. Enfin, cela permettait aussi de reconnaître des mots jamais rencontrés auparavant.

TAGGIT

Le successeur de Proto-Synthex est le système TAGGIT. Ce système a été conçu dans le but d'étiqueter le premier grand corpus de textes, le corpus Brown. Les étiqueteurs sont liés de près au corpus car c'est d'abord dans le but d'accélérer l'étiquetage de grand corpus que les étiqueteurs ont été conçus. Le corpus Brown sera donc présenté avant de voir plus en détail de système TAGGIT.

Comme il en a été fait mention dans la section 0, les recherches sur grands corpus, même si elles n'ont pas toujours été à la mode, n'ont jamais été abandonnées. En 1963, un groupe de chercheurs se réunit à l'Université Brown pour discuter des paramètres d'un corpus ayant pour but de saisir les caractéristiques de l'anglais américain contemporain. Il est établi, lors de cette réunion, qu'un million de mots serait, d'une part, suffisant pour différentes recherches et, d'autre part, serait gérable par un ordinateur de l'époque.

Un corpus tel que celui de Brown, devait, pour être utile, être étiqueté. Dès le départ, il était évident qu'il ne pouvait être étiqueté manuellement pour plusieurs raisons. D'abord, d'un point de vue essentiellement pratique, le financement pour cette recherche était limité. Ensuite, plus théoriquement, une telle tâche représentait un haut risque d'erreurs humaines, soit parce que plusieurs personnes différentes travaillaient de façon divergente, à cause de l'ennui et de la répétition. Il fallait donc écrire un programme informatique pour réaliser cette tâche afin d'éliminer le plus possible ce type d'erreurs. On confia donc à deux étudiants gradués de l'Université Brown, Barbara B. Greene et Gerald M. Rubin, l'élaboration de ce programme informatique. Greene & Rubin (1971) se sont basés sur les travaux de l'équipe de Proto-Synthex pour élaborer la méthode de base. Nous verrons plus en détails cette méthode et nous verrons aussi en quoi elle diffère de celle de Proto-Synthex.

Après avoir fait un certain travail de pré-édition, TAGGIT fonctionne de façon générale comme le CGC de Proto-Synthex. L'attribution des étiquettes se fait par la consultation d'une liste de mots d'environ 2 800 formes. Cette liste, tout comme dans Proto-Synthex, contient les mots des classes fermées et les mots qui seraient incorrectement étiquetés par les suffixes. Si le mot se trouve dans cette liste, alors TAGGIT lui attribue la ou les étiquettes qui accompagnent le mot dans la liste. Dans cette liste, à peu près 61% des mots sont identifiés par une seule étiquette. Le reste des mots se voit attribué de deux à quatre étiquettes chacun, rarement cinq. Si le mot ne se trouve pas dans cette liste, alors sa finale est comparée à une liste d'environ 450 suffixes d'une à cinq lettres de long. Dans le cas où plusieurs suffixes s'appliquent (e.g. -UDE et -TUDE dans *hebetude*), ce sont les étiquettes qui accompagnent le plus long suffixe qui sont attribuées au mot. Seulement 51% des suffixes correspondent à une seule étiquette. L'économie d'espace-mémoire a été le principal facteur de décision pour l'établissement de la liste de suffixes. S'il était plus avantageux d'avoir un suffixe, alors il était ajouté à la liste de suffixes, sinon, on ajoutait le ou les mots à la liste de mots. En plus de ces deux étapes principales, TAGGIT utilise certaines routines spécialisées pour identifier certains cas :

1. Les mots commençant par une lettre majuscule, sans être en début de phrase, sont étiquetés NOM PROPRE;
2. les mots commençant par UN- et se terminant par -ED sont étiquetés ADJECTIFS;
3. les mots unis par un trait d'union sont étiquetés séparément. Ensuite, le tout est étiqueté en rapport avec les étiquettes de chacune des parties;
4. les mots contenant une apostrophe (e.g. *can't*) subissent un traitement semblable à celui en c);
5. si les mots qui se terminent par *-s* ne sont pas reconnus par les listes de mots et de suffixes, alors le suffixe *-s* est éliminé ou remplacé (dans certains cas de pluriel en *-ies*) et le système essaie de reconnaître une forme du singulier ou de la troisième personne du singulier;

Finalement, si un mot passe à travers ces vérifications sans se voir attribuer une étiquette, alors le programme donne par défaut les étiquettes NOM-VERBE-ADJECTIF.

Là où se distingue TAGGIT du CGC de Proto-Synthex, c'est dans la désambiguïsation. Lors d'une première passe, les mots du corpus sont parfois identifiés par plusieurs étiquettes. Par exemple, pour *walk*, TAGGIT aura identifié le mot comme pouvant être un

NOM et un VERBE. Pour réduire le nombre d'étiquettes multiples, et donc l'ambiguïté, TAGGIT repasse sur le corpus et tente de sélectionner la bonne étiquette parmi les choix proposés à l'aide d'environ 200 règles contextuelles. Ces règles sont toutes du type :

a) dans le contexte « a b ? c d », ? doit être X

ou bien

b) dans le contexte « a b ? c d », ? ne doit pas être X

où a, b, c, d, et X sont des étiquettes. Les règles contextuelles de désambiguïsation sont de deux types : celles qui sont sûres à 100% et celles qui sont sûres à 95%. Dans ce cas, un astérisque est inclus dans l'*output* pour signifier une possible erreur. Les règles sont appliquées des plus spécifiques, c'est-à-dire celles dont le contexte est le plus restreint, aux plus générales, c'est-à-dire celles dont le contexte est moins restreint.

Tout comme dans le cas du CGC de Proto-Synthex, une certaine ambiguïté est conservée due au choix des étiquettes, ou plutôt, de l'ensemble des mots identifiés par une étiquette. Par exemple, les auteurs concluent que, parce qu'il est impossible de désambiguïser par le contexte le mot « does » (qui peut être auxiliaire ou verbe conjugué), ils lui donnent une seule étiquette, peu importe le cas : DOZ. Puisque les auteurs en sont conscients, ces étiquettes sont identifiées, dans leur ouvrage, par un astérisque.

Caradec & Saada

En faisant un saut dans le temps, on trouve qu'en 1982, Caradec & Saada (1982) proposent, pour le français, une approche somme toute semblable à celle de Greene & Rubin (1971). La méthode proposée utilise les terminaisons graphiques pour distinguer les adjectifs et les substantifs des verbes. Autrement dit, les adjectifs et les substantifs sont regroupés dans une même catégorie et cette catégorie s'oppose à celle des verbes. Tout comme pour les systèmes présentés plus tôt, l'algorithme consulte en premier lieu une liste de mots et, si le mot ne s'y trouve pas, consulte une liste de terminaisons pour identifier la nature grammaticale de ce mot. Ces terminaisons ont, au plus, cinq lettres (Caradec & Saada

1982 : 274). Cette limite correspond à celle qu'avait également imposée Greene & Rubin à TAGGIT pour des raisons d'économie d'espace-disque. Caradec & Saada ne vont pas plus loin dans la résolution de l'ambiguïté. À l'aide de 1772 terminaisons et d'un dictionnaire de 3 669 formes lexicales, ils réussissent à reconnaître correctement la nature grammaticale de 86% des mots d'un corpus de 152 120 mots. Ces chiffres sont toutefois trompeurs. En effet, les adjectifs et les substantifs ne constituent qu'une seule catégorie. De plus, les participes présents et les participes passés sont classés comme des adjectifs. C'est donc dire que l'étiquette NOM/ADJECTIF regroupe à la fois les substantifs, les adjectifs, les participes passés et les participes présents. À ce compte, l'ambiguïté n'est pas tellement réduite. En fait, la méthode proposée par Caradec & Saada ne permet que d'étiqueter les mots des classes fermées et les adverbes, d'une part, et d'autre part, de distinguer les verbes des autres catégories regroupées sous l'étiquette NOM/ADJECTIF. Les auteurs se défendent toutefois en disant que : « […] certaines applications informatiques nécessitant une analyse syntaxique, comme la documentation automatique par exemple, n'ont pas toujours besoin d'établir un distinguo subtil entre les substantifs et les adjectifs. C'est pourquoi nous n'avons pas, dans cette étude, tenté de les séparer en deux classes ».

S'il est probablement vrai que certaines applications ne requiert pas cette distinction, il existe de nombreux autres cas où une telle distinction est essentielle.

En résumé, le travail de Caradec & Saada est très sommaire et ne comporte comme seule originalité, que l'utilisation d'une règle spéciale pour la reconnaissance des adverbes. Dans ce cas, les auteurs font remarquer, à juste titre, que 80% des adverbes de leur corpus se terminent par *-ément*, *-ement* et *–mment*. Utilisées seules, ces terminaisons ne sont toutefois pas utiles car elles apparaissent aussi chez d'autres catégories grammaticales. Par contre, grâce à la règle :

adverbe = adjectif (souvent au féminin) + suffixe en *-ment*

ils arrivent à déterminer s'il s'agit d'un adverbe, d'un adjectif ou d'un substantif.

Les auteurs remarquent que les terminaisons graphiques fournissent des informations importantes qui sont, en français, le temps, la personne et le nombre pour les verbes; le nombre et parfois le genre pour les substantifs; le genre et le nombre pour les adjectifs.

Comme concluent les auteurs, leurs travaux permettent toutefois de constater que, malgré la réputation du français d'avoir de nombreuses exceptions à de toutes aussi nombreuses règles, la graphie semble respecter, dans une certaine mesure, la morphophonologie de la langue.

CLAWS

Le succès remporté par le corpus Brown et son étiqueteur, TAGGIT, ont poussé les Européens à refaire l'expérience sur un corpus d'anglais britannique. Au cours des années 1970, des chercheurs des Universités de Lancaster, d'Oslo et de Bergen ont regroupé 500 échantillons de texte de 2000 mots chacun pour constituer un corpus jumeau de celui de Brown, en anglais britannique cette fois. Le corpus LOB, un acronyme fait d'après le nom des villes logeant les universités participantes (Lancaster, Oslo et Bergen), a permis d'aller plus loin dans l'étiquetage. En effet, les chercheurs avaient comme but, non seulement de représenter dans un corpus l'anglais britannique contemporain, mais aussi d'augmenter la proportion de mots pouvant être correctement étiquetés automatiquement. Forts de l'expérience de TAGGIT et se basant sur les résultats obtenus sur le corpus Brown, ces chercheurs feront une découverte qui ouvrira la voix aux recherches actuelles.

La description de CLAWS est importante. D'abord parce que ce système a relancé l'intérêt des recherches sur l'étiquetage automatique, ensuite, à cause de son taux de succès impressionnant comparé aux étiqueteurs réalisés auparavant et finalement, parce qu'il permet d'illustrer le fonctionnement général des étiqueteurs probabilistes.

CLAWS, au début, reprend essentiellement la technique utilisée par TAGGIT, c'est-à-dire : utiliser d'abord une liste de mots et une liste de suffixes pour identifier les étiquettes potentielles des mots et ensuite, utiliser des règles de désambiguïsation. Johansson & Jahr (1982) reprennent les travaux de Greene & Rubin (1971) et tentent d'améliorer la liste des suffixes et des mots utilisés par l'algorithme de TAGGIT.

Johansson & Jahr bénéficiaient de la liste de suffixes utilisés par TAGGIT en plus du corpus Brown déjà étiqueté, ce que ne possédaient pas, évidemment, Greene & Rubin en 1971. Pour vérifier les suffixes, Johansson & Jahr ont d'abord utilisé une liste d'environ 75 000 mots écrits en ordre inverse (de la fin vers le début. Par exemple : *elppa* au lieu de *apple*) obtenus du corpus Brown et du corpus LOB et ensuite, ils l'ont comparée à une liste des fréquences d'apparition des étiquettes pour chaque suffixes utilisés par TAGGIT.

Par conséquent, ils ont pu repérer des erreurs, ajouter des suffixes et améliorer l'efficacité de certains suffixes. En tout, ils ont éliminé 80 suffixes jugés douteux ou superflus et ont ajouté 150 nouveaux suffixes à la liste de 450 utilisée par Greene & Rubin. Ils ont aussi éliminé près de 50 étiquettes reliées à des suffixes. En fait, si l'on tient compte des suffixes éliminés, il s'agit plutôt d'une addition de 240 suffixes.

Les auteurs, en réorganisant la liste de suffixes, ont ainsi fait passer la liste de mots de 3 000 à 5 000 entrées en y ajoutant des exceptions ce qui, d'un autre côté, éliminait de nombreuses ambiguïtés. Ils ont aussi ajouté à cette liste, environ 300 abréviations courantes et environ 500 mots communs commençant par une lettre majuscule ainsi que quelques unités syntaxiques (Garside 1987:36).

Certaines routines spéciales qu'utilisait TAGGIT ont aussi été modifiées pour CLAWS (Garside (1987:31)). En effet, pour l'étiquetage du corpus Brown, une unité syntaxique telle que *can't* se voyait assigner deux étiquettes représentant « le mode » + *not*. Pour le corpus LOB, CLAWS divise *can't* en deux unités : *can* et *n't* et les deux unités sont étiquetées séparément. Une marque indique toutefois que les deux unités sont liées dans le corpus. Également, dans certains cas, CLAWS identifie plusieurs mots par une seule étiquette. C'est le cas des unités *because of* et *such as*, toutes deux étiquetées PRÉPOSITION, ainsi que *as if* et *at once*, respectivement étiquetées CONJONCTION DE SUBORDINATION et ADVERBE. Un traitement spécial était aussi réservé à certaines unités graphiques. Par exemple, des unités telles *$37.00, £2, i, x", 27th, 1st, 1940s, ½, H_2SO_4*, etc. font l'objet d'un traitement spécial. Par ailleurs, les routines spéciales pour traiter le pluriel, le possessif (en anglais : *'s*) et les mots avec trait d'union sont maintenues.

Marshall (1983) fera une observation qui bouleversera l'avenir de l'étiquetage. En effet, en appliquant les règles de désambiguïsation de TAGGIT sur le corpus LOB, il s'aperçoit que, bien que seulement 25% des règles soient du type :

X Y → (not) A ou Y *X* → (not) A

c'est-à-dire des règles qui ne se préoccupent que du mot précédant ou suivant, ce type de règle est appliqué dans 80% des cas. Autrement dit, l'information la plus fréquemment utile pour la désambiguïsation est l'étiquette précédente. Cette observation suggéra à Marshall que l'utilisation des probabilités de co-occurrence entre des étiquettes successives serait une méthode efficace et peut-être plus adéquate pour étiqueter le corpus LOB. Les résultats auxquels il arriva sont étonnants. L'exemple qui suit aidera à bien saisir l'amélioration que les probabilités ont apportée aux règles linguistiques utilisées auparavant par TAGGIT.

Prenons deux règles de TAGGIT :

La première exprime le fait que devant un verbe conjugué à la troisième personne (VBZ), un mot ne peut pas être un nom au pluriel (NNS) :

X VBZ → not NNS

Inversement, la seconde spécifie qu'un nom au pluriel ne peut pas être suivi d'une forme verbale conjuguée à la troisième personne du singulier :

NNS *X* → not VBZ

Soit la phrase :

	Henry	NP
Henry likes stews.	likes	NNS VBZ
(Henri aime les ragoûts.)	stews	NNS VBZ
	.POINT	

Dans cette phrase, *likes* et *stews* sont ambigus : ils peuvent tous deux être une forme de nom pluriel ou une forme conjuguée de la troisième personne du singulier. Dans un tel cas, TAGGIT tente de désambiguïser les suites de mots ambigus en partant de chaque extrémité et en essayant graduellement de lever les ambiguïtés. Cependant, les règles mentionnées plus tôt ne sont d'aucun secours car l'une et l'autre sont bloquées par l'ambiguïté des mots *likes* et *stews*. En effet, pour qu'elles puissent s'appliquer, au moins un des mots étiquetés NNS et VBZ doit être non ambigu. Puisque ce n'est pas le cas, aucune règle ne s'applique. L'ambiguïté ne sera donc pas résolue par TAGGIT. Pourtant, il est beaucoup plus probable que la forme *likes* soit un verbe à la troisième personne du singulier et que la forme *stews* soit un nom.

CLAWS, de son côté, évaluera la probabilité de chacune des séquences d'étiquettes possibles :

Prob(NP NNS VBZ),

Prob(NP NNS NNS),

Prob(NP VBZ NNS) et

Prob(NP VBZ VBZ).

La probabilité de la suite NP VBZ NNS étant la plus élevée, la forme *likes* est donc identifiée comme une forme verbale et *stews*, comme un nom pluriel (c.f. Marshall (1983, 1987) pour plus de détails sur les calculs).

Le calcul de probabilité utilisé au départ était la probabilité de rencontrer l'étiquette A suivant l'étiquette B divisé par la probabilité de rencontrer l'étiquette A :

$$\text{Probabilité (AB)} = \frac{\text{fréquence (AB)}}{\text{fréquence (A)}}$$

Cependant, ce calcul avait l'inconvénient de privilégier les étiquettes très fréquentes au détriment des étiquettes plus rares. Dans certains cas, l'algorithme choisissait une mauvaise étiquette à cause de sa fréquence élevée plutôt que la bonne qui était moins fréquente. Le calcul a donc été modifié pour tenir compte de la fréquence de l'étiquette B :

$$\text{Probabilité (AB)} = \frac{\text{fréquence (AB)}}{\text{fréquence (A)} \times \text{fréquence (B)}}$$

En plus de calculer la valeur de la séquence la plus probable, CLAWS procède, dans certains cas, à un deuxième calcul. En effet, si, dans la plupart des cas, les étiquettes de la séquence la plus probable correspondent aux étiquettes correctes, il arrive que plusieurs séquences aient sensiblement la même valeur ce qui rend l'attribution des étiquettes très délicate. En effet, il arrive que l'étiquette attribuée par la séquence la plus probable ne soit pas la bonne. Dans ce cas, l'algorithme calcule pour chaque mot la probabilité de chacune des étiquettes que lui attribue l'ensemble des séquences probables. Autrement dit, l'algorithme détermine pour chaque mot, la probabilité de chacune de ses étiquettes, indépendamment des étiquettes qui les précèdent ou les suivent. Par exemple, dans le cas de *Henry likes stews*, quatre séquences d'étiquettes sont possibles. Ces séquences attribuent deux étiquettes à *likes*, soient NP et VBZ. CLAWS évalue la probabilité que l'étiquette soit VBZ en calculant la somme de la valeur des séquences qui prédisent cette étiquette, puis en divisant cette somme par la somme de la valeur de toutes les séquences :

$$\frac{\text{val (NP NNS VBZ .)+val(NP VBZ VBZ .)}}{\text{val (NP NNS VBZ .)+val(NP VBZ VBZ .)+val(NP NNS NNS .)+val(NP VBZ NNS .)}}$$

L'algorithme évalue ensuite de la même façon la probabilité que l'étiquette soit NNS. En comparant les résultats, CLAWS s'aperçoit que la probabilité que l'étiquette soit VBZ est plus élevée. Il sélectionne donc cette étiquette pour *likes*. Les étiquettes ayant un seuil de

probabilité supérieur à 90% sont sélectionnées comme étant la bonne étiquette. Les autres sont éliminées. Dans le cas où plusieurs étiquettes ont un seuil plus élevé que 90%, CLAWS ne prend aucune décision et les étiquettes sont listées en ordre décroissant de probabilité. La désambiguïsation est alors assurée par un réviseur humain à la fin du traitement.

Le calcul des probabilités est efficace, mais, dans certains cas, il arrive qu'il ne soit pas adéquat. En effet, il arrive que le mot qui précède soit d'une utilité presque nulle dans l'attribution d'une étiquette et que ce soit l'avant-précédant qui soit plus important. C'est pourquoi l'algorithme de CLAWS utilise certains triplets de probabilité. Trois cas sont recensés :

1. Il arrive que, dans la désambiguïsation de mots qui peuvent être des verbes conjugués au passé ou des participes passés (les formes en —ED), ces mots soient précédés d'un adverbe. Dans ce cas, le mot qui précède l'adverbe est plus utile que l'adverbe lui-même, en particulier, lorsque le mot qui précède l'adverbe est une forme du verbe *have* ou *be*. Il y a alors peu de chance que le mot qui suit l'adverbe soit autre chose qu'un verbe au passé. Dans ce contexte, une procédure a donc été implémentée pour favoriser l'étiquette « verbe au passé ».

2. Par ailleurs, la même observation s'applique à la particule *not* qui s'avère beaucoup moins utile que le mot qui la précède. Étant donné que cette particule est toujours étiquetée de façon non ambiguë XNOT, une fonction a tout simplement été créée pour ignorer cette étiquette et plutôt consulter celle qui la précède. Les citations entre guillemets et les parenthèses sont ignorées de la même façon.

3. Enfin, on observe également que, dans la majorité des cas, une conjonction de coordination (*and* et *or*) unit des mots de même catégorie. Lorsque de part et d'autre de la conjonction les mots sont ambigus, alors plusieurs séquences se voient attribuer une même probabilité d'occurrence. Par exemple, si les mots ceinturant la conjonction peuvent être des noms ou des adjectifs, alors quatre suites sont possibles : NN-CC-NN, JJ-CC-JJ, NN-CC-JJ et JJ-CC-NN. Pourtant, les deux dernières sont

beaucoup moins probables. Une procédure a donc été programmée pour gérer ces cas.

Avec le calcul des probabilités et les procédures spécifiques à certaines situations, CLAWS réussit à correctement étiqueter 97% des mots en utilisant 134 étiquettes.

Mis à part les quelques triplets qui viennent d'être présentés, la technologie de l'époque ne permettait pas d'utiliser les trigrams (Marshall 1983 :147). Cependant, de l'avis de Marshall, il serait étonnant que l'utilisation d'une matrice triple donne de meilleurs résultats que l'utilisation d'une matrice double. Une intuition qui sera confirmée, entre autres, par Schmid (1994), Cutting et al. (1992) et Jelinek (1985), qui, même en utilisant des modèles mathématiques plus sophistiqués (une approche par *neural networks* dans le cas de Schmid) n'ont pas obtenu de résultats significativement meilleurs.

En résumé, les travaux sur le corpus LOB ont permis d'augmenter le taux de succès pour l'étiquetage mais aussi d'ouvrir la voie à de nombreuses approches statistiques mettant à profit des variantes du Hidden Markov Model (Cutting et al. (1992), Rabiner (1986)).

Apprentissage automatique de grammaires

Pendant que les règles linguistiques et les statistiques étaient populaires dans les années 1990 (c.f. Chanod & Tapanainen (1995), Samuelsson & Voutilainen(1997), Qiao & Huang(1998)), Eric Brill a proposé une méthode par apprentissage automatique qui tranche avec les méthodes conventionnelles que nous avons vues jusqu'à présent.

Alors que les étiqueteurs « traditionnels » que nous avons vus jusqu'à maintenant sont conçus dans le seul et unique but d'attribuer des étiquettes aux mots d'un texte, le système de Brill est conçu pour extraire des relations grammaticales entre les mots d'un corpus déjà étiqueté. Une fois ces relations extraites, elles peuvent être utilisées pour déterminer la nature grammaticale des mots d'un texte non étiqueté.

Il s'agit d'avoir un même corpus en deux formats : un format étiqueté manuellement, c'est-à-dire vérifié par un être humain, et un format non étiqueté. L'algorithme d'apprentissage fonctionne ainsi. La première étape est d'attribuer des étiquettes au format non étiqueté.

Plusieurs façons sont possibles allant de l'attribution aléatoire d'une étiquette à l'attribution des étiquettes par un autre étiqueteur. Ce corpus sera appelé *corpus étiqueté « sommairement »* pour le distinguer du corpus étiqueté à la main.

Par la suite, le corpus étiqueté sommairement est comparé au corpus étiqueté manuellement. Une série d'hypothèses sont émises. Par exemple, « Changer l'étiquette DÉTERMINANT par celle de PRONOM si l'étiquette suivante est VERBE ». Ainsi, la séquence suivante :

Jean <NOM PROPRE> le <DÉTERMINANT> pense <VERBE>.

Sera remplacée par :

Jean <NOM PROPRE> le <PRONOM> pense <VERBE>.

Chaque hypothèse est appliquée au corpus étiqueté sommairement et le résultat est comparé au corpus étiqueté manuellement. Les erreurs produites sont compilées et l'hypothèse de laquelle résulte le moins d'erreurs est retenue comme une règle. Ensuite, toutes les autres hypothèses sont appliquées et celle par laquelle résulte le moins d'erreurs est retenue comme une deuxième règle et ainsi de suite, jusqu'à ce qu'aucune règle ne puisse faire diminuer le nombre d'erreurs contenues dans le corpus étiqueté sommairement. Le schéma suivant illustre l'ensemble de la procédure dans le cas où seulement quatre hypothèses (H1, H2, H3, H4) sont possibles. Dans le schéma, le corpus de départ correspond au corpus étiqueté sommairement.

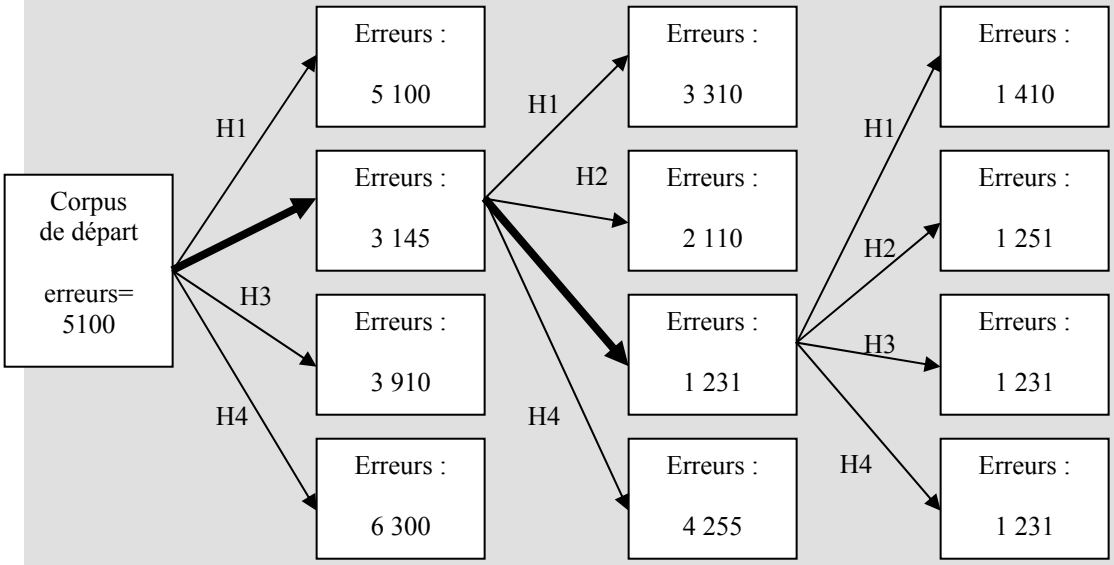


Figure 4-1 Processus d'apprentissage de la méthode de Brill.

Pour que le système d'apprentissage puisse émettre des hypothèses valables, les hypothèses ont une forme définie à laquelle le système n'a qu'à instancier des variables. Brill (1995) en utilise six :

Change l'étiquette A pour B si

1. Le mot précédent/suivant est étiqueté *z*.
2. Le deuxième mot avant/après est étiqueté *z*.
3. Un des deux mots précédents/suivants est étiqueté *z*.
4. Un des trois mots précédents/suivants est étiqueté *z*.
5. Le mot précédant est étiqueté *z* et le suivant est étiqueté *w*.
6. Le mot précédant est étiqueté *z* et le deuxième mot avant/après est étiqueté *w*.

Le système instanciera les variables avec toutes les valeurs possibles et testera chacune des hypothèses ainsi obtenues. Les règles retenues peuvent par la suite être utilisées pour étiqueter n'importe quel autre texte. Le taux de succès rapporté par Brill oscille autour de 96%.

Bien que le temps d'entraînement du système proposé par Brill soit long, cette méthode possède bien des avantages. D'abord, les phénomènes linguistiques sont représentés par des

règles faciles à comprendre, contrairement aux tables de probabilités des méthodes statistiques. Dans ces dernières, les phénomènes linguistiques sont difficiles à saisir et modifier les calculs pour ajuster les probabilités à certains phénomènes est plutôt difficile. De plus, les règles ainsi obtenues sont également moins volumineuses que les tables de statistiques. Par exemple, Brill (1995) rapporte que les 82 premières règles extraites atteignent un taux de réussite de 96.8%.

Utiliser les règles extraites par la méthode d'apprentissage automatique permet de reporter l'attribution d'une étiquette au moment où l'état du système sera le plus en mesure de prendre une décision. En effet, il est probable qu'il soit impossible d'étiqueter le mot *x* à un moment *z* parce que le mot qui précède *x* n'est pas encore étiqueté. Quand ce mot sera étiqueté, *x* pourra alors l'être si une règle s'applique. Sinon, les autres règles s'appliqueront jusqu'à ce que l'environnement de *x* soit favorable à son étiquetage. Avec les méthodes statistiques, l'étiquetage est plus « direct », c'est-à-dire que l'étiqueteur probabiliste ne change plus sa décision quand elle est prise. Lorsqu'il calcule la probabilité de l'ensemble des étiquettes d'une phrase, il ne peut reporter sa décision si un mot présente une probabilité faible. Le calcul est effectué et le meilleur résultat l'emporte.

Un autre avantage des règles « à la Brill » est la possibilité pour certaines règles de venir corriger le travail fait par des règles utilisées antérieurement par le système. Par exemple, si dans « Il la ferme », « ferme » a d'abord été étiqueté comme SUBSTANTIF parce qu'il était précédé de l'étiquette DÉTERMINANT, ultérieurement, le tout pourra être changé de nouveau si une autre règle vient s'appliquer. En effet, si l'étiquette DÉTERMINANT est remplacée par celle de PRONOM à cause de la présence de « Il », la règle qui change l'étiquette SUBSTANTIF par VERBE si précédé de PRONOM sera appliquée. Enfin, pour terminer l'inventaire des avantages majeurs pour l'étiquetage, les hypothèses peuvent tenir compte des mots environnants et non pas seulement des étiquettes environnantes. En effet, les méthodes statistiques calculent l'ensemble des probabilités de transitions pour passer d'une étiquette à une autre. Par contre, elles ne calculent pas la probabilité de passer d'un mot à un autre. Ces calculs seraient trop importants pour être manipulés par la machine et la table de probabilité prendrait des proportions démesurées. En revanche, les règles de Brill peuvent être adaptées pour tenir compte des mots qui, dans certains cas, peuvent aider à

désambiguïser là où les étiquettes sont insuffisantes. Les règles prennent alors l'allure suivante :

Change l'étiquette A pour B si :

1. Le mot précédant/suivant est *w*.
2. Le deuxième mot avant/après est *w*.
3. Un des deux mots avant/après est *w*.
4. Le mot courant est *w* et le précédant/suivant est *x*.
5. Le mot courant est *w* et le précédant/suivant est étiqueté *z*.
6. Le mot courant est *w*.
7. Le mot précédant/suivant est *w* et le précédant/suivant est étiqueté *z*.
8. Le mot courant est *w*, le suivant/précédant est *x* et le précédant/suivant est *z*.

L'étiquetage du français

Comme le laisse deviner les travaux mentionnés plus tôt, la majorité des recherches sur l'étiquetage automatique de textes se sont faites sur l'anglais. En effet, comme le dit Véronis (2000), en faisant référence au milieu de la décennie 1990 : « […] au moment où le British National Corpus propose 100 millions de mots étiquetés du point de vue grammatical et 10 millions de mots de parole transcrite […], la communauté francophone ne dispose même pas de l'équivalent du Brown Corpus (1 million de mots étiquetés), réalisé dans les années soixante ».

Ce fait implique qu'il n'y a pas si longtemps, aucun étiqueteur du français ne pouvait utiliser de statistiques car les ressources pour estimer ces dernières n'existaient pas. De fait, les premiers travaux d'envergure portant sur le français ont dû utiliser des méthodes novatrices pour étiqueter les premiers corpus car il leur fallait faire un premier pas. On se rappelle que ce premier pas s'est fait en anglais, par l'entremise de TAGGIT. (Voir aussi Habert et al 1997 qui dit également que rien de tel n'est disponible pour le français)

Les recherches sur l'étiquetage automatique du français ont débuté sérieusement au milieu des années 1990.

Problèmes relatifs à l'étiquetage du français

Forts de l'expérience de la communauté anglophone, les chercheurs qui se sont penchés sur l'étiquetage du français connaissaient déjà les différentes techniques utilisées et testées sur l'anglais. L'approche à la TAGGIT ne causait pas de problèmes puisque que tout linguiste peut éditer manuellement des règles de désambiguïsation. Par contre, ce processus est long et ne donne pas les meilleurs résultats s'il est utilisé comme seule technique de désambiguïsation. Les techniques probabilistes sont plus rapides à implémenter et donnent de meilleurs résultats dès la première implémentation. Elle ont également fait leur preuve sur l'anglais. Elles sont perfectibles, mais elles sont un premier pas plus facile et rapide à franchir. Par contre, elles doivent, pour être efficaces, compter sur de grands corpus, étiquetés et vérifiés par un expert humain. Un tel corpus, n'existait pas encore à l'époque où Tzoukermann et ses collègues conçoivent un système d'étiquetage du français. C'est pourquoi Tzoukermann et al. (1997) ont proposé l'utilisation des génotypes au lieu des probabilités lexicales que l'on connaît en anglais.

Les génotypes correspondent à l'ensemble des groupes d'étiquettes que peuvent recevoir les mots français. Par exemple, le mot « le » peut être un pronom et un article. Son génotype est donc [PRONOM, ARTICLE]. Puisque le mot « la » peut aussi être un pronom et un article, il correspond au même génotype. Cela permet à un ensemble de mots, ayant en commun un même génotype, de partager les mêmes probabilités. Autrement dit, les probabilités d'occurrence des étiquettes pour les mots d'un génotype s'additionnent et sont utilisées pour calculer les probabilités du génotype. Ainsi, si un mot faisant partie d'un génotype n'apparaît pas dans le corpus, il se voit attribuer les mêmes probabilités que les autres membres du génotype. Par exemple, dans l'expérience menée par l'équipe de Tzoukermann, l'ensemble des données statistiques des mots du génotype [NOM FÉMININ SINGULIER, VERBE 3ÈME PERSONNE SINGULIER], dont font partie, par exemple, les mots *laisse*, *masse*, *tâche*, *lutte*, *forme*, *zone* et *place*, a permis d'estimer la probabilité de chacune des étiquettes, NOM FÉMININ SINGULIER et VERBE 3ÈME PERSONNE SINGULIER à respectivement 89.15% et 10.85%. Par conséquent, le mot *danse*, absent de leur corpus, mais faisant partie de ce génotype, s'est fait attribuer les mêmes probabilités. L'équipe de Tzoukermann a tout de même dû utiliser des corpus, mais grâce aux génotypes, ces derniers pouvaient être de

taille beaucoup plus petite. En fait, ils ont utilisé trois corpus totalisant environ 60 000 mots.

Par ailleurs, il s'avère que selon les travaux de Tzoukermann et al. (1997), des génotypes identiques ont une même distribution. Conséquemment, il est possible d'utiliser les probabilités des génotypes pour calculer la probabilité des séquences d'étiquettes. C'est ce qu'ils ont fait, en modifiant le calcul de probabilité pour tenir compte des génotypes. Le calcul est en fait le même qu'utilisa CLAWS.

Nous avons vu que la probabilité de chaque génotype est calculée par la somme des occurrences de tous les mots qui font partie du même génotype. Par conséquent, $P(t|T)$ exprime la probabilité de l'étiquette t pour le génotype T. La probabilité des séquences correspond simplement à la probabilité de la suite des étiquettes étant donné leur génotype, ce qui s'exprime $P(t_i,t_{i+1}|T_i,T_{i+1})$ pour un bigram et $P(t_i,t_{i+1},t_{i+2}|T_i,T_{i+1},T_{i+2})$ pour un trigram.

Le système présenté par Tzoukermann et al. (1997) atteint un taux de succès de 91% seulement en appliquant le module statistique et atteint à 93% en appliquant successivement le module linguistique et le module statistique.

MÉTHODOLOGIE

Pour réaliser ce projet, il était envisageable d'utiliser un système d'étiquetage déjà développé et d'en étudier la répartition des succès et des erreurs en l'utilisant sur un corpus de textes. Cette façon de faire offrait l'avantage de ne pas avoir à disposer d'un corpus étiqueté. En effet, dans un premier temps, il aurait suffit de constituer un corpus de textes choisis pour leur variété lexicale et une certaine unité de genre ; d'y repérer les homographes de type verbe-substantif ; d'étiqueter les textes à l'aide de l'étiqueteur et d'identifier parmi les homographes, les succès et les échecs. Finalement, constituer un programme qui prend en entrée le résultat de l'étiqueteur et qui, en appliquant certaines heuristiques, en aurait amélioré les performances. Utiliser un tel système aurait permis d'économiser le temps de développement du système d'étiquetage. Toutefois, nous avons été confronté rapidement aux limites de cette méthodologie. D'abord, il était difficile d'obtenir les systèmes en question et encore plus difficile d'obtenir les codes sources pour les étudier. De plus, le principal problème du point de vue de notre projet était que les étiqueteurs que nous connaissons utilisent des heuristiques pour corriger les erreurs du modèle statistique sous-jacent (Abeillé et al. 2003, Reyes 1997, Clément 2001). Par conséquent, en utilisant un tel système, nos observations n'auraient pas porté uniquement sur le modèle mathématique, mais aussi sur le résultat de l'application des heuristiques faussant nos observations. Les rapports techniques et articles disponibles ne sont pas assez détaillés pour répondre à toutes nos questions par rapport à l'ensemble des procédés utilisés après l'application des statistiques, la façon dont ils sont combinés et leur impact sur le résultat final. Devant toutes ces contraintes, nous avons décidé de développer notre propre étiqueteur. De cette façon, nous pouvons utiliser la méthode de notre choix et seulement cette méthode, permettant de décrire le comportement des statistiques avant tout autre processus d'amélioration des performances. Pour ce faire, il nous a fallu adopter un modèle statistique et nous doter d'un dictionnaire ainsi que d'un corpus étiqueté.

Par ailleurs, nous avons choisi de restreindre nos observations aux homographes de type verbe-substantif car il s'agit de catégories importantes d'une part, et d'autre part, de catégories très distinctes pour un locuteur. Même sans connaissances approfondies en linguistique, un locuteur moyen sait distinguer un verbe d'une autre unité lexicale par

rapport à l'impression d' « action » réalisée par un agent. Il nous semble que de tels homographes devraient également être aussi évidents pour un étiqueteur et les erreurs engendrées par ces homographes peuvent mieux faire ressortir le comportement d'un modèle mathématique que celles causées par d'autres catégories d'homographes à la sémantique moins marquée. Il se peut que d'autres types d'homographes réagissent différemment au modèle probabiliste, mais nous supposons que le comportement mis en lumière par les homographes verbaux permet d'éclaircir suffisamment de caractéristiques du comportement de notre méthode pour le présent travail.

Choix du modèle mathématique

Parmi les méthodes probabilistes disponibles, c'est celle de Church (1988) qui fut retenue pour le présent travail. La démonstration mathématique est présentée dans Charniak et al. (1993). Les ngrams sont une simplification des chaînes de Markov et leur taux de réussite avoisine celui des modèles plus complexes. Puisque la courbe de croissance du taux de succès ne croit pas de la même façon que celle de la complexité des modèles, nous avions avantage à utiliser un modèle simple et celui-ci des ngrams était tout indiqué pour atteindre les objectifs de ce projet.

Choix du corpus

Un extrait du corpus Le Monde–Paris VII nous a été gracieusement prêté par Anne Abeillé, professeure à l'Université Paris VII. Le corpus est constitué d'extraits du quotidien parisien Le Monde, parus entre 1989 et 1993 et couvre une variété d'auteurs et de domaines tels que l'économie, la littérature, la politique, etc. (Abeillé et al. 2003). Le corpus totalise un million de mots dont les parties du discours ont été étiquetées et vérifiées par plusieurs annotateurs. Il s'agit du même corpus utilisé pour entraîner l'étiqueteur de l'Inalf, suivant la méthode de Brill (Lecomte 1998).

La méthodologie suivie pour l'étiquetage du corpus apparaît dans Abeillé et al. (2003).

Notre extrait consiste en 250 897 parties du discours, dont 202 498 mots et 48 399 ponctuations.

Le corpus identifie la nature grammaticale des mots du français à l'aide de 203 étiquettes. Les étiquettes identifient trois types d'information. D'abord, la catégorie principale, ensuite, une sous-catégorie si cela s'applique et enfin, les traits morphologiques comme le genre et le nombre ou la personne, lorsque applicable. L'ensemble des étiquettes et les détails de l'annotation du corpus sont présenté dans Abeillé et Clément (2002). Leur signification et leurs combinaisons sont présentées dans le tableau ci-dessous:

Catégorie principale	Sous-catégorie	Nombre, genre ou personne
Adjectif	cardinal	Masculin/féminin, singulier/pluriel
	indefini	
	interrogatif	
	ordinal	
	qualificatif	
Adverbe		
	démonstratif	
	exclamatif	
	interrogatif	
Abréviation		
Conjonction	Coordination	
	Subordination	
Clitique	objet	1/2/3e personne, masculin/féminin, singulier/pluriel
	réflexif	
	sujet	
Déterminant	Cardinal	Masculin/féminin, singulier/pluriel
	défini	
	démonstratif	
	exclamatif	
	indéfini	
	interrogatif	
	possessif	1/2/3e personne, masculin/féminin, singulier/pluriel
Interjection		
Substantif	commun	Masculin/féminin, singulier/pluriel
Préposition		
Pronom	cardinal	Masculin/féminin, singulier/pluriel
	démonstratif	1/2/3e personne, masculin/féminin, singulier/pluriel
	indéfini	
	interrogatif	
	ordinal	
	personnel	
	possessif	
	relatif	
Ponctuation	Forte	
	Faible	
Verbe	C	1/2/3e personne, singulier/pluriel
	F	
	G	
	I	
	J	
	K	
	P	
	Subjonctif présent	
	subjonctif imparfait	
	impératif	
	infinitif	

Tableau 0-1 Informations lexicales représentées par les étiquettes.

Dans le corpus, les mots composés sont considérés comme un seul et même mot dans le corpus. Par exemple, les trois unités lexicales « caisse de retraite » sont identifiées dans le corpus par la seule étiquette N-C-fs, soit, Nom Commun féminin singulier. Cependant, les constituants des mots composés sont également étiquetés ce qui permet de les décomposer.

Dans l'exemple précédant, les constituants sont identifiés comme tel : caisse/N-C-fs, de/P, retraite/N-C-fs.

```
<s>
<w lemma= »il » ei= »CL3ms » ee= »CL-suj-3ms » cat= »CL » subcat= »suj » mph= »3ms »>Il</w>
<w lemma=″être″ ei=″VI3s″ ee=″V—I3s″ cat=″V″ subcat=″″ mph=″I3s″>était</w>
<w lemma=″de″ ei=″P″ ee=″P″ cat=″P″>de</w>
<w lemma=″le″ ei=″Dfs″ ee=″D-def-fs″ cat=″D″ subcat=″def″ mph=″fs″>la</w>
<w     lemma= »génération »     ei= »NCfs »     ee= »N-C-fs »     cat= »N »     subcat= »C »
mph= »fs »>génération</w>
<w lemma=″qui″ ei=″PROR3fs″ ee=″PRO-rel-3fs″ cat=″PRO″ subcat=″rel″ mph=″3fs″>qui</w>
<w lemma=″avoir″ ei=″VI3s″ ee=″V—I3s″ cat=″V″ subcat=″″ mph=″I3s″>avait</w>
<w lemma=″vingt″ ei=″DCmp″ ee=″D-card-mp″ cat=″D″ subcat=″card″ mph=″mp″>vingt</w>
<w lemma=″an″ ei=″NCmp″ ee=″N-C-mp″ cat=″N″ subcat=″C″ mph=″mp″>ans</w>
<w lemma= »à » ei= »P » ee= »P » cat= »P »>à</w>
<w lemma= »le » ei= »Dfs » ee= »D-def-fs » cat= »D » subcat= »def » mph= »fs »>la</w>
<w     lemma= »libération »     ei= »NCfs »     ee= »N-C-fs »     cat= »N »     subcat= »C »
mph= »fs »>libération</w>
<w lemma=″.″ Ei=″PONCTS″ ee=″PONCT-S″ cat=″PONCT″ subcat=″S″>.</w>
</s>
```

Choix du dictionnaire

Nous ne disposions pas d'un dictionnaire complet pour effectuer notre travail de recherche. Nous avons dû le constituer à partir du matériel qui nous était accessible. Nous avions sous la main une version texte des entrées du Petit Robert (1996) sur CD-ROM. Il s'agit d'une version électronique de la version imprimée de l'ouvrage. Nous disposions ainsi de tous les mots y figurant ainsi que des parties du discours identifiées par l'ouvrage. Cependant, ce dictionnaire ne contient pas les formes fléchies des participes passés ni les formes conjuguées des verbes. Certains adjectifs féminins y sont également absents (par exemple : chanteur, -se). En ce qui concerne les formes féminines des adjectifs et des participes passés, un script fut utilisé pour corriger automatiquement ces lacunes. Les mots ainsi modifiés ou ajoutés ont été vérifiés à la main pour s'assurer de la qualité du travail effectué par le script.

En ce qui a trait aux formes verbales fléchies absentes, la banque de mots du Petit Robert sur CD-ROM fut fusionnée avec la liste des verbes et des formes déclinées du Bescherelle (1997), lequel avait été préalablement numérisé. La numérisation fut révisée pour réduire les erreurs dues à la reconnaissance de caractères. Les parties du discours des formes fléchies des verbes furent ajoutées automatiquement par la consultation des terminaisons. Les exceptions furent étiquetées et vérifiées séparément pour réduire le nombre d'erreurs. En tout, notre dictionnaire final contient 256 000 mots fléchis.

Ce dictionnaire a dû être adapté pour que les étiquettes des parties du discours correspondent à celles utilisées par le corpus Le Monde–Paris VII. Le plus souvent, il ne s'agit que d'une façon différente de noter la même information mais dans d'autres cas, les informations sont différentes. En effet, les étiquettes de notre dictionnaire reflètent la grammaire traditionnelle tandis que celles du corpus s'inspirent plutôt du distributionnalisme. Par conséquent, certains mots ont une étiquette dans le corpus qui ne se retrouve pas dans les grammaires traditionnelles. Par exemple, dans le syntagme « la roue avant droite », *avant* est étiqueté Afs (adjectif féminin singulier) alors que selon la grammaire traditionnelle, ce mot est soit une préposition, un adverbe ou un nom, mais il n'est jamais un adjectif. Dans cet exemple, la grammaire traditionnelle propose un adverbe car il ne s'agit n'y d'une proposition ni d'un nom. Pourtant, le corpus, dans l'esprit distributionnaliste, propose qu'il s'agisse d'un adjectif car une telle position ne peut pas être occupée par une préposition et en isolant « la roue avant » du syntagme, un adverbe ne peut en occuper la place sans en changer le sens.

Ces différences entre le dictionnaire de départ et le corpus dans l'attribution des catégories lexicales aux mots a demandé un travail minutieux d'adaptation du dictionnaire au corpus. Dans un premier temps, les étiquettes pour lesquelles une simple « traduction » d'une notation à l'autre ont été modifiées. Il s'agit principalement des étiquettes des catégories principales des verbes, des noms et des adjectifs, suivies de la personne, du genre et du nombre selon le cas qui ont été réécrites selon les spécifications du corpus. Cette tâche a été automatisée d'après un tableau de translation d'étiquettes et vérifiée manuellement.

En second lieu, il a fallu recatégoriser certains mots. D'une part, il a fallu remplacer les catégories du dictionnaire qui ne correspondaient pas à celles du corpus et d'autre part, retrancher les étiquettes du dictionnaire qui ne sont pas prévues par le corpus.

Parmi les changements les plus importants, mentionnons les pronoms faibles, c'est-à-dire les pronoms personnels et les pronoms réflexifs dont l'étiquette principale est devenue respectivement clitique sujet (CL-suj) et clitique objet (CL-obj). Ce genre de changements n'a concerné que des mots de catégories fermées et les conversions n'ont pas été nombreuses. Ce travail a été effectué manuellement, sans recours à un script

d'automatisation. La liste exhaustive des mots des catégories fermées selon les spécifications du corpus figure dans Abeillé et Clément (2002).

Une fois cette tâche accomplie, il s'est agi de retirer du dictionnaire les étiquettes non modifiées et non présentes dans l'ensemble d'étiquettes utilisé par le corpus.

Extraction des statistiques

En accord avec le modèle adopté, trois types de données ont été extraites du corpus. Cette extraction a porté sur les deux premiers tiers du corpus. Le dernier tiers a été réservé pour l'évaluation de l'étiqueteur. Les deux tiers comportent 165 605 unités dont 133 611 mots lexicographiques.

Les probabilités à priori

Il s'agit, pour un mot donné, du nombre de chacune des catégories grammaticales qui lui ont été attribuées sur un nombre donné d'occurrences. Ces valeurs sont relativisées entre elles pour donner des pourcentages. Ces pourcentages correspondent à la probabilité d'observer chacune des catégories sans connaissances à priori du contexte dans lequel elles sont susceptibles d'apparaître.

Pour chacune des étiquettes d'un mot donné dans le dictionnaire, le calcul suivant a été appliqué pour déterminer leur probabilité à priori :

> Nb d'occurrence de l'étiquette ÷ Nb d'occurrences total du mot.

Ce qui donne, par exemple :

```
je#CL1ms:0.748201 CL1fs:0.244604 PRO1ms:0.007194
industrie#NC:0.820896 N:0.149254 NP:0.014925 D:0.014925
moi#PRO1ms:0.555556 PROfs:0.222222 PRO1fs:0.166667 NP:0.055556
inscrits#VKmp:0.666667 A:0.166667 NC:0.166667
participants#NC:0.875000 A:0.125000
```

Malgré tout, de nombreux mots n'apparaissaient pas dans le corpus avec telle ou telle étiquette de sorte que les probabilités à priori n'étaient pas calculables pour chaque étiquette possible pour un mot. Pour évaluer les probabilités manquantes, nous avons suivi la recommandation de Church (1988) d'augmenter le nombre d'occurrences de toutes les étiquettes de un, y compris les étiquettes non observées (section 0).

Les génotypes

Habituellement, les corpus à partir desquels les probabilités sont tirées sont d'une taille d'environ un million de mots. Le nôtre ne représente qu'un peu plus du quart du corpus d'un million de mots Le Monde–Paris VII et l'extraction des probabilités ne provient que du deux tiers de cet échantillon. Plusieurs mots présents dans le dictionnaire n'ont aucune occurrence dans le corpus et donc aucune probabilité à priori. Pour contrer cette lacune, nous avons évalué les génotypes (section 0) présents de notre corpus et attribué ces probabilités à priori aux mots de notre dictionnaire qui n'apparaissent pas dans le corpus.

D'abord, tous les génotypes du dictionnaire ont été recensés. Ensuite, pour chacun des mots du corpus, le nombre d'occurrences de l'étiquette du mot en question a été augmenté de 1 pour le génotype correspondant. Après avoir compilé toutes les occurrences du corpus, la probabilité a priori de chacun des génotypes a été calculée de la même façon que celle des mots. Finalement, les mots n'ayant pas d'occurrence dans le corpus se sont fait attribuer les probabilités a priori de leur génotype.

Les bigrams et trigrams

Enfin, les probabilités pour chacun des bigrams et des trigrams présents dans le corpus ont été compilées. Les bigrams et les trigrams comprennent les ponctuations et les début et fin de phrase.

Pour les bigrams, le nombre d'occurrences de l'étiquette Y précédée de X divisé par le nombre d'occurrences de l'étiquette Y correspond à la probabilité d'observer l'étiquette Y sachant X. Autrement dit, il s'agit de la probabilité d'occurrence du bigram XY sachant X.

Voici un extrait des probabilités de quelques bigrams :

```
D       + PROR3ms : 0.00612144955926
PROR3ms + P       : 0.660611065235
PROfp   + VG      : 0.680272108844
CC      + VG      : 0.222640863162
PROR3ms + D       : 12.0561519405
```

Pour les trigrams, le nombre d'occurrences de l'étiquette Z précédée de la suite d'étiquette XY divisé par le nombre d'occurrences de la suite d'étiquettes XY correspond à la

probabilité d'observer l'étiquette Z sachant XY. Autrement dit, il s'agit de la probabilité d'occurrence du trigram XYZ sachant XY. En voici quelques exemples :

```
VG    + ADV  + VW     : 2.65486725664
VI3s  + D    + PONCTW : 0.671140939597
VF3p  + A    + P      : 33.3333333333
CL3fp + VP3p + VKmp   : 0.625
CL3mp + VKms + ADV    : 33.3333333333
```

Prototype d'étiqueteur

L'étiqueteur développé pour ces travaux prend en entrée une suite de mots déjà segmentée sous forme d'une liste présentant un mot par ligne. Cette façon de faire offrait l'avantage de ne pas avoir à développer de segmenteur de mots ni de segmenteur de phrase, c'est-à-dire de script qui identifie les frontières de mots et de phrases. D'une part, le présent travail ne traite pas de la problématique entourant la segmentation d'un texte en mots et en phrases et d'autre part, le fait de pouvoir s'en passer permet d'éliminer la répercussion des erreurs de segmentation sur l'étiquetage. Le présent travail s'intéresse à la performance du modèle des ngrams appliqué à l'étiquetage automatique de texte et le fait de partir d'une segmentation exempte d'erreur permet de mieux atteindre cet objectif.

L'étiqueteur utilise trois fichiers de données : d'abord un dictionnaire des mots du français accompagnés de leurs probabilités lexicales à priori, ensuite une liste des bigrams et de leurs probabilités d'occurrence et enfin, un fichier semblable pour les probabilités des trigrams.

L'algorithme de l'étiqueteur prend la forme générale suivante :

```
La liste des étiquettes attribuées par l'algorithme est prise en note. Elle est nécessaire
pour pouvoir tenir compte des étiquettes précédentes pour le calcul des trigrams et des
bigrams.

Pour chaque mot en entrée:
    Les étiquettes possibles sont tirées du dictionnaire

    S'il n'y a qu'une étiquette possible alors le terme n'est pas ambigu:
        l'étiquette du mot est conservée dans une liste

    Si plusieurs étiquettes sont possibles alors le terme est ambigu:
        Pour chaque étiquette possible
            La probabilité du trigram formé des deux étiquettes précédentes suivies de
             l'étiquette en question est extraite du fichier de données;
            La probabilité du trigram est multipliée par la probabilité à priori du mot
             courant;
            La probabilité la plus élevée l'emporte et détermine l'étiquette la plus
             probable;
            L'étiquette la plus probable est ajoutée à la liste.

    Si aucun trigram n'est dans la base de données, alors les bigrams sont consultés
        Pour chaque étiquette possible
           La probabilité du bigram formé de l'étiquette précédente suivie de
            l'étiquette en question est extraite du fichier de données;
           La probabilité du bigram est multipliée par la probabilité à priori du mot
            courant;
           La probabilité la plus élevée l'emporte et détermine l'étiquette la plus
            probable;
           L'étiquette la plus probable est ajoutée à la liste.

    Si aucun bigram n'est dans le fichier de données, alors l'étiquette la plus
     probable est ajoutée à la liste.
```

Évaluation de la qualité du prototype

Avant d'utiliser le prototype en vue de tirer des conclusions sur le modèle utilisé, il a fallu d'abord s'assurer de la qualité du prototype. Par conséquent, il a fallu s'assurer que l'algorithme du prototype ne contienne pas d'erreur. À cette fin, nous nous sommes assuré de deux choses. D'abord, que l'algorithme faisait bel et bien ce qui lui était demandé. Pour ce faire, nous avons calculé manuellement une centaine de trigrams et de bigrams et comparé ces valeurs avec celles du prototype en nous assurant qu'elles étaient équivalentes. Ensuite, il nous est apparu que comparer le prototype aux résultats publiés d'autres étiqueteurs était une autre façon de vérifier si notre prototype était bien conçu. Nous avons émis l'hypothèse que si le prototype était bien conçu nos résultats seraient comparables à ceux publiés.

Cette évaluation a porté sur un tiers du corpus, l'autre deux tiers ayant été réservé à l'extraction des données statistiques.

Cette évaluation a été faite en faisant varier l'ensemble d'étiquettes et la façon d'évaluer les probabilités a priori dans le but de la comparer avec des résultats déjà publiés. Cette façon de faire a également permis de détailler le rendement de notre prototype et ces données sont réutilisées plus loin dans le travail.

Les performances du prototype ont été évaluées avec deux ensembles d'étiquettes, un premier dit « étendu » et un second dit « restreint ». L'ensemble d'étiquettes étendu se compose de toutes les étiquettes possibles pour le vocabulaire alors que l'ensemble restreint correspond aux étiquettes principales (verbe, nom, adjectif, etc.), sans considération de sous-catégorie ni de traits morphologiques de genre, nombre et personne. L'ensemble étendu comprend un total de 203 étiquettes alors que l'ensemble restreint en compte 38 au total.

Également, le prototype fut testé en utilisant trois types de probabilités à priori :

1. les probabilités telles qu'observées dans le corpus d'entraînement sans génotypes pour estimer celles des mots inconnus;
2. les probabilités tirées du corpus complétées de celles des génotypes pour les mots n'apparaissant pas dans le corpus et enfin;
3. l'utilisation exclusive des génotypes en guise de probabilité à priori.

Taux de succès du prototype

Les deux tiers du corpus ont servi d'ensemble d'entraînement du prototype. Cet ensemble était constitué de 165 605 unités comprenant des étiquettes pour les débuts et les fins de phrases ainsi que tous les signes de ponctuation. De ces unités, 133 611 étaient des mots au sens lexicographique. Le dernier tiers a constitué l'ensemble de test sur lequel les mesures ont été prises. Cet ensemble était constitué de 85 292 unités, dont 68 837 mots. Les taux de succès du prototype figurent dans le Tableau 0-2.

	Ensemble restreint (38 étiquettes)	Ensemble étendu (203 étiquettes)
Vocabulaire du corpus	90.33%	84.54%
Vocabulaire du corpus + Génotypes	92.41%	87.35%
Génotypes seulement	73.54%	71.59%

Tableau 0-2 Taux de succès du prototype.

D'abord, en n'utilisant que les mots compris dans le corpus d'entraînement, le prototype obtient un taux de réussite de 90.33% avec l'ensemble d'étiquettes restreint alors qu'en utilisant l'ensemble étendu d'étiquettes, ce taux chute à 84.54%. Dans ce cas, le prototype est désavantagé par les mots inconnus, c'est-à-dire les mots apparaissant dans le corpus d'évaluation qui n'apparaissent pas dans celui d'entraînement.

Ensuite, une seconde évaluation a été réalisée en utilisant les génotypes (voir section 0) pour donner une probabilité à priori aux mots inconnus. Dans ce cas, le prototype a atteint un taux de réussite de 92.41% avec l'ensemble d'étiquette restreint et un taux de 87.35% à l'aide de l'ensemble étendu. Une amélioration respective de 2.08% et de 2.81% est observable par rapport à l'utilisation exclusive du vocabulaire du corpus d'entraînement.

Enfin, l'utilisation exclusive des génotypes a été utilisée comme probabilité à priori. Il était intéressant en effet de se faire une idée de l'utilité des génotypes. Étonnamment, les taux de réussite ne sont pas élevés. Le prototype atteint 73.54% en utilisant l'ensemble restreint d'étiquettes et 71.59% avec l'ensemble étendu. Ces résultats ne corroborent pas ceux rapportés dans TZOUKERMANN et al. (1995) et TZOUKERMANN et al. (1997), en se rapportant respectivement aux lignes 16, 17 et 19 du tableau Tableau 0-3 Comparatif des étiqueteurs du français publiés. Cela nous a obligé à revisiter le code du prototype pour s'assurer que des erreurs n'étaient pas demeurées incorrigées. Pourtant, le code n'est pas fautif. Dans ce cas, nous supposons que les corpus d'entraînement (10 000 et 86 000 mots),

les corpus d'évaluation (1000 et 2500 mots) ainsi que les découpage différent des unités complexes sont responsables de la différence du taux de réussite.

Comparaison avec d'autres étiqueteurs français

Le Tableau 0-3 présente l'ensemble des systèmes d'étiquetage du français que nous avons recensés dans la littérature. Lorsque cela est possible, les informations suivantes y figurent : le nombre d'étiquettes utilisées par le système, le nombre de mots ayant servi à entraîner et évaluer le système, le taux de réussite et le type d'algorithme utilisé. Lorsque les nombres de mots du corpus d'entraînement sont égaux à ceux du corpus d'évaluation, il s'agit du même corpus ayant servi aux deux tâches. Cinq types sont présents :

1. Système d'apprentissage de règles à la Brill
2. N-grams
3. Règles implémentées manuellement
4. Chaînes de Markov
5. Combinaison de système

L'utilisation des génotypes est indiquée lorsque ces derniers sont utilisés pour évaluer les probabilités à priori. Pour faciliter la comparaison, les valeurs du Tableau 0-2 concernant notre prototype ont été répétées aux lignes A à D dans le Tableau 0-3.

#	Étiqueteur	Nb d'étiquettes	Nb de mots du corpus		Taux de réussite %	Type
			Entraînement	Évaluation		
A	Notre prototype	203	165 605	85 292	**84.54%**	3-grams>2-grams>unigrams
B	Notre prototype	38	165 605	85 292	**90.33%**	3-grams>2-grams>unigrams
C	Notre prototype	203	165 605	85 292	**87.35%**	3-grams>2-grams>unigrams + génotypes
D	Notre prototype	38	165 605	85 292	**92.41%**	3-grams>2-grams>unigrams + génotypes
1	Talana1[1]	110	50 210	50 210	92.00%	Brill
2	2-grams[1]	110	50 210	50 210	86.06%	2-grams
3	2-grams[1]	110	50 210	50 210	92.22%	2-grams + génotypes
4	**3-grams[1]**	**110**	**50 210**	**50 210**	**88.37%**	**3-grams**
5	**3-grams[1]**	**110**	**50 210**	**50 210**	**94.62%**	**3-grams + génotypes**
6	4-grams[1]	110	50 210	50 210	90.44%	4-grams
7	4-grams[1]	110	50 210	50 210	95.76%	4-grams + génotypes
8	5-grams[1]	110	50 210	50 210	92.54%	5-grams
9	5-grams[1]	110	50 210	50 210	96.21%	5-grams + génotypes
10	Talana2[1]	25	50 210	38 801	95.40%	Brill + règles manuelles
11	Talana3[1]	110	50 210	38 801	92.90%	Brill + règles manuelles
12	Test A[2]	88[3]	n/a	5752	96.80%	Chaînes de Markov
13	Test A[2]	88[3]	n/a	5752	98.70%	Règles manuelles
14	Test B[2]	88[3]	n/a	12 000	95.00%	Chaînes de Markov
15	Test B[2]	88[3]	n/a	12 000	97.50%	Règles manuelles
16	Genos[4]	253	10 000	1 000	89.30%	Genotype Unigrams
17	Genos[4]	67	10 000	1 000	90.40%	Genotype Unigrams
18	**Genos[5]**	**72**	**76 000**	**1 500**	**95.00%**	**3-grams>2-grams>unigrams + génotypes**
19	Genos[6]	67	86 000	2 500	91.00%	Unigrams Genotypes
20	**Genos[6]**	**67**	**86 000**	**2 500**	**93.00%**	**3-grams>2-grams>Unigrams + génotypes**

Tableau 0-3 Comparatif des étiqueteurs du français publiés

1. ABEILLÉ et al. (1998)
2. CHANOD & TAPANAINEN (1995a)
3. CHANOD & TAPANAINEN (1995b)
4. TZOUKERMANN et al. (1995)
5. TZOUKERMANN et al. (1996)
6. TZOUKERMANN et al. (1997)

Parmi tous ces systèmes, seuls ceux rapportés aux lignes 4, 5, 18 et 20 sont comparables à notre prototype.

Dans les cas 18 et 20, l'algorithme de base est le même, mais les probabilités à priori ne sont pas évaluées de la même façon. Les génotypes sont utilisés pour évaluer ces dernières

alors que dans le prototype, les génotypes ne sont utilisés que pour compléter les probabilités à priori calculées dans le corpus d'entraînement. Le nombre d'étiquettes utilisées est également différent, 72 et 67 étiquettes pour TZOUKERMANN et al. (1996) et TZOUKERMANN et al. (1997) et 38 et 203 pour le prototype. Dans ce cas, le prototype atteint 84.54% avec l'ensemble d'étiquettes étendu et 90.33% avec l'ensemble restreint comparativement à 93% pour TZOUKERMANN et al. (1997) avec 72 étiquettes et 95% pour TZOUKERMANN et al. (1996) avec 67 étiquettes.

En ce qui concerne les systèmes présentés aux lignes 4 et 5, l'algorithme diffère légèrement dans le fait que seuls les 3-grams sont utilisés. Dans le prototype, si aucun 3-gram n'est utile, la banque de 2-grams est consultée et si cela est également inutile, alors les probabilités à priori sont alors utilisées telles quelles. À la ligne 4, le système n'utilise que le vocabulaire du corpus d'entraînement pour évaluer les probabilités à priori tandis que celui de la ligne 5 utilise les génotypes en guise de probabilité à priori. Le nombre d'étiquettes utilisées varie également. Le système de la ligne 4, qui n'utilise pas les génotypes, réussit à 88.37% avec 110 étiquettes. Le prototype lui aussi sans utiliser les génotypes, réussit à 84.54% avec 203 étiquettes et à 90.33% avec 38 étiquettes. Le système de la ligne 5, qui utilise les génotypes, réussit à 94.62% avec 110 étiquettes et le prototype, avec les génotypes, réussit à 87.35% avec 203 étiquettes et à 92.41% avec 38 étiquettes. Cependant, les systèmes des lignes 4 et 5 ont été entraînés sur l'ensemble des valeurs qui ont servi pour évaluer les systèmes. Si le même exercice est fait avec le prototype, alors celui-ci réussit à 93.25% avec 203 étiquettes et à 96.85% avec 38 étiquettes.

Enfin, ce qui se dégage vraiment ici, c'est qu'une équivalence ne peut être établie clairement en comparant les résultats de notre prototype et ceux des autres systèmes car trop de paramètres diffèrent d'un système à l'autre. Cependant, les valeurs des taux de réussite, dans l'absolu, varient de 86.06% à 98.70% pour les systèmes répertoriés et ceux du prototype varient de 84.54% à 92.41% et grimpent de 93.25% à 96.85% s'il est entraîné sur l'ensemble du corpus. Les performances du prototype semblent acceptables.

HYPOTHÈSES, RÉSULTATS ET DISCUSSION

Plusieurs mesures ont été prises quant aux aptitudes de désambiguïsation du prototype en vue de répondre aux questions présentées à la page 14 des objectifs présentés au début de ce travail. Ces mesures sont présentées et discutées dans les sections qui suivent et apportent des éléments de réponse aux questions soulevées par les objectifs. Les résultats sont présentés en trois parties. Chacune de ces parties correspond à un objectif.

Les objectifs étant très larges, on ne peut apporter toutes les réponses aux questions qu'ils soulèvent. Par conséquent, des questions plus précises ont été utilisées pour conduire les observations.

Le succès des ngrams

Dans un premier temps, nous avons voulu mesurer l'apport de la désambiguïsation, c'est-à-dire déterminer jusqu'à quel point il y a désambiguïsation de la part des ngrams par rapport à l'attribution de l'étiquette la plus probable selon les probabilités à priori.

Ensuite, nous avons voulu expliquer le succès des ngrams. Quels phénomènes linguistiques les ngrams permettent-ils de désambiguïser? Autrement dit, qu'est-ce que saisissent les ngrams qui leur donne une aptitude à désambiguïser?

Apport de la désambiguïsation dans le taux de succès

Parler de succès des méthodes probabilistes pour la désambiguïsation, c'est prendre pour acquis, dans le cas de notre prototype, que les ngrams réussissent à bien désambiguïser. Cependant réussissent-ils à désambiguïser? Si oui, quelle est l'ampleur de ce succès? Dans quelle proportion l'ambiguïté est-elle réduite?

Pour répondre à ces questions, un prototype d'étiqueteur sans désambiguïsation fut réalisé pour déterminer le taux de réussite minimal à partir duquel peut être calculé l'apport des ngrams. Ce prototype consulte le dictionnaire et donne à chaque mot, l'étiquette la plus probable selon les probabilités à priori compilées. Ce prototype a eu droit à la même rigueur que le prototype d'étiqueteur dont il est question au 0 quant à la vérification de son efficacité et la fiabilité de ses résultats. Puisqu'il n'y a aucune autre désambiguïsation que le fait de choisir l'étiquette la plus probable à priori, le taux de succès de ce prototype

correspond au niveau « plancher » de la désambiguïsation, c'est-à-dire qu'il s'agit du rendement minimum que le prototype d'étiquetage doit atteindre car autrement, la désambiguïsation via les ngrams donnerait des résultats pires que la simple attribution de l'étiquette la plus probable.

Le niveau plancher fut calculé avec deux types de probabilités à priori compilées à partir du corpus d'entraînement, soit le vocabulaire du corpus et ce même vocabulaire complété des génotypes. Le taux de reconnaissance du niveau plancher paraît dans le Tableau 0-1.

	Ensemble restreint (38 étiquettes)	**Ensemble étendu (203 étiquettes)**
Vocabulaire du corpus	89.39%	81.55%
Vocabulaire du corpus + Génotypes	91.84%	86.08%

Tableau 0-1 Taux de réussite de l'attribution de l'étiquette la plus probable.

En comparant ce niveau plancher avec les taux de succès atteint par le prototype utilisant les ngrams, on peut calculer l'amélioration apportée par l'étape de désambiguïsation. Ces valeurs sont présentées au Tableau 0-2 suivant.

Taux de réussite du prototype – taux de réussite plancher	**Ensemble restreint (38 étiquettes)**	**Ensemble étendu (203 étiquettes)**
Vocabulaire du corpus	90.33% - 89.39% = **0.94%**	84.54% - 81.55% = **2.99%**
Vocabulaire du corpus + Génotypes	92.41% - 91.84% = **0.57%**	87.35% - 86.08% = **1.27%**

Tableau 0-2 Amélioration du taux de réussite dû à la désambiguïsation utilisant les ngrams

Le prototype utilisant les ngrams améliore donc le taux de succès de l'étiquetage dans son ensemble. Mais quelle est la proportion de cette amélioration sur l'ensemble de l'ambiguïté? Autrement dit, combien d'ambiguïtés demeurent irrésolues? Pour le savoir, il serait possible de prendre l'écart entre le niveau plancher et le taux réussite parfait (100%)

et calculer la proportion présentée dans le Tableau 0-2 sur cet écart, mais nous avons préféré connaître la capacité maximum du modèle et calculer la proportion de taux de succès gagnée par rapport à ce maximum. D'une part, il est intéressant de connaître le maximum que peuvent atteindre les ngrams et d'autre part, calculer une proportion en incluant une partie d'ambiguïté inatteignable par le modèle serait ignorer une partie importante de son comportement.

Pour approximer ce maximum, le prototype utilisant les ngrams a été entraîné sur l'ensemble du corpus et réévalué sur le même tiers du corpus ayant servi à l'évaluation de l'entraînement réalisé sur les deux tiers du corpus. Cette pratique transgresse les pratiques scientifiques habituelles car elle ouvre la porte au surentraînement, c'est-à-dire à une spécialisation sur le corpus d'entraînement. Ce surentraînement empêche alors de généraliser les performances et de supposer un même taux de réussite sur des données fraîches. En effet, « quel est le taux de succès sur des données n'ayant jamais été rencontrées » est la question à laquelle ces évaluations tentent de répondre. Par contre, dans notre cas, la question est plutôt de savoir quel est le taux maximum auquel on peut s'attendre du prototype si les conditions étaient idéales. Nous croyons qu'un vocabulaire entièrement connu et l'absences de données nouvelles constituent des conditions idéales et en cela, utiliser le même corpus pour entraîner et évaluer, constitue une façon de réduire au minimum les inconnus et par conséquent, d'approximer la limite supérieure d'efficacité du modèle utilisé.

Ce taux de succès optimum est présenté dans le Tableau 0-3.

	Ensemble restreint (38 étiquettes)	Ensemble étendu (203 étiquettes)
Vocabulaire du corpus	96.85%	93.25%
Vocabulaire du corpus + Génotypes	96.85%	93.25%

Tableau 0-3 Maximum théorique du taux de succès du prototype.

Étant donné que l'ensemble du vocabulaire était connu, l'utilisation des génotypes pour évaluer les probabilités à priori des mots nouveaux était inutile. Cela explique pourquoi les résultats apparaissant dans le Tableau 0-3 sont égaux pour les deux types de probabilités à priori.

On y constate que même avec de meilleures conditions, le prototype ne parvient pas à s'approcher du 100% davantage que les autres systèmes recensés dans la littérature.

Par ailleurs, on peut calculer la proportion d'ambiguïté résolue par le prototype par rapport au maximum qu'il peut atteindre. Cette proportion apparaît dans le Tableau 0-4.

Amélioration du taux de succès/(maximum théorique-niveau plancher)*100	**Ensemble restreint (38 étiquettes)**	**Ensemble étendu (203 étiquettes)**
Vocabulaire du corpus	0.94/(96.85-89.39) = 12.60%	2.99/(93.25-81.55) = 25.49%
Vocabulaire du corpus + Génotypes	0.57/(96.85-91.84) = 11.38%	1.27/(93.25-86.08) = 17.71%

Tableau 0-4 Proportion du gain en taux de succès attribué à la désambiguïsation des ngrams.

Avec l'utilisation exclusive du vocabulaire rencontré dans le corpus d'entraînement, on constate une réduction de l'ambiguïté de 12.60% en utilisant l'ensemble restreint d'étiquettes. La réduction de l'ambiguïté double à 25.49% en utilisant l'ensemble étendu d'étiquettes.

Ce qui est étonnant, c'est que les résultats avec l'utilisation des génotypes sont légèrement inférieurs avec une réduction de l'ambiguïté de 11.38% en utilisant l'ensemble restreint d'étiquettes et une réduction de 17.71% avec l'utilisation de l'ensemble étendu ce qui correspond à une amélioration d'environ 33%.

Il est intéressant de noter que malgré le fait que l'utilisation de l'ensemble d'étiquette restreint conduit à des taux de succès plus élevés en général (Tableau 0-1), l'utilisation de l'ensemble étendu de son côté, permet une désambiguïsation plus efficace (Tableau 0-4).

Cela suggère que plus il y a d'informations sur les mots en question, meilleure est la désambiguïsation, car les combinaisons d'étiquettes sont plus contraintes étant donné l'accord notamment entre le genre et le nombre des étiquettes qui doivent être licites à l'intérieur des ngrams.

Les tableaux Tableau 0-1, Tableau 0-2 et Tableau 0-3 ainsi que l'analyse faite des données qui y sont présentées suggèrent que les ngrams améliorent la désambiguïsation. Cependant, le taux d'amélioration de l'application seule des ngrams est relativement bas tel que présenté par le Tableau 0-4. Pour un ensemble restreint d'étiquettes, ce taux de désambiguïsation oscille autour de 12% de l'ambiguïté qu'il est théoriquement possible de résoudre avec les ngrams. Avec un ensemble d'étiquettes plus étendu, cette valeur grimpe à environ 25%. Il reste donc place à l'amélioration d'autant plus qu'il s'agit de valeurs calculées par rapport au maximum du modèle (Tableau 0-3) et non à partir du taux de désambiguïsation parfait (100%).

Ce taux de désambiguïsation est-il réparti uniformément sur l'ensemble du vocabulaire? Autrement dit, est-il le même pour tous les types d'ambiguïté? Pour s'en donner une idée, le taux de désambiguïsation des homographes de type verbe/substantif a été calculé pour vérifier si en ciblant un type d'ambiguïté, le taux de désambiguïsation varie.

Les mêmes valeurs ont été calculées en ne considérant que les homographes de type verbe/substantif. Le Tableau 0-5 présente le taux de succès atteint par le prototype sur ce type d'ambiguïté.

	Ensemble restreint (38 étiquettes)	**Ensemble étendu (203 étiquettes)**
Vocabulaire du corpus	91.74% **(1.41%)**	88.15% **(3.61%)**
Vocabulaire du corpus + Génotypes	92.34% **(-0.07%)**	90.20% **(2.85%)**

Tableau 0-5 Taux de succès du modèle de trigrams.

Entre parenthèses et en caractères gras apparaît la différence de valeur par rapport au taux de désambiguïsation général atteint par le prototype (Tableau 0-1), c'est-à-dire sur l'ensemble du vocabulaire, sans cibler un type d'ambiguïté en particulier. On remarque que dans un seul cas (ensemble d'étiquettes restreint + génotypes), la désambiguïsation est inférieure à la moyenne générale, quoique bien peu inférieure (-0.07%). Les meilleurs gains de désambiguïsation se trouvent en utilisant l'ensemble étendu d'étiquettes, soit 3.61% en n'utilisant que le vocabulaire du corpus et 2.85% en utilisant les probabilités à priori des génotypes pour les mots nouveaux.

	Ensemble restreint (38 étiquettes)	Ensemble étendu (203 étiquettes)
Vocabulaire du corpus	88.77% **(-2.97%)**	84.76% **(-3.39%)**
Vocabulaire du corpus + Génotypes	89.51% **(-2.83%)**	87.97% **(-2.23%)**

Tableau 0-6 Taux de succès du modèle plancher sur les homographes de type verbe/substantif.

Le Tableau 0-6 présente la performance sur ce type d'ambiguïté du modèle de base, c'est-à-dire l'attribution de l'étiquette la plus probable à priori. Entre parenthèses, apparaissent les différences entre le taux de désambiguïsation général et le taux de désambiguïsation des homographes en question. On remarque que le prototype est en dessous de sa moyenne dans tous les cas. Par conséquent, **cela suggère que l'utilisation des probabilités à priori comme seule méthode de désambiguïsation n'est pas adaptée aux homographes de type verbe/substantif.**

	Ensemble restreint (38 étiquettes)	Ensemble étendu (203 étiquettes)
Vocabulaire du corpus	97.12% (5.38%)	96.44% (8.29%)
Vocabulaire du corpus + Génotypes	97.12% (4.78%)	96.44% (6.24%)

Tableau 0-7 Taux de succès théoriquement maximal du prototype sur les homographes de type verbe/substantif.

Le Tableau 0-7 quant à lui présente le taux de succès maximal que pourrait potentiellement atteindre les ngrams dans le cas de la désambiguïsation d'homographes de type verbe/substantif. Les valeurs entre parenthèses représentent la différence entre le maximum théorique que pourraient atteindre les ngrams sur l'ensemble du lexique et le maximum qu'ils pourraient atteindre s'ils étaient appliqués sur les homographes de type verbe/substantif seulement. On constate un gain important par rapport à la moyenne, **ce qui suggère que l'utilisation des ngrams est plus appropriée pour les homographes de type verbe/substantif que pour le reste du lexique.**

En valeurs absolues, les gains affichés par le Tableau 0-5 sont supérieurs à ceux calculés sur l'ensemble du vocabulaire (Tableau 0-2) exception faite de l'ensemble restreint d'étiquettes + génotype. Mais ces valeurs absolues sont-elles proportionnellement plus élevées que dans le cas général? Le Tableau 0-8 présente la proportion de l'ambiguïté que peut résoudre le modèle qui est effectivement résolue.

Amélioration du taux de succès/(maximum théorique-niveau plancher)*100	Ensemble restreint (38 étiquettes)	Ensemble étendu (203 étiquettes)
Vocabulaire du corpus	1.41/(97.12-88.77) = 16.89% **(4.29%)**	3.61/(96.44-84.76) = 30.91% **(5.42%)**
Vocabulaire du corpus + Génotypes	-0.07/(97.12-89.51) = -0.92% **(-12.30%)**	2.85/(96.44-87.97) = 33.65% **(15.94%)**

Tableau 0-8 Proportion du gain en taux de succès attribué à la désambiguïsation des ngrams pour les homographes de type verbe/substantif.

Entre parenthèses apparaît la différence entre le taux d'amélioration général (Tableau 0-4) et celui dont il est question ici, soit le taux d'amélioration concernant les homographes. Il y a amélioration dans tous les cas, sauf lorsque les génotypes sont utilisés pour les mots nouveaux en utilisant l'ensemble restreint d'étiquettes. D'une part, les génotypes nuisent lorsque l'ensemble d'étiquettes est restreint et d'autre part, ils améliorent de façon appréciable le taux de succès lorsqu'ils sont calculés avec l'ensemble d'étiquettes étendu. On peut donc émettre l'hypothèse que plus les étiquettes exprimant les génotypes sont précises, plus ces derniers sont efficaces. En effet, peut-être qu'en utilisant des étiquettes trop sommaires (ensemble restreint) les génotypes groupent des mots qui ont un comportement syntaxique moins homogène ce qui détériore les généralisations que peuvent faire les génotypes.

Les valeurs présentées par les tableaux Tableau 0-5, Tableau 0-6, Tableau 0-7 et Tableau 0-8 suggèrent donc que les ngrams sont plus adaptés à la désambiguïsation des homographes de type verbe/substantif qu'au reste des mots ambigus.

L'adéquation des ngrams dans la désambiguïsation

Si les ngrams permettent d'améliorer la désambiguïsation, cela porte à croire qu'ils permettent de modéliser une information linguistique. Puisque les ngrams expriment des combinaisons de catégories syntaxiques, ce qu'ils peuvent saisir ne peut avoir de lien qu'avec la syntaxe. Par conséquent, à quel aspect de la syntaxe correspondent-ils et quelle partie de la syntaxe leur échappe? En effet, s'ils saisissaient l'ensemble du modèle syntaxique, les résultats seraient près du taux de succès parfait.

Que sait-on de la syntaxe, qui puisse être utile dans ce cas? D'abord, l'ordre des mots, en français, est important. En effet, la phrase déclarative, qui sert de référence, présente le sujet qui précède le verbe qui précède lui-même l'objet. On dit ainsi du français que c'est une langue de type SVO : Sujet-Verbe-Objet. Les phrases ne sont pas seulement une juxtaposition de mots. Les unités lexicales peuvent former ensemble des unités plus grandes que le mot mais plus petites que la phrase : ce sont les syntagmes. En fait, ce sont ces unités, les syntagmes, qui sont d'abord régies par l'ordre SVO, et les unités constituant les syntagmes sont à leur tour contraintes par un ordre à l'intérieur même d'un syntagme, en fonction de la nature de ce syntagme.

Puisque les ngrams prédisent des parties du discours pour de très courtes séries de mots (trois au maximum pour notre prototype), il semble raisonnable de penser que les ngrams saisissent en plus grande partie des données sur la structure interne des syntagmes puisque de petites séries sont plus susceptibles de se trouver à l'intérieur d'un constituant que si elles chevauchent la frontière entre deux constituants.

C'est sur cette prémisse que, pour expliquer le succès des méthodes probabilistes, nous avons émis l'hypothèse que les structures syntaxiques, les syntagmes, trouvées en corpus sont suffisamment régulières et récurrentes et que ces structures peuvent être représentées par des séries de modèles simples (ngrams). Autrement dit, en étant plus nombreux, les ngrams représentant des suites internes aux syntagmes auront statistiquement plus de poids pour prédire les suites à l'intérieur des syntagmes et les ngrams chevauchant les syntagmes et dont la structure devrait représenter des suites illicites à l'intérieur des syntagmes devraient avoir un poids statistique supérieur seulement aux limites des syntagmes.

L'inventaire des suites syntaxiques du corpus d'entraînement dans lesquelles apparaissent les noms et les verbes est présenté au Tableau 0-9. Puisque les syntagmes ne sont pas identifiés dans le corpus, il a fallu approximer cette information. Pour les syntagmes nominaux, l'ensemble des suites comprises entre une frontière de syntagme (début de phrase, virgule) ou un déterminant et un nom ont été retenues. Pour les syntagmes verbaux, il s'agit des suites comprises entre une ponctuation, un pronom en position sujet (cette information est disponible dans le corpus) ou un syntagme nominal sujet.

Pour faciliter la généralisation, les adverbes n'ont été pris en compte dans les structures du Tableau 0-9. En tout, 43 596 noms et verbes ont été recensés dont 30 960 noms et 12 636 verbes.

Rang	Pct %	Syntagmes nominaux
1	47.00% (20521/43596)	Det + **Nom**
2	15.37% (6713/43596)	Divers + **Nom**
3	4.65% (2031/43596)	Sujet + **V. conj.**
4	4.25% (1854/43596)	Prep + **Infinitif**
5	4.11% (1795/43596)	**Nom**
6	3.90% (1703/43596)	Det + Nom + Prep + **Nom**
7	3.33% (1454/43596)	Det + Nom + **V. conj.**
8	3.17% (1385/43596)	**V. conj.**
9	2.84% (1242/43596)	Divers + **V. conj.**
10	2.01% (876/43596)	Nom propre + **V. conj.**
11	1.48% (646/43596)	PRO rel. + **V. conj.**
12	1.03% (451/43596)	Refl. + **V. conj.**
13	0.78% (340/43596)	Det + Nom + Adj + **V. conj.**
14	0.72% (315/43596)	Sujet + Objet + **V. conj.**
15	0.66 % (287/43596)	Obj. + **Infinitif**
16	0.64 % (279/43596)	Nom + **V. conj.**
17	0.52 % (228/43596)	Det + Adj. card. + **Nom**
18	0.52 % (228/43596)	Sujet + V. conj. + **Infinitif**
19	0.50 % (217/43596)	Refl. + **Infinitif**
20	0.45 % (197/43596)	Det + Nom + PRO rela. + **V. conj.**
21	0.35 % (155/43596)	Sujet + PRO refl. + **V. conj.**
22	0.34 % (148/43596)	Det + Nom + PRO refl. + **V. conj.**
23	0.29 % (125/43596)	Det + Adj + Nom + **V. conj.**
24	0.27 % (120/43596)	Det + Nom + Prep + Nom + **V. conj.**
25	0.25 % (111/43596)	Det + Nom + DetConcat + Nom + **V. conj.**
26	0.22 % (96/43596)	V. conj. + Refl. + **Infinitif**
27	0.18 % (79/43596)	Sujet + PRO rela. + **V. conj.**

Tableau 0-9 Séquences d'étiquettes se terminant par un nom ou un verbe les plus fréquentes dans le corpus d'entraînement.

On constate que les noms sont en plus grand nombre que les verbes. En effet, une proposition doit normalement contenir un verbe, mais elle peut contenir plus d'un nom. Par conséquent, il est normal qu'a priori, plus de nom se trouvent dans le corpus. On remarque, grâce à la ligne 2 du Tableau 0-9 qu'un relativement grand nombre de syntagmes nominaux ont une structure (suites de catégories syntaxiques) rare qui ne se généralise pas avec les modèles plus courants. En fait, il s'agit du deuxième cas le plus courant du Tableau 0-9.

Également, la ligne 9 du Tableau 0-9 nous renseigne sur les suites de catégories qui ne se généralisent pas pour les verbes conjugués. Elles sont parmi les plus nombreuses.

Les tableaux Tableau 0-10 et Tableau 0-11 reprennent la même information mais permettent de visualiser les données par rapport respectivement aux syntagmes nominaux et verbaux.

Rang	Pct %	Syntagmes nominaux
1	66.28% (20521/30960)	Det+ **Nom**
2	21.68% (6713/30960)	Divers + **nom**
3	5.80% (1795/30960)	**Nom**
4	5.50% (1703/30960)	Det+ Nom+ Prep+ **Nom**
5	0.74% (228/30960)	Det+ Adj. card. + **Nom**

Tableau 0-10 Séquences d'étiquettes se terminant par un substantif dans le corpus d'entraînement.

Le Tableau 0-10 permet de constater que **près des deux tiers des noms sont précédés d'un déterminant. Cette situation est très bien saisie par les trigrams** bien que dans ce cas, des bigrams seraient suffisants. Avec cette fréquence, la probabilité à priori de rencontrer un nom après un déterminant est très élevée et dans le calcul faisant intervenir les probabilités à priori lexicales, un homographe de type verbe/substantif verra son étiquette SUBSTANTIF favorisée par rapport à celle de verbe.

Par contre, dans le même Tableau 0-10, on constate qu'environ un nom sur cinq n'est pas précédé d'un déterminant (ligne 2). Également, environ un nom sur vingt n'est pas introduit. Dans ce cas, il suit une virgule et est le plus souvent le premier mot d'une incise (« […] M. Labelle, *président de la chambre de commerce*, est fier de […] »).

Enfin, le reste des suites observées correspond à des unités nominales complexes (ligne 4) et, en faible fréquence, à des dates exprimées littéralement (ex. : le 12 janvier) ou à des adjectifs cardinaux quantifiant un substantif (ex. : les deux vice-présidents).

Rang	Pct %	Syntagme verbal
1	15.98% (2031/12636)	Sujet+ **V. conj.**
2	14.59% (1854/12636)	Prep+ **Infinitif**
3	11.44% (1454/12636)	Det+ Nom+ **V. conj.**
4	10.90% (1385/12636)	**V. conj.**
5	9.77% (1242/12636)	Divers + **verbes**
6	6.89% (876/12636)	Nom propre+ **V. conj.**
7	5.08% (646/12636)	PRO rel. + **V. conj.**
8	3.55% (451/12636)	Refl. + **V. conj.**
9	2.68% (340/12636)	Det+ Nom+ Adj+ **V. conj.**
10	2.48% (315/12636)	Sujet+ Objet+ **V. conj.**
11	2.26% (287/12636)	Obj. + **Infinitif**
12	2.20% (279/12636)	Nom+ **V. conj.**
13	1.79% (228/12636)	Sujet+ V. conj. + **Infinitif**
14	1.71% (217/12636)	Refl. + **Infinitif**
15	1.55% (197/12636)	Det+ Nom+ PRO rela. + **V. conj.**
16	1.22% (155/12636)	Sujet+ PRO refl. + **V. conj.**
17	1.16% (148/12636)	Det+ Nom+ PRO refl. + **V. conj.**
18	0.98% (125/12636)	Det+ Adj+ Nom+ **V. conj.**
19	0.94% (120/12636)	Det+ Nom+ Prep+ Nom+ **V. conj.**
20	0.87% (111/12636)	Det+ Nom+ DetConcat+ Nom+ **V. conj.**
21	0.76% (96/12636)	V. conj. + Refl. + **Infinitif**
22	0.62% (79/12636)	Sujet+ PRO rela. + **V. conj.**

Tableau 0-11 Séquences d'étiquettes se terminant par un verbe dans le corpus d'entraînement.

Le Tableau 0-11 présente les données relatives aux suites syntaxiques précédant les verbes dans le corpus d'entraînement. Les types de suites sont plus diversifiés que ceux des syntagmes nominaux au Tableau 0-10. Le cas le plus fréquent est celui d'un verbe précédé d'un pronom sujet. Ce cas est très bien saisi par les trigrams bien qu'un bigram serait suffisant. Le deuxième cas en ordre de fréquence est celui des verbes infinitifs précédés d'une préposition (à, de). Ce genre de séquence est lui aussi, bien traduit par les ngrams. En troisième lieu, on retrouve les verbes immédiatement précédés d'un sujet exprimé par un syntagme nominal de type déterminant + nom. La totalité de ce cas est saisie par les trigrams. En quatrième position, on retrouve un cas similaire à un de ceux trouvés avec les noms, c'est-à-dire, un verbe au début d'une préposition, c'est-à-dire en début de phrase ou précédé d'une virgule. Ce cas est cependant plus diversifié que celui du nom car les noms se trouvaient en grande partie dans une incise. Enfin, en cinquième position, viennent les suites qui ne se sont pas généralisées avec les autres. Ces cas comptent pour près de 60% des cas.

Avec une plus faible fréquence, on retrouve des suites syntaxiques tout de même très intéressantes pour la désambiguïsation. En effet, les séquences 8, 10, 11, 14, 16, 17, 21 et 22 contiennent des éléments révélateurs d'un verbe. Les pronoms réflexifs (séquences 8, 14, 16, 17 et 21) en effet, ne se trouvent que dans l'environnement d'un verbe (Silberztein 1993).

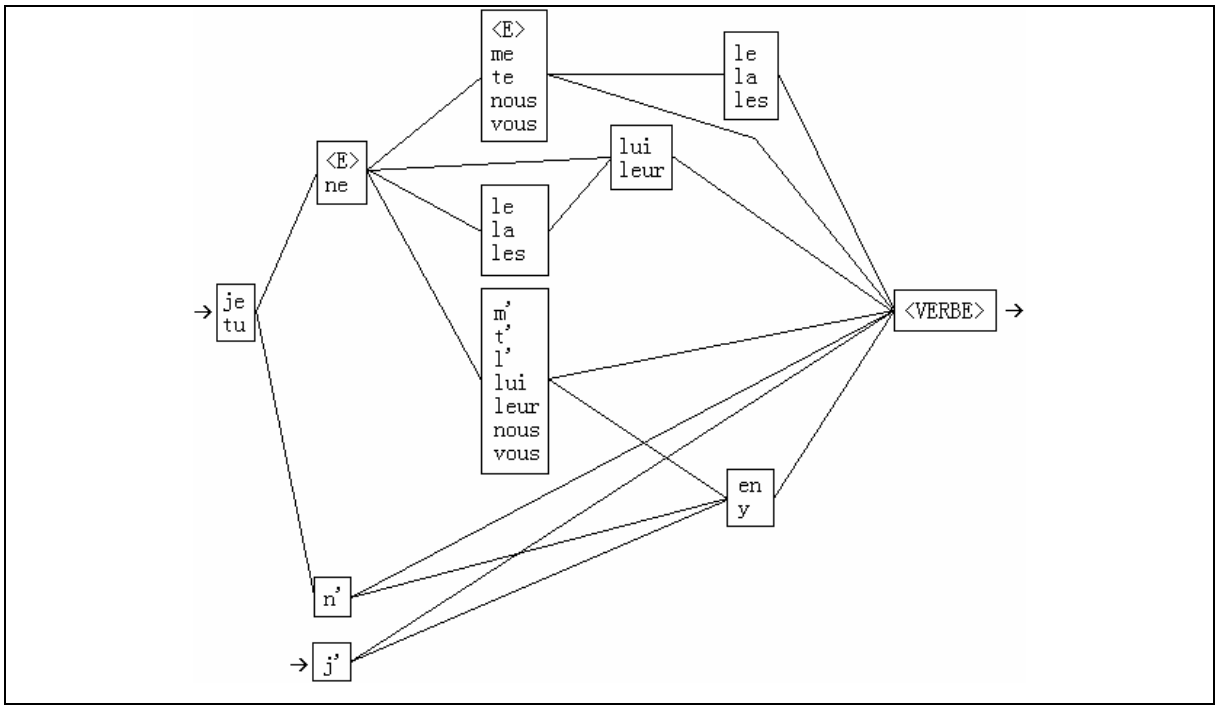


Figure 6-1 Grammaire locale des pronoms je et tu

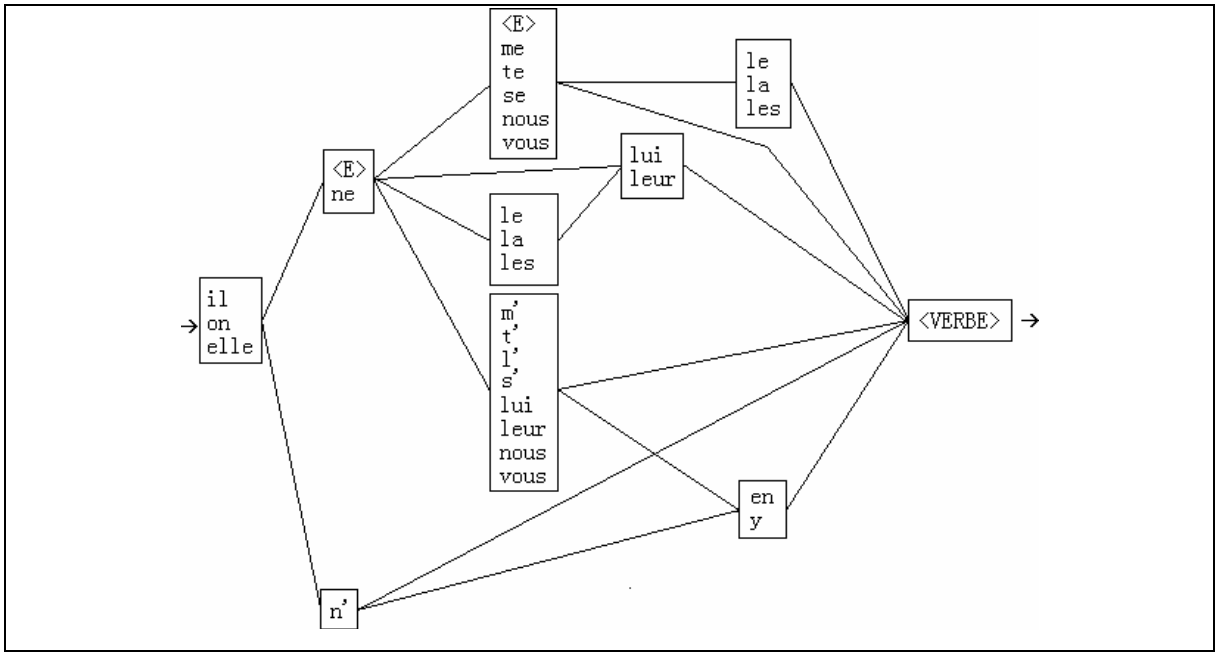


Figure 6-2 Grammaire locale des pronom il, on et elle

Également, dans le cas des séquences 10 et 11, les pronoms objets sont indicateurs de verbes. Enfin, à la séquence 22, la suite pronom personnel suivi d'un pronom objet est également très prédictive d'un verbe.

Il reste les autres types de séquences, soient les séquences 6, 7, 12, 13, 15, 18, 19 et 20.

Les séquences 18 à 20 représentent des verbes précédés d'un syntagme nominal. Les séquences 7 et 15 contiennent un pronom relatif, ce qui indique également un verbe à proximité, à condition que le pronom soit correctement désambiguïsé d'avec les déterminants. Enfin, dans la séquence 13, les pronoms personnels suivis du verbe conjugué ne peuvent être ambigus et devraient aider à la désambiguïsation du verbe infinitif.

Bref, les tableaux Tableau 0-9, Tableau 0-10 et Tableau 0-11 montrent que pour la plupart, les séquences syntaxiques impliquant des noms et des verbes sont régulières et saisissables par des ngrams. Là où il reste un flou à éclaircir est dans les séquences 2 et 9 du Tableau 0-9. Dans ces cas, il s'agit de séquences moins régulières et plus inusitées. Il semble que les données aillent dans le même sens que l'hypothèse que nous avons émise. C'est-à-dire que les séquences les plus fréquentes répondent bien au modèle des ngrams. Cependant, il reste

un ensemble de séquences où le pouvoir des ngrams est peut-être limité. Cela reste à être élucidé.

Les limites des ngrams

Les données précédentes montrent que les ngrams réussissent à désambiguïser les mots ambigus. Plus particulièrement, elles montrent que les ngrams sont plus doués pour la désambiguïsation des homographes de type verbe/substantif que pour les autres types d'ambiguïté. Les observations présentées en 5.1.2 suggèrent que cela est dû à l'abondance de structures syntaxiques qui répondent bien aux ngrams dans les cas des noms et des verbes. Les données suggèrent aussi que les ngrams ne peuvent désambiguïser parfaitement même dans des conditions idéales. Cela est donc d'autant plus vrai dans des conditions non idéales. Dans ce cas, la portion d'ambiguïté résolue est relativement mince comparée à la capacité maximale des ngrams.

Qu'est-ce qui empêche les ngrams d'obtenir de meilleurs résultats? Pour répondre à cette question, nous avons d'abord mis en lumière un comportement des ngrams jusqu'alors masqué par le mode d'évaluation présenté à la 0 et ensuite, nous avons identifié des cas où la désambiguïsation ne devrait pas être un problème pour un locuteur et nous avons confronté le prototype à ces cas.

La section 0 montre que le succès de la désambiguïsation n'est pas toujours logique par rapport aux ngrams et la section 0 montre que des cas supposés évidents du point de vue du système linguistique ne le sont pas toujours pour les ngrams, montrant ainsi que ces derniers ne peuvent saisir qu'une partie de ce système.

Limites dues aux erreurs de désambiguïsation précédentes

Sachant que le prototype s'appuie sur les deux étiquettes précédentes pour calculer la probabilité de la suivante (celle en cours de désambiguïsation), nous avons émis l'hypothèse que des erreurs dans les étiquettes précédentes sont peut-être responsables de l'échec de la désambiguïsation des homographes de type verbe/substantif. Pour vérifier l'hypothèse, les homographes ont été recensés ainsi que l'état de leur désambiguïsation : succès ou échec. De même, pour chacun de ces homographes, l'état de la désambiguïsation des deux étiquettes précédentes, soit celles ayant servi au calcul de la probabilité de

l'étiquette de l'homographe, a été recensé. Ces données sont présentées aux tableaux Tableau 0-12 à Tableau 0-15 et ont été compilées d'après les combinaisons possibles d'ensemble d'étiquettes et de l'utilisation ou non des génotypes.

Ensemble étendu d'étiquettes + prob. à priori du corpus seulement				
Étiquette antéprécédente	**Étiquette précédente**	**Étiquette homographe**	**Nb et %**	**Total**
Succès	Succès	Échec	410 **(6.32%)**	**769 (11.85%)**
Succès	Échec	Échec	174 **(2.68%)**	
Échec	Succès	Échec	130 **(2.00%)**	
Échec	Échec	Échec	55 **(0.85%)**	
Succès	Succès	Succès	4334 **(66.77%)**	**5722 (88.15%)**
Succès	Échec	Succès	717 **(11.05%)**	
Échec	Succès	Succès	540 **(8.32%)**	
Échec	Échec	Succès	131 **(2.02%)**	
			Total :	**6491**

Tableau 0-12 État de désambiguïsation des étiquettes des trigrams ayant servi au calcul de probabilité des homographes.

Ensemble étendu d'étiquettes + génotypes				
Étiquette antéprécédente	**Étiquette précédente**	**Étiquette homographe**	**Nb et %**	**Total**
Succès	Succès	Échec	385 **(5.93%)**	**636 (9.80%)**
Succès	Échec	Échec	130 **(2.00%)**	
Échec	Succès	Échec	87 **(1.34%)**	
Échec	Échec	Échec	34 **(0.52%)**	
Succès	Succès	Succès	4523 **(69.68%)**	**5855 (90.20%)**
Succès	Échec	Succès	759 **(11.69%)**	
Échec	Succès	Succès	483 **(7.44%)**	
Échec	Échec	Succès	90 **(1.39%)**	
			Total :	**6491**

Tableau 0-13 État de désambiguïsation des étiquettes des trigrams ayant servi au calcul de probabilité des homographes.

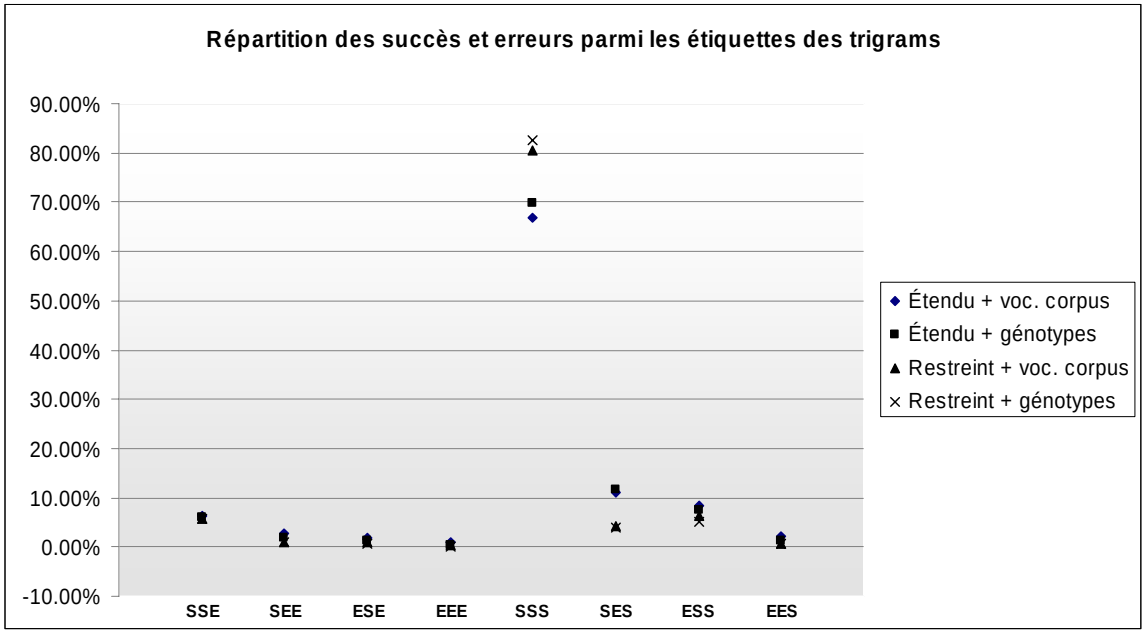
Ensemble restreint d'étiquettes + prob. à priori du corpus seulement				
Étiquette antéprécédente	**Étiquette précédente**	**Étiquette homographe**	**Nb et %**	**Total**
Succès	Succès	Échec	381 **(5.87%)**	**536 (8.26%)**
Succès	Échec	Échec	72 **(1.11%)**	
Échec	Succès	Échec	62 **(0.96%)**	
Échec	Échec	Échec	21 **(0.32%)**	
Succès	Succès	Succès	5217 **(80.37%)**	**5955 (91.74%)**
Succès	Échec	Succès	282 **(4.34%)**	
Échec	Succès	Succès	409 **(6.30%)**	
Échec	Échec	Succès	47 **(0.72%)**	
			Total :	**6491**

Tableau 0-14 État de désambiguïsation des étiquettes des trigrams ayant servi au calcul de probabilité des homographes.

Ensemble restreint d'étiquettes + génotypes				
Étiquette antéprécédente	**Étiquette précédente**	**Étiquette homographe**	**Nb et %**	**Total**
Succès	Succès	Échec	376 **(5.79%)**	**497 (7.66%)**
Succès	Échec	Échec	65 **(1.00%)**	
Échec	Succès	Échec	42 **(0.65%)**	
Échec	Échec	Échec	14 **(0.22%)**	
Succès	Succès	Succès	5368 **(82.70%)**	**5994 (92.34%)**
Succès	Échec	Succès	263 **(4.05%)**	
Échec	Succès	Succès	327 **(5.04%)**	
Échec	Échec	Succès	36 **(0.55%)**	
			Total :	**6491**

Tableau 0-15 État de désambiguïsation des étiquettes des trigrams ayant servi au calcul de probabilité des homographes.

Ces proportions sont sensiblement les mêmes pour toutes les combinaisons possibles comme le montre le graphique 1 qui suit. « S » signifie succès tandis que « E » signifie échec. Ainsi, « SSE » signifie que les étiquettes antéprécédente et précédente ont été désambiguïsées correctement mais que l'étiquette de l'homographe ne l'a pas été.



Graphique 0-1 Répartition des succès et erreurs parmi les étiquettes des trigrams selon toutes les combinaisons d'ensemble d'étiquette et de probabilités à priori possible.

Il est étonnant de constater que dans tous les cas, le prototype présente une répartition semblable de désambiguïsation selon le patron de succès et d'échec parmi les deux étiquettes précédentes.

En isolant les homographes désambiguïsés de ceux qui ne l'ont pas été adéquatement, on obtient les valeurs présentées aux tableaux Tableau 0-16 et Tableau 0-17 qui suivent. Le Tableau 0-16 présente d'abord les homographes ayant été désambiguïsés correctement.

Succès et erreurs parmi les étiquettes précédant un succès				
3gram	Étendu + voc. corpus	Étendu + génotypes	Restreint + voc. corpus	Restreint + génotypes
SSS	4334 (75.74%)	4523 (77.25%)	5217 (87.61%)	5368 (89.56%)
SES	717 (12.53%)	759 (12.96%)	282 (4.74%)	263 (4.39%)
ESS	540 (9.44%)	483 (8.25%)	409 (6.87%)	327 (5.46%)
EES	131 (2.29%)	90 (1.54%)	47 (0.79%)	36 (0.60%)
Total	**5722**	**5855**	**5955**	**5994**

Tableau 0-16 Succès et erreurs parmi les deux étiquettes précédant le succès de la désambiguïsation d'un homographe.

En regardant les données du Tableau 0-16, il est étonnant de constater le grand nombre d'homographes ayant été correctement désambiguïsés malgré une erreur parmi les deux étiquettes précédentes. En effet, les cas de trigrams SES, ESS et EES représentent 24.26% (1ère colonne : 12.53% + 9.44% + 2.29%) en utilisant l'ensemble étendu d'étiquettes et 22.75% (2ème colonne : 12.96% + 8.25% + 1.54%) en utilisant les probabilités à priori du corpus et les génotypes. Dans le même ordre, les mêmes cas représentent respectivement 12.40% (3ème colonne : 4.74% + 6.87% + 0.79%) et 10.45% (4ème colonne : 4.39% + 5.46% + 0.60%) en utilisant l'ensemble d'étiquettes restreint. C'est donc dire que ces erreurs n'ont pas eu d'impact sur la désambiguïsation. Également, on remarque que ces erreurs ont un impact moins négatif en utilisant l'ensemble d'étiquettes étendu plutôt que l'ensemble restreint. On voit aussi qu'une suite de deux étiquettes correctement identifiées mène plus souvent à un succès avec l'ensemble d'étiquettes restreint que l'ensemble étendu.

Le Tableau 0-17 reprend les résultats pour les homographes n'ayant pas été correctement désambiguïsés.

Succès et erreurs parmi les étiquettes précédant un échec				
3gram	Étendu + voc. corpus	Étendu + génotypes	Restreint + voc. corpus	Restreint + génotypes
SSE	410 **(53.32%)**	385 **(60.53%)**	381 **(71.08%)**	376 **(75.65%)**
SEE	174 **(22.63%)**	130 **(20.44%)**	72 **(13.43%)**	65 **(13.08%)**
ESE	130 **(16.91%)**	87 **(13.68%)**	62 **(11.57%)**	42 **(8.45%)**
EEE	55 **(7.15%)**	34 **(5.35%)**	21 **(3.92%)**	14 **(2.82%)**
Total	**769**	**636**	**536**	**497**

Tableau 0-17 Succès et erreurs parmi les deux étiquettes précédant l'échec de la désambiguïsation d'un homographe.

En étudiant la distribution des erreurs et succès des étiquettes précédant l'échec de la désambiguïsation d'un homographe à la lumière du Tableau 0-17, on remarque d'abord que le succès parmi les deux étiquettes précédentes n'est pas garant d'une désambiguïsation réussie. En effet, dans ces conditions, en utilisant l'ensemble étendu d'étiquettes, 53.32%

des homographes ne sont pas correctement désambiguïsés en utilisant les probabilités à priori du corpus seulement et 60.53% ne le sont pas lorsque ces dernières sont complétées par celles des génotypes. En utilisant l'ensemble restreint d'étiquettes, dans le même ordre, c'est 71.08% et 75.65% des cas qui ne sont pas désambiguïsés. On remarque ainsi que l'utilisation de l'ensemble étendu d'étiquettes mène moins souvent à un échec après deux succès consécutifs qu'en utilisant l'ensemble restreint. En contrepartie, l'ensemble restreint semble moins sensible aux échecs précédents que l'ensemble étendu (SEE, ESE et EEE).

En résumé, ce qui est étonnant, c'est de constater que ce ne sont pas tous les succès qui reposent sur une désambiguïsation adéquate des étiquettes ayant servi au calcul de la probabilité de la nature de l'homographe. Il est tout aussi étonnant de constater que ce ne sont pas tous les échecs qui sont dus à une mauvaise désambiguïsation parmi les deux étiquettes précédentes. En fait, dans ce cas, un plus grand nombre d'échecs survient après le succès de la désambiguïsation des deux étiquettes précédentes que dans le cas où une seule erreur est survenue dans la désambiguïsation de ces deux étiquettes.

On ne peut donc affirmer que les erreurs de désambiguïsation sur les homographes de type verbe/substantif sont dues à d'autres erreurs de désambiguïsation parmi les mots qui précèdent. L'hypothèse n'est donc pas confirmée.

Le poids des substantifs

Des résultats aussi étonnants nous ont poussé à nous poser une autre question : « comment autant d'homographes peuvent-ils être correctement désambiguïsés malgré des erreurs dans les étiquettes précédentes ? » Nous avons émis l'hypothèse que les noms étant en plus grand nombre dans le corpus, les probabilités a priori ainsi que les ngrams avaient un biais favorable envers les noms. Si le nombre de noms dans l'ensemble d'homographes est également plus élevé, en bénéficiant de ce biais des probabilités, cela expliquerait pourquoi tant d'homographes sont correctement désambiguïsés. Ce biais devrait se refléter dans le taux de succès des homographes dont la nature est un nom. Dans leur cas, le taux de succès devrait être élevé peu importe le succès de la désambiguïsation des étiquettes précédentes tandis que les autres homographes devraient présenter un taux de succès faible lorsque il y a au moins une erreur parmi les deux étiquettes précédentes.

Pour vérifier l'hypothèse, nous avons départagé dans les données des tableaux Tableau 0-12 à Tableau 0-15, l'étiquette de l'homographe en question. Les tableaux Tableau 0-18 à Tableau 0-21 en donnent les détails.

Ensemble étendu d'étiquettes + génotypes									
	Noms	Verbes conj.	Infinitifs	Part. passés	Part. présents	Adj. qual	Adv	Prep	Autres
SSE	101 (3.38%)	63 (3.95%)	9 (2.58%)	89 (10.93%)	9 (10.00%)	81 (55.86%)	19 (6.88%)	3 (2.94%)	11 (78.57%)
SEE	57 (1.91%)	49 (3.07%)	3 (0.86%)	6 (0.74%)	0 (0.00%)	9 (6.21%)	4 (1.45%)	0 (0.00%)	2 (14.29%)
ESE	11 (0.37%)	18 (1.13%)	2 (0.57%)	19 (2.33%)	2 (2.22%)	27 (18.62%)	6 (2.17%)	1 (0.98%)	1 (7.14%)
EEE	2 (0.07%)	24 (1.51%)	0 (0.00%)	4 (0.49%)	0 (0.00%)	4 (2.76%)	0 (0.00%)	0 (0.00%)	0 (0.00%)
SSS	2045 (68.39%)	1138 (71.39%)	289 (82.81%)	570 (70.02%)	56 (62.22%)	12 (8.28%)	213 (77.17%)	83 (81.37%)	0 (0.00%)
SES	562 (18.80%)	132 (8.28%)	21 (6.02%)	13 (1.60%)	11 (12.22%)	3 (2.07%)	12 (4.35%)	5 (4.90%)	0 (0.00%)
ESS	157 (5.25%)	149 (9.35%)	24 (6.88%)	106 (13.02%)	11 (12.22%)	9 (6.21%)	18 (6.52%)	9 (8.82%)	0 (0.00%)
EES	55 (1.84%)	21 (1.32%)	1 (0.29%)	7 (0.86%)	1 (1.11%)	0 (0.00%)	4 (1.45%)	1 (0.98%)	0 (0.00%)
Total:	**2990**	**1594**	**349**	**814**	**90**	**145**	**276**	**102**	**14**

Tableau 0-18 Répartition des erreurs selon la nature de l'homographe pour l'ensemble étendu d'étiquettes et les probabilités à priori complétées des génotypes.

Ensemble étendu d'étiquettes + vocabulaire du corpus seulement									
	Nouns	Conj. verbs	Infinitivs	Past Part.	Pres. Part.	A-qual	Adverbes	Prep	Autres
SSE	106 (3.55%)	50 (3.14%)	13 (3.72%)	120 (14.74%)	17 (18.89%)	74 (28.24%)	16 (5.80%)	3 (2.94%)	11 (78.57%)
SEE	85 (2.84%)	53 (3.32%)	3 (0.86%)	9 (1.11%)	6 (6.67%)	14 (5.34%)	2 (0.72%)	0 (0.00%)	2 (14.29%)
ESE	23 (0.77%)	26 (1.63%)	4 (1.15%)	34 (4.18%)	6 (6.67%)	28 (10.69%)	7 (2.54%)	1 (0.98%)	1 (7.14%)
EEE	17 (0.57%)	24 (1.51%)	0 (0.00%)	6 (0.74%)	0 (0.00%)	8 (3.05%)	0 (0.00%)	0 (0.00%)	0 (0.00%)
SSS	1954 (65.35%)	1140 (71.52%)	275 (78.80%)	520 (63.88%)	43 (47.78%)	125 (47.71%)	206 (74.64%)	71 (69.61%)	0 (0.00%)
SES	509 (17.02%)	125 (7.84%)	23 (6.59%)	17 (2.09%)	7 (7.78%)	4 (1.53%)	19 (6.88%)	13 (12.75%)	0 (0.00%)
ESS	215 (7.19%)	147 (9.22%)	29 8.31% ()	98 (12.04%)	10 (11.11%)	7 (2.67%)	22 (7.97%)	12 (11.76%)	0 (0.00%)
EES	81 (2.71%)	29 (1.82%)	2 (0.57%)	10 (1.23%)	1 (1.11%)	2 (0.76%)	4 (1.45%)	2 (1.96%)	0 (0.00%)
Total:	**2990**	**1594**	**349**	**814**	**90**	**262**	**276**	**102**	**14**

Tableau 0-19 Répartition des erreurs selon la nature de l'homographe pour l'ensemble étendu d'étiquettes et les probabilités a priori du corpus seulement.

Ensemble restreint d'étiquettes + vocabulaire du corpus seulement						
Ngram	**Nouns**	**Verbes***	**A-qual**	**Adverbs**	**Prep**	**Others**
SSE	93 (3.11%)	171 (6.01%)	84 (32.06%)	17 (6.16%)	4 (3.92%)	7 (77.78%)
SEE	25 (0.84%)	39 (1.37%)	5 (1.91%)	1 (0.36%)	1 (0.98%)	1 (11.11%)
ESE	25 (0.84%)	20 (0.70%)	13 (4.96%)	3 (1.09%)	0 (0.00%)	1 (11.11%)
EEE	8 (0.27%)	12 (0.42%)	1 (0.38%)	0 (0.00%)	0 (0.00%)	0 (0.00%)
SSS	2493 (83.38%)	2258 (79.31%)	150 (57.25%)	237 (85.87%)	79 (77.45%)	0 (0.00%)
SES	149 (4.98%)	118 (4.14%)	2 (0.76%)	3 (1.09%)	10 (9.80%)	0 (0.00%)
ESS	179 (5.99%)	200 (7.02%)	7 (2.67%)	15 (5.43%)	8 (7.84%)	0 (0.00%)
EES	18 (0.60%)	29 (1.02%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)
Total:	**2990**	**2847**	**262**	**276**	**102**	**9**

Tableau 0-20 Répartition des erreurs selon la nature de l'homographe pour l'ensemble restreint d'étiquettes et les probabilités a priori du corpus seulement.

*incluant Conj.verbs + Infinitifs + Part. passé + Part. présent

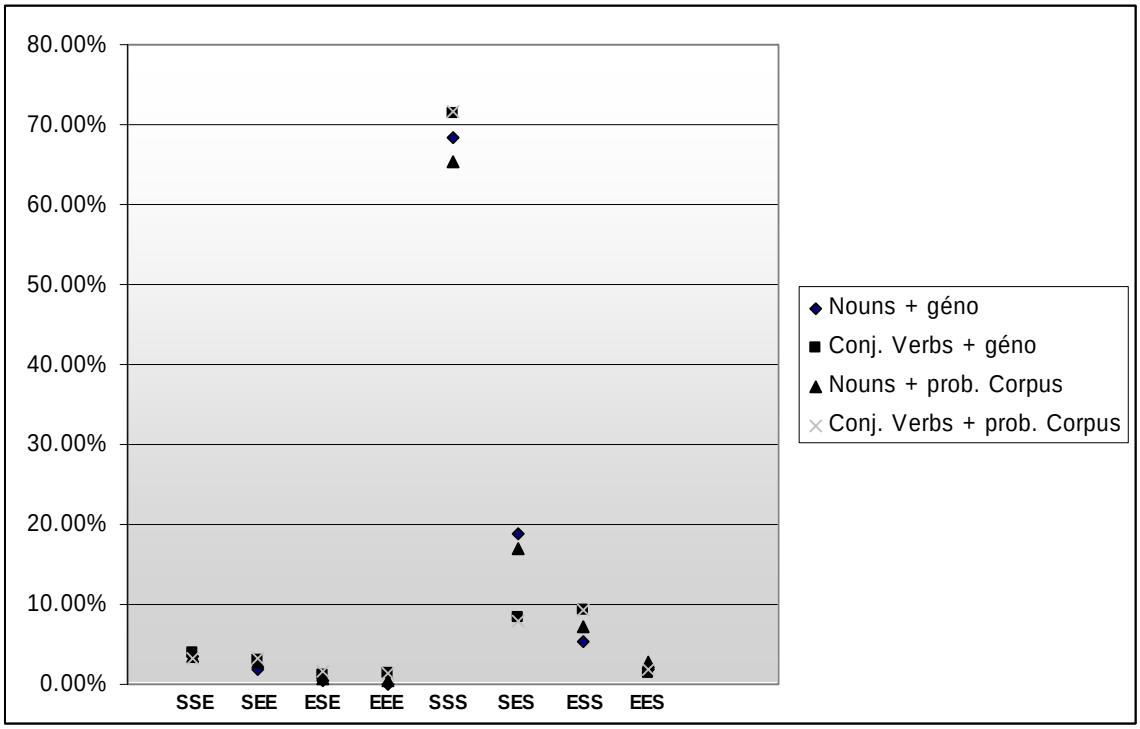
Ensemble restreint d'étiquettes + génotypes						
ngram	**Nouns**	**Verbes***	**A-qual**	**Adverbs**	**Prep**	**Others**
SSE	106 (3.55%)	147 (5.16%)	90 (34.35%)	17 (6.16%)	4 (3.92%)	7 (77.78%)
SEE	17 (0.57%)	39 (1.37%)	6 (2.29%)	2 (0.72%)	0 (0.00%)	1 (11.11%)
ESE	15 (0.50%)	14 (0.49%)	10 (3.82%)	2 (0.72%)	0 (0.00%)	1 (11.11%)
EEE	3 (0.10%)	10 (0.35%)	1 (0.38%)	0 (0.00%)	0 (0.00%)	0 (0.00%)
SSS	2573 (86.05%)	2315 (81.31%)	149 (56.87%)	242 (87.68%)	89 (87.25%)	0 (0.00%)
SES	145 (4.85%)	111 (3.90%)	1 (0.38%)	2 (0.72%)	4 (3.92%)	0 (0.00%)
ESS	121 (4.05%)	185 (6.50%)	5 (1.91%)	11 (3.99%)	5 (4.90%)	0 (0.00%)
EES	10 (0.33%)	26 (0.91%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)
Total:	**2990**	**2847**	**262**	**276**	**102**	**9**

Tableau 0-21 Répartition des erreurs selon la nature de l'homographe pour l'ensemble restreint d'étiquettes et les probabilités à priori complétées des génotypes.

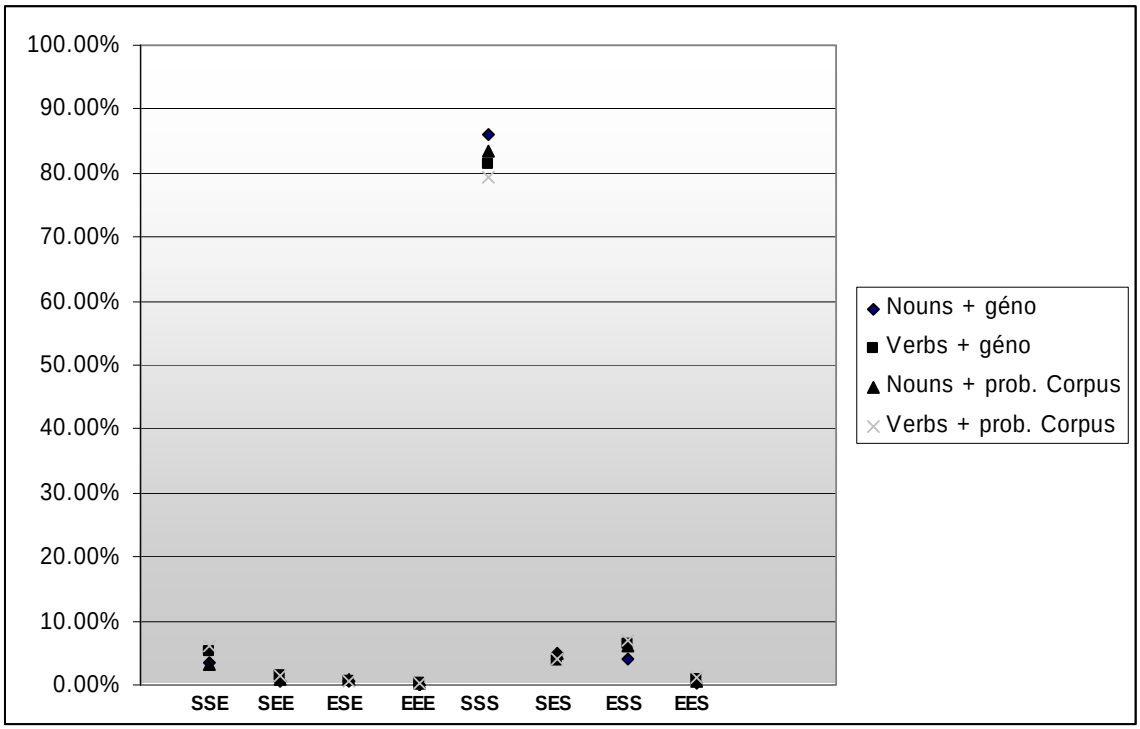
*incluant Conj.verbs + Infinitifs + Part. passé + Part. présent

En général, on remarque que toutes les instances d'homographes, peu importe leur nature grammaticale, sont correctement ou incorrectement désambiguïsées selon toutes les combinaisons de succès et d'échec parmi les étiquettes précédentes. Autrement dit, il existe des cas EEE, ESE, SEE, SSE, EES, ESS, SES et SSS pour toutes les natures d'homographes. En ce qui concerne plus particulièrement les noms et les verbes, les deux graphiques suivant illustrent plus clairement que les succès et échecs sont distribués de la

même façon pour les deux natures, que ce soit en utilisant l'ensemble étendu d'étiquettes ou l'ensemble restreint.



Graphique 0-2 Distribution des succès et échecs parmi les étiquettes des ngrams ayant servi à désambiguïser les noms et verbes homographes en utilisant l'ensemble étendu d'étiquettes.



Graphique 0-3 Distribution des succès et échecs parmi les étiquettes des ngrams ayant servi à désambiguïser les noms et verbes homographes en utilisant l'ensemble restreint d'étiquettes.

En définitive, les données indiquent que ni les noms, ni les verbes ne sont responsables des succès et insuccès étonnants parmi les trigrams ayant mené à la désambiguïsation des homographes. Cela se vérifie par le fait que la distribution des erreurs selon les échecs ou succès des étiquettes précédentes est sensiblement la même pour les noms et les verbes alors qu'elle devrait être avantageuse pour les noms et ne pas l'être pour les verbes selon l'hypothèse. Il semble plutôt que combinaisons d'échecs et de succès des trigrams soient relativement distribuées uniformément parmi toutes les catégories grammaticales d'homographes.

Apport des connaissances linguistiques

Les sections 0 et 0 on fait ressortir certaines forces et faiblesses de l'utilisation des ngrams quant à leur habileté à désambiguïser les homographes de type verbe-substantif. D'une part, il a été constaté que les structures syntaxiques présentes dans le corpus d'entraînement sont suffisamment récurrentes pour être statistiquement significatives. D'autre part, il a été

montré que les erreurs de désambiguïsation ne sont pas toutes responsables de l'échec de la désambiguïsation des homographes tandis que les succès sont parfois difficiles à expliquer. Il y a même une part de comportement inusité pour tous les types d'homographes et toutes les situations de succès et d'échec parmi les étiquettes précédant un homographe. Ces observations nous renseignent un peu plus sur le comportement des trigrams lors de la désambiguïsation. En général, les trigrams saisissent une partie de l'information syntaxique du corpus, mais d'un autre côté, l'application de ces informations à de nouveaux cas mène à des situations difficilement compréhensibles.

Quel rôle la linguistique peut-elle avoir dans l'étiquetage automatique de textes. Est-ce que des notions de linguistique peuvent améliorer ces performances? Si oui, quelles sont-elles et comment peuvent-elles être appliquées? C'est en réfléchissant à ces questions que des observations supplémentaires ont mené à de nouveaux constats et à des idées pour améliorer l'utilisation des ngrams. Les sections suivantes présentent ces observations supplémentaires, la réflexion et les idées qui les accompagnent.

Les connaissances linguistiques appropriées

Quelle est la nature des connaissances qui peuvent être utiles pour l'étiquetage automatique?

D'abord, il faut faire la part des choses entre le langage et la langue écrite. En effet, la linguistique a pour objet d'étude le langage humain. Or, ce dernier est d'abord et avant tout un système oral car bien avant l'arrivée de l'écriture, les êtres humains communiquaient entre eux. La compétence et les connaissances des êtres humains en terme de langage se situent donc au niveau de la production, de la réception et de la compréhension de la parole. Pour s'en convaincre, il suffit de penser que ce n'est que bien plus tard dans l'histoire de l'humanité que l'écriture est apparue. De même, les enfants apprennent à parler avant d'écrire sans compter que même de nos jours, certaines personnes analphabètes peuvent mener une vie entière sans lire ni écrire.

Les systèmes d'écriture essaient tant bien que mal de représenter le langage oral, mais de nombreuses informations sont perdues : les systèmes d'écriture ne représentent pas l'intonation; certains mots ont la même orthographe bien que leur signification soit

différente; ni la prononciation ni la phonologie ne sont représentées; pas plus que ne le sont les liaisons en français; etc. Il ne faudrait cependant pas croire que tous les torts vont à l'écriture, car, par exemple, dans la langue orale, il existe également des mots qui se prononcent de la même façon mais qui signifient des choses différentes. Cependant, il est très important de garder en considération que le traitement automatique du langage s'exerce sur deux types de données : des données orales et des données écrites. Les deux types de données comportent leur lot de difficultés, mais le traitement du langage écrit dont fait partie l'étiquetage automatique de texte a ceci de particulier qu'il opère sur une version partielle du système linguistique. Il doit aussi composer non seulement avec des phénomènes linguistiques, mais avec des phénomènes qui n'ont rien à voir avec ce système. Par exemple, en français, le son /o/ s'écrit de trois façons : *o*, *au* et *eau*. Si on ajoute les finales plurielles *os* et *aux*, il s'agit alors de cinq façons différentes d'écrire un seul et même son. D'autres éléments sont aussi exclusifs à l'écrit tels que certains caractères de ponctuation (parenthèses, tirets, etc.), les nombres exprimés en chiffre arabe, les formules mathématiques, les numéros de téléphone, les adresses, les changements de paragraphes, etc. Par conséquent, la notion de linguistique s'appliquant à l'étiquetage automatique est quelque peu altérée par le fait de manipuler seulement un reflet partiel et non entièrement systématique du langage.

Par conséquent, les connaissances dites « linguistiques » en question dans cette section réfèrent donc à la fois à des caractéristiques du langage ainsi qu'aux standards d'orthographe et de grammaticalité du français écrit tels que dictés par les grammaires françaises traditionnelles. Plus particulièrement dans le cas de l'étiquetage automatique, seuls le niveau lexical et le niveau syntaxique (dont le concept est restreint aux suites de mots licites en surface) ne sont accessibles.

Au début de ce projet, l'idée d'utiliser des connaissances linguistiques pour améliorer la performance des ngrams devait prendre la forme de règles supplémentaires appliquées avant ou après la désambiguïsation des ngrams pour venir corriger certaines erreurs du prototype. Le tout aurait pris la forme d'un module externe de préparation pré-prototype des données ou de correction post-prototype des étiquettes. C'est d'ailleurs le genre de règles et d'application de règles que certains étiqueteurs rapportés dans les publications

utilisent (Chanod & Tapanainen (1995a), Tzoukermann et al. (1995, 1997), Abeillé et al. (1998)).

En cherchant des erreurs du prototype pouvant être corrigées par des règles éditées manuellement, nous avons constaté que les situations où, à l'inverse, le prototype ne devait pas faire d'erreur étaient plus révélatrices de pistes de recherche. Cela a mené à de nouvelles observations dont il est question dans les paragraphes qui suivent.

Nous avons identifié quelques situations évidentes pour des locuteurs du français. Il appert que les situations que nous mentionnons font toutes appel à des notions de syntaxe dans la perspective où l'environnement immédiat d'un homographe contient des indices forts de la nature de l'homographe en question. Dans ces situations, une erreur de la part du prototype est plus difficile à pardonner étant donné l'évidence de la nature de l'homographe pour un locuteur. Pour chacune de ces situations, nous avons évalué si les ngrams devraient être en mesure d'extraire des statistiques représentant ces phénomènes ou si, dû à la nature des situations linguistiques, ils ne devraient pas être en mesure de tirer des statistiques permettant de désambiguïser efficacement ces cas. Nous avons émis l'hypothèse que si les ngrams extraient des informations de nature linguistique, le taux de succès de la désambiguïsation des homographes devrait avoisiner le taux de succès des mots de même nature non ambigus (environ 100% puisqu'ils ne sont pas ambigus) dans les situations linguistiques accessibles aux ngrams tandis que pour les autres, le taux de succès pour les homographes devrait être inférieur au taux de succès général.

Les sections suivantes présentent cinq cas d'homographes évidents pour un locuteur et mentionnent pour chacun d'eux s'il sont également avantagés ou désavantagés du point de vue du modèle des trigrams. La performance du prototype est commentée en rapport à l'hypothèse émise.

Les homographes linguistiquement avantagés

Voici les cas où nous avons jugé que les homographes ne devraient pas être difficiles à désambiguïser pour un locuteur puisqu'ils sont dans un environnement syntaxique où se trouvent des indices forts indiquant leur nature.

Tous les tableaux présentent les résultats en groupant les cas par succès ou insuccès des deux étiquettes précédentes comme c'est le cas dans les sections qui ont précédé. La lettre *E* signifie échec tandis que la lettre *S* signifie succès. Par exemple, *ES* signifie que l'étiquette antéprécédente n'a pas été désambiguïsée correctement alors que l'étiquette précédente a correctement été identifiées. Les observations ont été limitées à l'utilisation complémentaire des génotypes par le prototype et des ensembles étendus et restreints d'étiquettes.

Pour chacune de ces situations, nous avons évalué si les trigrams étaient en mesure de saisir cette information.

Les homographes ayant comme sujet un pronom personnel.

Les verbes dans une structure syntaxique débutant par un pronom personnel incluant le pronom relatif « qui », soit l'ensemble {je, tu, il, on, elle, nous, vous, ils, elles, qui} sont intéressants car le sujet réalisé par le pronom est un indice fort pour un locuteur de la nature de l'homographe. En fait, un locuteur ne se trompe pas. Syntaxiquement, l'homographe ne peut être qu'un verbe, puisque seul un verbe peut servir de support à un sujet.

De plus, comme le montrent les grammaires locales illustrées par les figures 1 et 2 de la section 0, les mots pouvant s'interposer entre le sujet exprimé par un pronom personnel et le verbe forment un ensemble limité : {ne, n, me, te, se, nous, vous, le, la, les, m, t, l, s, lui, leur, y, en}. Cela veut donc dire que lorsqu'un homographe est précédé par un pronom personnel et qu'il n'est séparé de celui-ci que par des mots de l'ensemble mentionné ci-dessus, il ne peut s'agir que d'un verbe.

Cette situation est un avantage également pour le prototype car tous les indices pertinents sont présents dans les deux mots qui précèdent l'homographe et peuvent ainsi être saisis par les trigrams.

Hypothèse : Les verbes homographes dont le sujet est un pronom personnel devraient présenter un taux de réussite élevé avoisinant celui des verbes qui ne sont pas homographes de type verbe/substantif.

Le Tableau 0-22 suivant fait l'inventaire des succès et échecs du prototype quant aux homographes dans une telle situation. Ces résultats sont comparés au taux de désambiguïsation des verbes qui ne sont pas homographes de type verbe-substantif[1].

		Ensemble restreint				Ensemble étendu			
		Succès	%	Échecs	%	Succès	%	Échecs	%
Homographe verbe/substantif	EE	1	0.31%	0	0.00%	3	0.93%	1	0.31%
	SE	3	0.93%	1	0.31%	39	12.04%	2	0.62%
	ES	27	8.33%	1	0.31%	35	10.80%	3	0.93%
	SS	279	86.11%	12	3.70%	231	71.30%	10	3.09%
Total:		**310**	**95.68%**	**14**	**4.32%**	**308**	**95.06%**	**16**	**4.94%**
Autres	EE	1	0.11%	0	0.00%	26	2.86%	0	0.00%
	SE	26	2.86%	0	0.00%	175	19.27%	2	0.22%
	ES	94	10.35%	0	0.00%	106	11.67%	0	0.00%
	SS	784	86.34%	3	0.33%	593	65.31%	6	0.66%
Total:		**905**	**99.67%**	**3**	**0.33%**	**900**	**99.12%**	**8**	**0.88%**

Tableau 0-22 Résultats de l'hypothèse 5.3.3.1.

On remarque d'abord que l'utilisation de l'ensemble étendu et restreint donne des résultats presque identiques. On aurait pu s'attendre à ce que l'utilisation de l'ensemble étendu mène à un taux de désambiguïsation moins élevé à cause d'erreurs dans le mode ou le temps (par exemple, présent de l'indicatif au lieu du subjonctif ou inversement), mais cela n'est pas le cas car le corpus contient peu d'ambiguïté entre les modes et les temps.

Par ailleurs, en lien directement avec l'hypothèse, on remarque que 95.68% et 95.06% des homographes ne sont pas désambiguïsés correctement en utilisant respectivement l'ensemble restreint et étendu d'étiquettes. Malgré l'assurance que peut fournir l'environnement syntaxique, le prototype ne désambiguïse pas mieux ces homographes.

Ce qui est étonnant avec ces résultats, ce sont les cas d'échecs malgré le succès de la désambiguïsation des deux étiquettes précédentes (*SSE*). En effet, dans une telle situation, deux scénarios sont possibles :

Si le mot qui précède immédiatement est un pronom personnel, alors il devrait être impossible de trouver autre chose qu'un verbe conjugué à la suite de ce pronom;

[1] Les résultats pour ces mots ne sont pas de 100% car certains d'entre eux sont parfois ambigus, cependant, sans que leur nature ne puisse être substantive. Par exemple, un homographe de type verbe/adjectif comme *abasourdis*.

Si le mot qui précède n'est pas un pronom personnel, alors les mots qui précèdent sont tout de même d'un ensemble restreint et non ambigus ce qui devrait donner des trigrams forts et conduire à un taux de désambiguïsation élevé.

En regardant ces cas de plus près, il a été constaté que certaines informations linguistiques contenues par les dictionnaires sont exactes, mais insuffisantes pour conduire à une désambiguïsation adéquate par l'utilisation des trigrams. Aucune erreur n'est liée directement aux pronoms personnels. Elles sont dues au pronom *qui* et à la négation *ne*.

D'abord, le pronom « qui » est ambigu contrairement aux autres pronoms personnels. Selon le dictionnaire et le corpus utilisés par le prototype, il peut être un pronom relatif pourvu d'un genre et d'un nombre ainsi qu'un pronom interrogatif également pourvu d'un genre et d'un nombre, ce qui est exact du point de vue de la grammaire traditionnelle. Par contre, cela est insuffisant pour le prototype. En effet, certaines erreurs sont dues à la confusion entre le pronom relatif et le pronom interrogatif. Par exemple, l'homographe *échange* mal désambiguïsé suivant, accompagné des deux mots avec lesquels il forme un trigram:

```
 [...]
.                       PONCT-S
Qui                     PRO-inter-ms
échange         N-C-ms  (au lieu de V-P3s)
 [...]
?                       PONCT-S
```

Dans cette question, les étiquettes précédentes sont correctement désambiguïsées : Ponctuation forte (PONCT-S) suivie d'un pronom interrogatif masculin singulier (PRO-inter-ms). La confusion vient du fait que les nombreux cas du pronom interrogatif *quel* (PRO-inter-ms) suivi d'un substantif masculin singulier comme dans les expressions *Quel dommage!, Quel culot!* ont donné un trigram statistiquement fréquent : PONCT-S + PRO-inter-ms + N-C-ms. Combiné avec la probabilité à priori la plus fréquente pour *échange*, soit nom masculin singulier (N-C-ms), le résultat est que l'étiquette la plus probable de l'homographe a été évaluée comme étant celle d'un nom masculin singulier plutôt qu'un verbe conjugué au présent de l'indicatif, troisième personne du singulier même si, dans ce cas, seule l'option du verbe conjugué était valable.

Les autres cas d'erreurs sont dus à la négation *ne*. Syntaxiquement, la négation ne peut pas précéder autre chose qu'un verbe conjugué. Aucun homographe ne devrait alors être étiqueté comme un nom. Pourtant, cela est le cas pour certains homographes selon le prototype. En investiguant, la raison apparaît fort simple : le dictionnaire et le corpus identifient *ne* comme un adverbe (ADV). Cela n'est pas faux, mais c'est insuffisant pour le prototype car de nombreux adverbes précèdent des noms dans le corpus. Par exemple :

```
[...]
Jean-Marie          N-P-ms
Cavada              N-P-ms
,                   PONCT-W
alors               ADV
directeur           N-C-ms
d'                  P
antenne             N-C-fs
de                  P
FR3                 N-P-ms
[...]
```

Le fait de considérer *ne* comme un simple adverbe conduit le prototype à identifier les homographes suivants comme des noms alors que même sans contexte, il est évident pour un locuteur qu'il s'agit de verbes :

ne cadre

ne change

ne coupe

Dans les deux types d'erreurs mentionnés, les informations fournies par le dictionnaire ne sont pas mauvaises en soit du point de vue de la grammaire traditionnelle, mais il leur manque des précisions linguistiques pour que les trigrams reflètent les restrictions syntaxiques à la source des observations rapportées par le Tableau 0-22. Dans le premier cas, il suffirait d'ajouter à l'étiquette du pronom interrogatif la précision qu'il s'agit également d'un pronom sujet. En effet, cette caractéristique est exclusive au pronom interrogatif *qui* et n'est par conséquent pas partagée par le pronom interrogatif *quel*. Dans le deuxième cas, le simple fait de distinguer la négation *ne* des autres adverbes par une

étiquette exclusive, comme par exemple NEG (pour *négation*), aurait aussi permis d'éviter les erreurs.

Ces deux cas illustrent la sensibilité des trigrams quant à la nature des informations mises à leur disposition. Dans ces cas, les trigrams ne sont pas en soi la cause des erreurs, mais plutôt, ce sont de mauvais choix d'étiquettes qui en sont à l'origine. Ces choix peuvent être éclairés par des connaissances linguistiques.

Les verbes immédiatement précédés d'un nom propre comme sujet.
Un verbe précédé d'un nom propre est plutôt évident à désambiguïser pour un locuteur. En effet, il est syntaxiquement impossible de faire suivre un nom propre d'un nom commun.

Jean ferme la porte.

*Jean ferme agricole.

De son côté, le prototype ne peut utiliser que les noms propres du corpus d'entraînement. Si les noms propres sont inconnus, c'est-à-dire ne font pas partie du dictionnaire, alors ils héritent de toutes les étiquettes possibles pour le calcul du trigram. Il se peut que la désambiguïsation soit correcte, mais les probabilités sont fortes qu'elle échoue.

Hypothèse : Le taux de succès de la désambiguïsation des verbes homographes immédiatement précédés d'un nom propre comme sujet devrait être faible.

Les résultats du prototype dans cette situation sont présentés au Tableau 0-23 suivant.

		Ensemble restreint d'étiquettes				Ensemble complet d'étiquettes			
		Succès	%	Échecs	%	Succès	%	Échecs	%
Homographe verbe/substantif	EE	15	8.93%	3	1.79%	7	4.17%	13	7.74%
	SE	48	28.57%	15	8.93%	37	22.02%	26	15.48%
	ES	5	2.98%	0	0.00%	7	4.17%	0	0.00%
	SS	80	47.62%	2	1.19%	75	44.64%	3	1.79%
Total:		**148**	**88.10%**	**20**	**11.90%**	**126**	**75.00%**	**42**	**25.00%**
Autres	EE	22	8.53%	0	0.00%	29	11.24%	0	0.00%
	SE	109	42.25%	5	1.94%	109	42.25%	6	2.33%
	ES	7	2.71%	0	0.00%	8	3.10%	0	0.00%
	SS	115	44.57%	0	0.00%	106	41.09%	0	0.00%
Total:		**253**	**98.06%**	**5**	**1.94%**	**252**	**97.67%**	**6**	**2.33%**

Tableau 0-23 Résultats pour l'hypothèse 5.3.3.2.

Comme le prédit l'hypothèse, le taux de succès est inférieur au taux de succès général du prototype. L'utilisation de l'ensemble restreint mène à un taux de succès de 88.10% alors que l'ensemble étendu n'obtient que 75.00%. Il est étonnant de constater que les cas les plus fréquents sont ceux où les deux étiquettes précédentes ont été correctement désambiguïsées (*SS*) et ceux où l'antéprécédente a été correctement désambiguïsée et que l'étiquette précédente, soit celle du nom propre, ne l'a pas été (*SE*). C'est cette dernière situation qui compte pour la majorité des échecs (75.00% pour l'ensemble restreint et 61.90% pour l'ensemble étendu). Si l'on regarde tous les cas où l'étiquette précédente n'a pas été correctement désambiguïsée, ils constituent 90.00% des cas d'erreurs avec l'ensemble restreint et 92.96% pour l'ensemble étendu. Cependant, la non désambiguïsation de l'étiquette précédente permet tout de même de désambiguïser correctement 42.57% et 34.92% respectivement en utilisant l'ensemble restreint et étendu. Il semble donc que les performances soient, dans ce cas, affectées négativement par un échec de la reconnaissance du nom propre, quoique cette lacune ne mène pas nécessairement à un échec car malgré tout, dans les cas où les étiquettes précédentes sont correctement identifiées, le prototype se trompe 2 fois en utilisant l'ensemble d'étiquettes restreint et 3 fois avec l'ensemble étendu.

Dans ce cas aussi, le prototype devrait être avantagé car l'information pertinente fait partie des deux mots qui précèdent. Le prototype est cependant désavantagé par le fait que son dictionnaire ne contient pas de nom propre. Il est donc plus probable qu'il fasse des erreurs parmi les étiquettes qui précèdent l'homographe en question réduisant ainsi ses chances de

bien désambiguïser l'homographe. Malgré tout, il semble que le succès des deux étiquettes précédentes ne soit pas garant du succès de la désambiguïsation comme on peut le voir dans le Tableau 0-23.

Dans ce cas, une solution envisageable serait d'intégrer, à un moment ou un autre du processus d'étiquetage, une méthode de reconnaissance des noms propres pour augmenter la reconnaissance de ces derniers en souhaitant que cela ait un impact positif sur la désambiguïsation des homographes.

Les noms immédiatement précédés d'un déterminant.

Hypothèse : Le déterminant devant un homographe devrait être un indice fort indiquant un substantif plutôt qu'un verbe et cela devrait se refléter par un taux de désambiguïsation élevé.

Le taux de désambiguïsation des homographes dans une telle position est détaillé dans le Tableau 0-24 qui suit.

		Ensemble restreint d'étiquettes				Ensemble complet d'étiquettes			
		Succès	%	Échecs	%	Succès	%	Échecs	%
Homographe verbe/substantif	EE	7	0.38%	3	0.16%	19	1.04%	2	0.11%
	SE	63	3.45%	0	0.00%	375	20.54%	36	1.97%
	ES	56	3.07%	7	0.38%	52	2.85%	2	0.11%
	SS	1631	89.32%	59	3.23%	1295	70.92%	45	2.46%
Total:		**1757**	**96.22%**	**69**	**3.78%**	**1741**	**95.35%**	**85**	**4.65%**
Autres	EE	51	0.69%	4	0.05%	103	1.39%	6	0.08%
	SE	442	5.95%	10	0.13%	1593	21.45%	62	0.83%
	ES	241	3.24%	5	0.07%	221	2.98%	7	0.09%
	SS	6556	88.27%	118	1.59%	5363	72.21%	72	0.97%
Total:		**7290**	**98.16%**	**137**	**1.84%**	**7280**	**98.02%**	**147**	**1.98%**

Tableau 0-24 Résultats pour l'hypothèse 5.3.3.3.

Ce Tableau 0-24 montre que le taux de désambiguïsation des homographes n'atteint pas celui du reste du vocabulaire dans les mêmes circonstances. Il est intrigant toutefois de constater que lorsque les deux étiquettes précédentes ont été correctement désambiguïsées en utilisant l'ensemble restreint d'étiquettes, le taux de succès de désambiguïsation des homographes (89.32%) est plus élevé que celui du reste du vocabulaire (88.27%).

Cependant, les erreurs dans les mêmes conditions sont aussi plus élevées du côté des homographes (3.23%) que de celui du reste du vocabulaire (1.59%).

On remarque également que l'écart entre le taux de succès de la désambiguïsation des homographes et du reste du vocabulaire est moins élevé que dans les cas précédents (c.f. sections 0 et 0).

En regardant de plus près les échecs de la désambiguïsation lorsque les deux étiquettes précédentes sont correctement désambiguïsées, on remarque que, comme le prédit notre hypothèse, aucun homographe n'est confondu avec un verbe conjugué. Par contre, certains des homographes ont été reconnus, à tort, comme des adjectifs lorsque cette étiquette fait partie de leur génotype. Par exemple, les homographes suivants précédés des quatre unités lexicales de leur contexte gauche:

Homographe et contexte gauche	Étiquette du prototype	Étiquette attendue
[...] été opposée par le carré [...]	A-qual-ms	N-C-ms
[...] toujours se trouver un toqué [...]	A-qual-ms	N-C-ms
[...]. Non , ces allumés [...]	A-qual-mp	N-C-mp
[...] sa force . Ces allumés	A-qual-mp	N-C-mp
[...] f) est le double [...]	A-qual-ms	N-C-ms
[...] ces congrès , ces assises [...]	A-qual-fp	N-C-fp
[...] propres mythes . Les Anglais [...]	A-qual-mp	N-C-mp
[...] la majorité . Les expatriés [...]	A-qual-mp	N-C-mp
[...] l' exception d' un extraverti [...]	A-qual-ms	N-C-ms

Tableau 0-25 Exemples d'erreurs commises par le prototype.

Malgré des indices forts, les homographes sont tout de même mal désambiguïsés dans des circonstances évidentes pour un locuteur.

Dans le cas des homographes reconnus comme des adjectifs, cela est dû au fait que des adjectifs peuvent suivre un déterminant et précéder un nom. Par exemple : « Les beaux avions ». Dans ce cas, la combinaison des trigrams et de la probabilité à priori peut mener à des erreurs dans le calcul de la nature d'un homographe. Dans ces cas, la séquence est valide syntaxiquement. Il faudrait que le prototype puisse consulter le contexte droit pour éclairer sa décision. Aussi, sachant que ce ne sont pas tous les adjectifs qui peuvent

précéder un nom (*La rouge automobile. *Une rapide automobile.), cette information (peut précéder ou non un substantif) pourrait être encodée dans les étiquettes.

Les verbes immédiatement suivis d'un complément d'objet direct introduit par un déterminant.

Il est en effet difficile de faire suivre un nom d'un déterminant alors que dans le cas d'un verbe, cela introduit un complément d'objet direct.

*La ferme la porte.

Il ferme la porte.

Cet indice devrait être suffisant pour indiquer un verbe plutôt qu'un nom. Le Tableau 0-26 suivant présente les détails du taux de réussite dans de telles circonstances. Bien sûr, les trigrams ne contiennent d'informations que sur ce qui précèdent et le prototype ne regarde pas l'environnement qui suit un homographe lors de la désambiguïsation. Par conséquent, le prototype ne peut bénéficier de ces informations.

Hypothèse : Le pourcentage de désambiguïsation des homographes suivis d'un complément d'objet direct devrait être inférieur au taux général du prototype.

		Ensemble restreint d'étiquettes				Ensemble complet d'étiquettes			
		Succès	%	Échecs	%	Succès	%	Échecs	%
Homographe verbe/substantif	EE	2	0.65%	4	1.31%	2	0.65%	5	1.63%
	SE	12	3.92%	0	0.00%	18	5.88%	4	1.31%
	ES	14	4.58%	2	0.65%	30	9.80%	1	0.33%
	SS	247	80.72%	25	8.17%	221	72.22%	25	8.17%
Total:		**275**	**89.87%**	**31**	**10.13%**	**271**	**88.56%**	**35**	**11.44%**
Autres	EE	21	1.52%	0	0.00%	40	2.89%	0	0.00%
	SE	80	5.78%	2	0.14%	134	9.68%	4	0.29%
	ES	92	6.64%	2	0.14%	136	9.82%	3	0.22%
	SS	1173	84.69%	15	1.08%	1051	75.88%	17	1.23%
Total:		**1366**	**98.63%**	**19**	**1.37%**	**1361**	**98.27%**	**24**	**1.73%**

Tableau 0-26 Résultats pour l'hypothèse 5.3.3.4.

On voit que le taux de désambiguïsation des homographes n'atteint pas celui du reste du vocabulaire dans le même environnement syntaxique. Encore une fois, malgré des indices forts, le prototype produit des erreurs dans un contexte pourtant évident pour un locuteur.

Dans ce cas tout comme dans le précédent, le modèle des ngrams n'a aucune façon d'accéder à l'information concernant les mots qui suivent. Le modèle entier repose sur les étiquettes déjà identifiées. Dans le cas des trigrams, il s'agit des deux étiquettes précédentes. Le modèle des ngrams dans ce cas, ne tire pas partie de toutes les informations disponibles pour sélectionner l'étiquette de l'homographe.

Les verbes précédés d'un syntagme sujet du type DET + NOM.

Hypothèse : Le fait qu'un homographe soit précédé d'un déterminant suivi d'un substantif devrait fortement favoriser la reconnaissance d'un verbe au détriment d'un nom.

Le Tableau 0-27 suivant montre les détails de la désambiguïsation des homographes dans cette situation.

		Ensemble restreint d'étiquettes				Ensemble complet d'étiquettes			
		Succès	%	Échecs	%	Succès	%	Échecs	%
Homographe verbe/substantif	EE	0	0.00%	0	0.00%	3	0.83%	1	0.28%
	SE	13	3.58%	2	0.55%	13	3.58%	3	0.83%
	ES	9	2.48%	2	0.55%	43	11.85%	5	1.38%
	SS	310	85.40%	27	7.44%	267	73.55%	28	7.71%
Total:		**332**	**91.46%**	**31**	**8.54%**	**326**	**89.81%**	**37**	**10.19%**
Autres	EE	2	0.29%	0	0.00%	22	3.24%	0	0.00%
	SE	49	7.23%	1	0.15%	44	6.49%	3	0.44%
	ES	36	5.31%	2	0.29%	135	19.91%	8	1.18%
	SS	572	84.37%	16	2.36%	445	65.63%	21	3.10%
Total:		**659**	**97.20%**	**19**	**2.80%**	**646**	**95.28%**	**32**	**4.72%**

Tableau 0-27 Résultats pour l'hypothèse 5.3.3.5.

Le Tableau 0-27 montre que malgré un environnement favorisant l'hypothèse d'un substantif, le prototype n'atteint pas le niveau de désambiguïsation du reste du vocabulaire qui ne peut être confondu avec un verbe.

Encore une fois, dans un contexte plutôt évident, le prototype commet des erreurs difficilement compréhensibles du point de vue d'un locuteur. En y regardant de plus près

par contre, ce qui est évident pour un locuteur l'est moins pour un prototype utilisant des trigrams. Le prototype confond les verbes et les noms dont voici cinq exemples :

Homographe et contexte gauche	Étiquette du prototype	Étiquette attendue
[…] Mais le magazine conserve […]	N-C-fs	V-P3s
[…] La passion aide […]	N-C-fs	V-P3s
[…] Mais le temps presse […]	N-C-fs	V-P3s
[…] et la firme figure […]	N-C-fs	V-P3s
[…] Or le temps presse […]	N-C-fs	V-P3s

Tableau 0-28 Exemples d'homographes en position de verbe conjugué confondus avec des substantifs.

Le fait que la suite DÉTERMINANT + NOM + NOM soit autorisée dans le corpus explique ce cas. En effet, on retrouve des suites comme :

```
1. le profil type            D-def-ms   + N-C-ms + N-C-ms
2. Une carte maitresse       D-indef-fs + N-C-fs + N-C-fs
3. l' épargne retraite       D-def-fs   + N-C-fs + N-C-fs
4. l' ingénieur conseil      D-def-ms   + N-C-ms + N-C-ms
5. le lendemain matin        D-def-ms   + N-C-ms + N-C-ms
```

Lorsque la probabilité a priori d'un homographe avantage fortement l'étiquette du nom, alors le trigram est débalancé par cette forte tendance et la nature la plus probable alors, selon les calculs du prototype, est celle du nom.

Le fait d'ajouter aux étiquettes l'information selon laquelle un nom qui suit un autre nom est attribut du premier semble une avenue possible.

L'autre type d'erreur est dû à la confusion avec les adjectifs. En effet, les homographes dont le génotype contient aussi la nature adjectivale sont parfois reconnus comme tel au détriment de la véritable nature, soit celle d'un verbe conjugué. Cela se comprend facilement car, en français, les adjectifs sont post-posés aux substantifs dans la grande majorité des cas (contrairement à l'anglais par exemple, où les adjectifs sont antéposés au noms : *the blue car*). Par conséquent, le biais est très favorable aux adjectifs dans ces cas et le prototype reconnaît souvent un adjectif au lieu d'un verbe. Cette situation aurait dû être

prévue par l'hypothèse et par conséquent, cette dernière est plutôt faible par rapport aux autres.

Dans ce cas aussi, le prototype aurait avantage à consulter le contexte droit pour vérifier et modifier si nécessaire sa décision.

L'utilité de la linguistique dans l'étiquetage automatique de texte

Les résultats exposés au cours des sections qui précèdent permettent de constater que les situations les plus utiles pour mettre à jour le comportement des trigrams sont celles où les deux étiquettes précédentes ont correctement été désambiguïsées car elles permettent d'émettre des hypothèses précises et faciles à valider. Les autres cas, c'est-à-dire ceux où une ou deux erreurs figurent parmi les étiquettes précédentes n'ont pas fait l'objet d'observations dans le cadre de ce travail. Néanmoins, elles ont sûrement leur utilité dans la description du comportement du prototype, mais elles sont probablement plus difficiles à analyser et expliquer.

Les cinq cas énumérés dans les sections 0 à 0 ont permis de cibler trois avenues d'amélioration au prototype :

1. Ajouter/modifier des informations lexicales dans les dictionnaires et le corpus.
2. Ajouter un module de reconnaissance des noms propres.
3. Permettre la consultation du contexte droit.

Ces suggestions n'ont pas été implémentées et testées pour vérifier si elles améliorent les résultats, mais elles proviennent d'observations basées sur des contextes syntaxiques qui ont révélé certains aspects du comportement du prototype et qui laissent croire qu'ils pourraient être corrigés de la sorte.

Ces trois avenues sont discutées plus en détails dans les sections qui suivent.

Ajout et modification de propriétés lexicales dans les dictionnaires et le corpus

L'un des problèmes soulevés par l'observation des contextes syntaxiques de ce chapitre est le manque de précision de l'information lexicale contenue dans le corpus et le dictionnaire.

Nous avons supposé que l'ajout de certaines propriétés aux mots pouvait corriger le type de problème observé en 5.3.3.1. Cependant, il serait nécessaire de pousser le concept d'ajout de propriétés à tout le vocabulaire selon les connaissances actuelles de la linguistique. Le fait que certaines étiquettes ne soient pas assez détaillées conduit à considérer que des mots ayant un comportement syntaxique différent soient de même nature. Cette supposition a des répercussions importantes dans le cas d'un étiqueteur dont le modèle ne considère que les relations syntaxiques entre les mots : le comportement syntaxique observé par l'entremise de ngrams peut être appliqué à des mots dont une certaine position syntaxique est interdite. Le fait que deux mots partagent la même nature suppose qu'ils sont interchangeables dans un même contexte (Dubois (1969 :9-10), Dubois et al. (1994 :156)). Le fait de pouvoir faire commuter deux mots dans le même environnement suppose qu'ils partagent les mêmes caractéristiques. Autrement, quelque chose diffère dans leur structure intrinsèque qui empêche leur insertion dans un même contexte.

Nous croyons que plus il y aura d'information sur les caractéristiques intrinsèques du lexique, plus les données apprises automatiquement seront précises. Beaucoup de caractéristiques pourraient être ajoutées. Cependant, il doit être possible d'identifier lesquelles sont pertinentes parmi toutes celles qui pourraient être ajoutées.

Le continuum ci-dessous montre où se situe cet ajout de caractéristiques. À l'une des extrémités, représentée par le rectangle 1, un ensemble d'étiquettes minimaliste, où trois étiquettes sont possibles (NOM, VERBE, AUTRE). À l'opposé, le rectangle 2 illustre le maximum d'informations auxquelles les ngrams pourraient avoir accès : le mot complet. En effet, en utilisant le tout, c'est-à-dire le mot complet, on se trouve à pouvoir manipuler toutes les caractéristiques intrinsèques disponibles, sans avoir à les identifier.

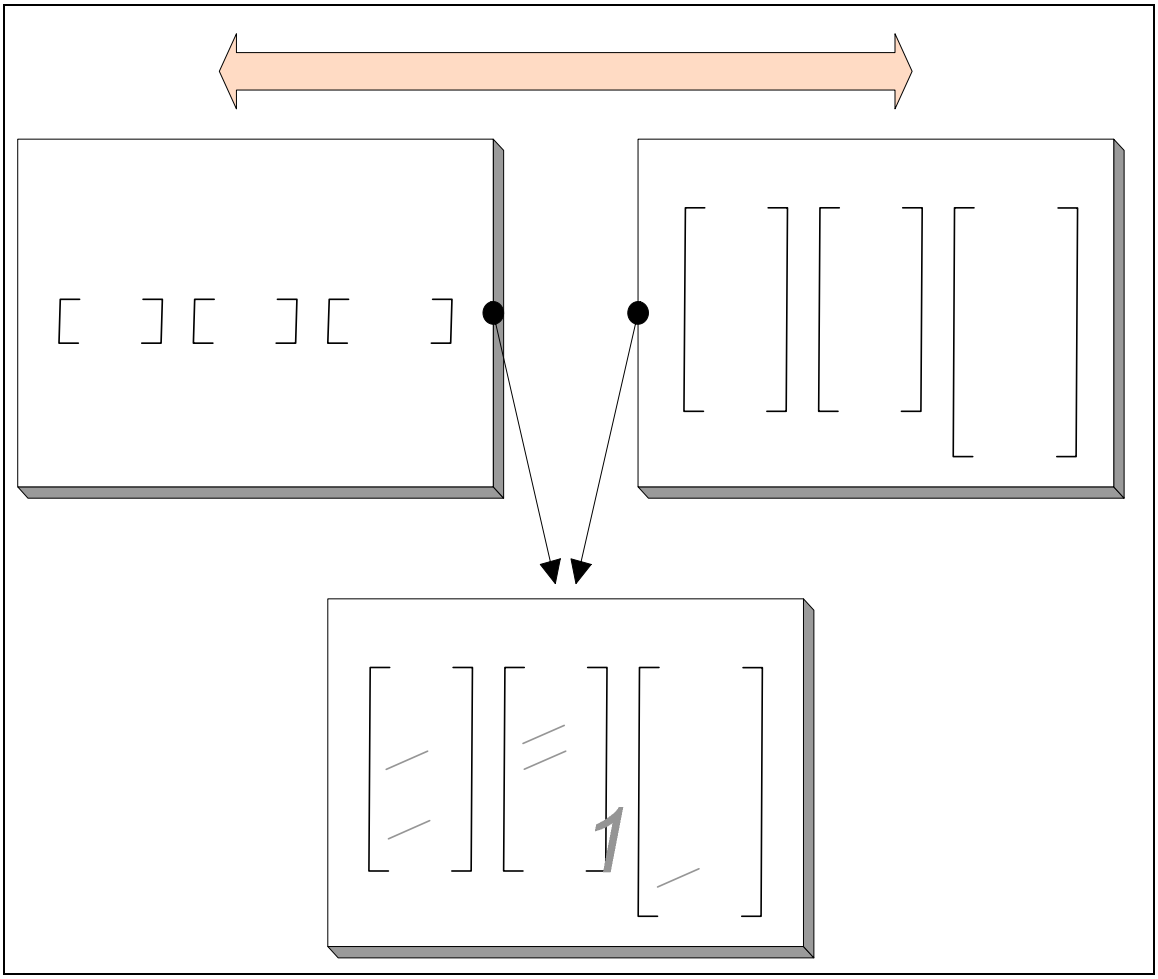


Schéma 0-1 Continuum des informations représentées par les étiquettes.

Ces situations extrêmes ne sont pas réalistes, mais elles illustrent par l'absurde où se situent les compétences linguistiques des prototypes pouvant être extraites des corpus actuels et quel est la position qui doit être visée, à notre avis, par les recherches.

Si, comme le représente le rectangle 1 par exemple, les corpus ne contenaient que trois étiquettes (NOM, VERBE et AUTRE), le modèle linguistique extrait par des ngrams aurait un pouvoir de désambiguïsation faible. En effet, il est fort probable que de nombreux homographes ne seraient pas correctement désambiguïsés vu le peu de ngrams possibles et leur trop grande généralité. On peut également supposer que de nombreux succès de la désambiguïsation seraient probablement dus au hasard. De toute façon, l'attribution au hasard d'une étiquette aux mots ambigus réduirait l'ambiguïté de 33% étant donné que

seuls trois choix sont disponibles (pour chaque homographe, il n'y a qu'une chance sur trois d'attribuer la bonne étiquette)[2]. Cette situation illustre le fait que le manque de caractéristiques lexicales ne mène pas à l'utilisation d'un modèle linguistiquement adéquat.

En contrepartie, tel que représenté par le rectangle 2 du schéma, s'il était possible d'utiliser des ngrams constitués de mots entiers, les trigrams contiendraient l'ensemble des caractéristiques disponibles pour la désambiguïsation et ce, sans avoir à les énumérer. Il est en effet possible d'imaginer ce que serait la désambiguïsation si, dans le cas des trigrams, les deux mots précédents étaient utilisés pour désambiguïser le mot suivant. Dans l'exemple illustré par le schéma, le trigram en question se traduirait par la probabilité de retrouver un verbe au présent de l'indicatif, à la troisième personne du singulier précédé de *le* et *chat*. Dans ce cas, *le* et *chat* seraient liés par tous les traits lexicaux possibles autorisant leur séquence et la prédiction de la séquence qu'ils forment en serait d'autant plus précise qu'aucune méprise sur leur nature ne serait possible. Une telle situation est cependant irréaliste car il faudrait disposer de quantités déraisonnablement volumineuses de textes étiquetés pour obtenir un nombre satisfaisant de cas pour en tirer des statistiques fiables.

Ces situations extrêmes existent en théorie. En pratique par contre, la situation des prototypes se situe quelque part entre les deux. Les prototypes présentés dans ce travail se trouvent vers la gauche du continuum, ceux utilisant l'ensemble d'étiquettes restreint étant plus à la gauche que ceux qui utilisent l'ensemble étendu d'étiquettes. Actuellement, l'ensemble étendu d'étiquettes est celui qui indique le maximum d'information lexicale (catégorie grammaticale, sous-catégorie, genre, nombre et personne) et parfois syntaxique comme dans le cas des clitiques dont le fait d'être sujet ou objet, lorsque cela s'applique, est indiqué par l'étiquette. La section 0 de ce travail montre que le taux de désambiguïsation est plus élevé en utilisant l'ensemble d'étiquettes étendu qu'en utilisant l'ensemble restreint. Par conséquent, nous croyons que l'ajout d'informations supplémentaires aux étiquettes utilisées par le corpus, le dictionnaire et le prototype pourrait permettre de s'approcher de la droite du continuum et ce, sans modifier le modèle

[2] Ces résultats sont meilleurs que ceux des prototypes présentés dans ce mémoire. Cependant, l'étiquette AUTRE étant intrinsèquement ambiguë, il n'y aurait pas vraiment de désambiguïsation. Cette situation farfelue n'est considérée que pour servir d'exemple.

d'apprentissage statistique, les ngrams. Nous croyons que pour l'instant, les trigrams n'atteignent pas le maximum de leur efficacité car il existe encore des incohérences dans les données qui leur sont fournies. Le modèle à atteindre, tel qu'illustré par le rectangle 3 du schéma, serait en quelque sorte l'ensemble des propriétés dont seulement celles utiles à l'étiquetage ne seraient retenues. Nous croyons donc que des recherches devraient être entreprises pour la découverte de ces propriétés.

Ceci dit, il est évident que le nombre de propriétés intrinsèques des mots pouvant être ajoutées aux étiquettes est important et départager les propriétés ayant un impact sur la désambiguïsation de celles qui n'en ont pas demanderait une somme importante de travail. Pour y parvenir, le type de situations évidentes pour les locuteurs discutées dans les sections 0 à 0 sont de bons outils pour tester l'ajout d'information aux étiquettes. Ce ne sont pas les seules situations possibles concernant les homographes de type verbe/substantif et il est fort probable d'en trouver pour tous les types d'ambiguïté. Faire l'inventaire de ces contextes syntaxiques pourrait également faire l'objet de futures recherches.

Il faudrait cependant faire attention de ne pas ajouter d'information dont le seul avantage est de contribuer à la réussite du prototype. Les informations devraient être appuyées par des évidences linguistiques et concorder avec le modèle linguistique de locuteurs natifs. Également, il sera intéressant d'observer et évaluer la pertinence de faire côtoyer des propriétés d'ordre syntaxique, sémantique, phonologique, etc. si jamais les propriétés ajoutées n'étaient pas de même palier linguistique.

Enfin, l'inconvénient d'ajouter de nouvelles étiquettes sera de disposer de corpus d'apprentissage plus volumineux pour permettre l'observation d'un nombre suffisant de cas, ce qui ne règle pas le problème des coûts reliés à l'étiquetage manuel de corpus. Cependant, ces coûts seront justifiés si de nouveaux corpus permettent de nouvelles percées dans le domaine de l'étiquetage automatique et d'autres domaines, notamment en linguistique où les corpus sont utiles à de nombreuses recherches.

Ajout d'un module de reconnaissance des noms propres

Le problème causé par les noms propres observés dans ce travail concerne en fait l'absence de noms propres dans le dictionnaire (outre ceux provenant du corpus d'entraînement).

Dans ce cas, les noms propres sont considérés comme des mots nouveaux et la technique utilisée par le prototype est de leur attribuer par défaut, toutes les étiquettes possibles. Par conséquent, la possibilité d'erreur est élevée. L'une des solutions serait d'ajouter au dictionnaire une liste de noms propres. Cependant, il est fort probable qu'une liste exhaustive soit difficile, voire impossible, à obtenir. En effet, les noms propres incluent non seulement des noms de personnes, mais aussi des noms de lieux, d'organisations, de produits, etc. Il serait alors avantageux de recourir à d'autres procédés pour reconnaître les noms propres.

La reconnaissance des noms propres ou entités nommées (named entities) est un problème complexe en soit comme en témoignent les publications à ce sujet (Kosseim et Poibeau (2001), Volk et Clematide (2001), Mikheev et al. (1999)). Cependant, les difficultés rapportées dans ces publications concernent davantage la classification des entités nommées que leur reconnaissance. Plusieurs indices peuvent être mis à profit pour reconnaître le statut d'entité d'un ou plusieurs mots. Par contre, une fois l'entité reconnue, il est plus délicat d'en identifier la nature (nom de lieu, nom de personne, nom d'organisation, etc.). En ce qui a trait à l'étiquetage, la reconnaissance des entités, sans avoir à en identifier la nature (outre que grammaticale), serait probablement suffisante pour améliorer les résultats des étiqueteurs.

Un autre des problèmes potentiellement causés par les entités nommées pour l'étiquetage automatique survient lorsque des noms communs sont « dénaturés » lorsqu'ils se retrouvent au sein d'entités nommées. Cette situation n'a pas été observée dans nos travaux, ce qui ne veut pas dire qu'elle ne s'est pas produite. Par exemple, le nom de famille *Plante* peut se trouver suivant une conjonction de subordination : « Il se peut que Plante soit blessé ». Dans ce cas, le nom commun *plante* ne peut y prendre place sans être précédé d'un déterminant : « Il se peut que la plante soit fanée ». Par conséquent, reconnaître les entités nommées est important car de faibles statistiques peuvent tout de même affecter l'équilibre fragile des trigrams et se manifester par des erreurs occasionnelles lorsque des conditions permettent de favoriser l'étiquette *nom propre*.

Il apparaît donc que l'étude de l'impact de la reconnaissance des entités nommées serait intéressante pour vérifier si elle améliore ou non la qualité de la désambiguïsation.

Consultation du contexte droit

Il a été suggéré aux sections 0, 0 et 0 que le prototype aurait avantage à consulter le contexte droit pour éclairer ses décisions. Le modèle utilisé par le prototype, les trigrams, n'utilise que le contexte gauche pour étiqueter. Utiliser d'autres procédés en permanence où sous forme d'exceptions ne serait plus uniquement l'application du modèle des trigrams. Par conséquent, l'idée d'utiliser le contexte droit évoque la possibilité d'utiliser plusieurs modèles mathématiques dépendamment de la nature du mot à désambiguïser. C'est pourquoi une avenue qui serait, à notre avis, des plus intéressantes à enquêter serait de sélectionner le modèle statistique à utiliser selon la nature de l'unité lexicale en cours de désambiguïsation, autrement dit, baser tout le processus de désambiguïsation sur les génotypes.

En effet, nous avons observé à maintes reprises, et cela est également rapporté dans d'autres publications (Church (1998), Marshall (1987)) que l'élément clé de la désambiguïsation est souvent le mot précédent. Dans ces cas, en utilisant des trigrams, la probabilité de l'étiquette de l'homographe est inutilement diluée par la probabilité de l'étiquette antéprécédente. Dans d'autres cas, c'est l'étiquette antéprécédente qui est importante et l'étiquette précédente ne fait que diluer l'efficacité de sa précédente. Par exemple, dans le cas d'un participe passé séparé de l'auxiliaire par un adverbe (... *il est rapidement arrivé*...), le pouvoir de désambiguïsation est davantage porté par l'auxiliaire que l'adverbe.

Il nous apparaît probable de faire avancer les recherches et d'améliorer la qualité de l'étiquetage automatique en étudiant le meilleur modèle pour désambiguïser un type d'homographe en particulier. Ce mémoire a porté sur les homographes de type verbe/substantif, mais en fait, tous les types d'ambiguïté (génotypes) pourraient être étudiés de la même façon.

En ce qui concerne les ngrams, il serait possible d'évaluer automatiquement quel niveau de ngrams serait le plus utile à la désambiguïsation pour tous les types d'ambiguïté dans un corpus. On découvrirait probablement que dans certains cas, il s'agit de bigrams, dans d'autres, de trigrams et peut-être même, dans certains cas, de ngrams de niveau supérieur.

Ce concept s'applique aussi à d'autres modèles que les ngrams et il serait possible de sélectionner plusieurs modèles (ngrams, chaîne cachées de Markov (Hidden Markov Models), réseaux neuronaux, apprentissage « à la Brill », etc.), et de sélectionner le plus adéquat pour chacun des types d'homographes (génotypes) possibles. Cela ferait un bon sujet de doctorat...

Confiner les ngrams à l'intérieur des frontières de syntagme

L'un des problèmes de l'utilisation telle quelle des trigrams non abordés dans ce mémoire, est le fait que certains d'entre eux chevauchent les frontières de syntagmes et de phrases. Par exemple, chaque premier mot d'une phrase ne retire aucun bénéfice à voir la probabilité de son étiquette calculée à partir des unités qui le précèdent, soient, pour notre prototype, le dernier mot de la phrase précédente et la ponctuation finale de cette phrase. Autrement dit, le succès de l'attribution de l'étiquette du premier mot de chaque phrase tient presque du hasard. Le même phénomène se produit avec certains types de syntagmes, comme les syntagmes nominaux. Par exemple, puisque la plupart des syntagmes nominaux doivent être introduits par un déterminant, la structure du syntagme précédent n'influence pas la structure du syntagme suivant. Autrement dit, la désambiguïsation des deux premiers mots du début d'un syntagme nominal ne bénéficie pas des deux derniers mots du syntagme précédent. Dans ce cas, heureusement, les déterminants sont peu ambigus et la désambiguïsation du nom qui suit le déterminant est moins aléatoire que dans l'exemple précédent.

Il est théoriquement intéressant d'imaginer un système de ngrams qui ne seront calculés qu'à l'intérieur des syntagmes. Cependant, en pratique, il serait difficile de le faire car, d'une part, il faudrait identifier ces frontières au moment de la compilation des statistiques, ce qui est faisable, mais, plus important encore, il faudrait le faire au moment d'appliquer les ngrams à la désambiguïsation, ce qui est plus difficile, voire impossible à réaliser. En effet, cela ressemble au problème de l'oeuf et de la poule : lequel survient avant l'autre? Les ngrams auraient ainsi besoin d'un parseur syntaxique alors que les parseurs syntaxiques ont besoin d'un étiqueteur pour déterminer la nature grammaticale des mots pour qu'ils puissent à leur tour déterminer la structure syntaxique des phrases.

Bien que cette suggestion ne découle pas d'observations quantifiées dans ce mémoire, il serait peut-être intéressant d'explorer dans cette direction ne serait-ce que pour éclaircir les limites de l'utilisation actuelle des ngrams à ce sujet.

Comparer l'efficacité des différents modèles de désambiguïsation à partir des mêmes contextes syntaxiques

En identifiant des situations syntaxiques évidentes pour les locuteurs et donc, évidentes du point de vue linguistique, nous croyons avoir identifié des éléments méthodologiques qui pourraient être utilisés pour comparer des systèmes d'étiquetage entre eux. Il s'agit d'une méthodologie qui permettrait de comparer la capacité de différents systèmes pour la désambiguïsation de types d'ambiguïté (ou génotypes). On pourrait alors voir quelles sont les différences d'efficacité des différents systèmes et ensembles d'étiquettes. D'un point de vue pratique, les résultats pourraient être exprimés en terme de succès ou d'échecs comme ce fut le cas dans ce travail. Ce faisant, les difficultés dues à la comparaison d'ensembles d'étiquettes différents pourraient être ignorées. Par ailleurs, ces comparaisons pourraient mener à des observations qui pourraient renforcir l'usage de certaines étiquettes, le remplacement d'autres ou la création de nouvelles étiquettes.

Le futur des trigrams pour l'étiquetage automatique

Nous voyons beaucoup d'avenir à poursuivre les recherches du côté des génotypes. Pour l'instant, le concept de génotype réfère au regroupement de l'ensemble des mots partageant les mêmes étiquettes possibles. Cependant, selon les étiquettes actuelles, les génotypes ne semblent pas tous homogènes du point de vue de leur comportement syntaxique.

Partant du principe du structuralisme (Dubois (1969 :9-10), Dubois et al. (1994 :156)) selon lequel les unités lexicales qui sont de même nature peuvent se substituer les unes aux autres dans un même environnement, les génotypes devraient également pouvoir tous se substituer les uns aux autres dans l'ensemble des environnements dans lesquels ils peuvent apparaître. Ainsi, un mot ne pouvant apparaître là où les autres le peuvent ne devrait par faire partie du même génotype. Cette situation est celle de la négation *ne* qui se trouvait dans le même génotype que les autres adverbes. Il est évident dans ce cas que la négation n'avait pas le même comportement syntaxique que les autres membres du génotype. Cela s'est traduit par

des erreurs générées par le prototype. Cela devrait être le cas pour tous les autres mots. Ils ne devraient partager que les mêmes caractéristiques syntaxiques dans un même génotype.

Il est probable que la scission de tous les génotypes en sous-ensembles cohérents du point de vue syntaxique mène à un grand nombre de génotypes. On peut même envisager de nouvelles dénominations pour nommer de nouvelles sous-catégories. Ce que l'on peut anticiper, c'est d'atteindre un stade où les informations requises pour distinguer les mots et les regrouper selon leur comportement syntaxique ne seront plus de nature syntaxique, mais appartenant à d'autres paliers linguistiques tels que la sémantique, la phonologie, etc. Par exemple, il est possible que ce qui distingue certains verbes soit le fait d'être transitif ou non, ce qui est une information qui a des implications syntaxiques car elle autorise ou non un complément d'objet direct. D'un autre côté, peut-être atteindrons nous le stade ou des différences seront marquées par le fait que des verbes acceptent exclusivement des sujets animés ou inanimés. Il y a probablement beaucoup à découvrir dans cette voix et malgré les réticences que nous pouvons avoir à complexifier les données, cela ne ferait que refléter le modèle linguistique acquis par un locuteur natif et s'en approcher. Également, il faudrait prendre garde d'inclure des informations seulement pour favoriser un prototype ou un modèle mathématique ou linguistique. Nous croyons que l'objectif à long terme doit être la description du modèle linguistique et de le fournir aux modèles d'apprentissage automatique. Il est sûr que plus les génotypes seront détaillés, plus la taille des corpus devra être importante pour inclure un nombre significativement représentatif de chacun des phénomènes syntaxiques. Par contre, nous croyons que :

- L'ampleur des corpus serait optimisée. Toutes les occurrences de génotypes pourraient être comptabilisées et réutilisées de façon fiable pour les mots non observés dont le génotype est connu.
- Les modèles d'apprentissage pourraient être comparés entre eux car les informations contenues dans les corpus seraient linguistiquement adéquates et correctes;
- De nouvelles catégories grammaticales pourraient voir le jour;
- Les ngrams seraient probablement suffisants et les probabilités a priori devraient être réservées aux cas d'ambiguïté où le recours au contexte n'est pas suffisant;
- On s'approcherait alors de la capacité maximale des trigrams;
- Enfin, rappelons l'idée exprimée ci-dessus d'utiliser les génotypes (type d'ambiguïté) comme base de calcul des ngrams ou autres modèles.

7. CONCLUSION

Une revue de la littérature permet de constater que les modèles mathématiques sont aptes à saisir des données qui conduisent à des résultats satisfaisants. Ils ont l'avantage d'être très rapides à le faire ce qui leur donne un avantage certain par rapport aux méthodes d'édition manuelle de règles, dites linguistiques. En contrepartie, ils doivent disposer de données préalablement étiquetées pour en déduire mathématiquement les relations et ce, en quantité représentative. Par contre, une fois ces données disponibles, elles peuvent être utilisées à volonté pour tester toutes les variantes de prototypes et de modèles sous-jacents possibles. Un de leurs inconvénients demeure toutefois la difficulté d'en étudier le comportement et leur sensibilité aux ajustements manuels des données qu'ils utilisent. Sur ce plan, les règles linguistiques offrent l'avantage d'être plus compréhensibles et de pouvoir être manipulées plus facilement.

Cependant, bien au-delà des avantages et inconvénients de chacune des méthodes, que ce soit en n'utilisant que des modèles mathématiques, des règles linguistiques ou les deux pour étiqueter automatiquement du texte, il reste une marge d'erreur que les recherches s'efforcent de réduire depuis le début. Cette marge ne s'est amoindrie que bien peu depuis et demeure loin de la performance d'un locuteur humain.

En nous intéressant à la question de l'étiquetage automatique de texte, nous nous sommes aperçu que peu de place était accordée à la linguistique parmi les recherches dans le domaine. L'emphase de la recherche dans ce domaine est davantage mise sur le développement de prototypes qui utilisent des modèles d'apprentissage statistiques plus ou moins différents dans le but d'obtenir les meilleurs résultats possibles sur un ensemble de test. Malheureusement, peu d'attention est portée sur la correspondance des modèles utilisés avec les résultats des recherches en linguistique et pourtant, les modèles mathématiques performent bien. C'est pourquoi, ce projet s'est articulé autour de trois objectifs :

1. Expliquer une part du succès des trigrams
2. Identifier une part des limites des trigrams
3. Établir dans quelle mesure la linguistique peut aider à améliorer les performances des étiqueteurs utilisant des trigrams.

Pour enquêter sur ces objectifs, nous avons réalisé un prototype d'étiqueteur automatique, basé sur un modèle mathématique simple : les trigrams (Church 1988, Charniak et al. 1993). Ce modèle offre l'avantage d'être simple et de donner des résultats semblables aux techniques mathématiques plus complexes. Aucun autre procédé que l'application telle quelle du modèle des trigrams n'a été utilisé dans le prototype ce qui a permis d'étudier strictement le comportement du modèle mathématique et non pas celui d'un prototype qui inclut d'autres méthodes pour améliorer ses résultats.

Le succès des trigrams

En cherchant à expliquer le succès des trigrams, il a été constaté que, de façon générale, les structures syntaxiques présentes dans le corpus d'entraînement sont suffisamment récurrentes pour être statistiquement significatives au su de techniques statistiques simples comme les trigrams. Du moins, elles le sont assez pour permettre de meilleurs résultats de la part des trigrams lorsque comparées à la simple attribution de l'étiquette la plus probable.

En ce qui a trait aux homographes de type verbe/substantif, le modèle des trigrams est clairement plus efficace que celui de l'attribution de l'étiquette la plus probable, ce qui, par conséquent, suggère que l'utilisation des ngrams est plus efficace pour les homographes de type verbe/substantif que pour le reste du lexique.

En faisant l'inventaire des structures des syntagmes nominaux et verbaux, il a été constaté que la majorité des séquences répondent bien au modèle des trigrams, quoique dans plusieurs cas, l'utilisation de bigrams semblerait suffisante.

Bien que les résultats suggèrent que les ngrams améliorent la désambiguïsation, le taux d'amélioration de l'application seule des ngrams demeure relativement bas. En effet, il fut également constaté que même avec de meilleures conditions, le prototype ne parvient pas à s'approcher du 100% plus que les autres systèmes recensés dans la littérature.

L'utilisation des génotypes a également permis de constater que plus les étiquettes exprimant les génotypes sont précises, plus ils sont efficaces. En effet, les génotypes

constitués d'étiquettes de l'ensemble étendu étaient deux fois supérieurs à ceux constitués d'étiquettes de l'ensemble restreint. Cela suggère que plus il y a d'informations sur les mots en question, meilleure est la désambiguïsation

Les limites des trigrams

La limite des trigrams rapportée dans ce mémoire se situe au niveau de « l'incohérence » des succès et des échecs de l'application des trigrams. En effet, il fut observé que plusieurs homographes sont désambiguïsés correctement malgré des erreurs de désambiguïsation parmi les deux mots précédents alors que d'autres homographes, malgré le succès de la désambiguïsation des mots précédents, ne sont pas correctement désambiguïsés. Autrement dit, ce ne sont pas tous les succès qui reposent sur une désambiguïsation adéquate des étiquettes ayant servi au calcul de la probabilité de la nature de l'homographe tout comme ce ne sont pas tous les échecs qui sont dus à une mauvaise désambiguïsation parmi les deux étiquettes précédentes.

En isolant le modèle statistique, cela a permis de mesurer son efficacité à discerner le comportement syntaxique des unités lexicales d'un corpus. **Cette recherche a permis de constater qu'un taux de succès général cache des incongruités importantes.**

L'utilité de la linguistique dans le domaine de l'étiquetage automatique

Tout au long de ce travail, nous avons été préoccupé par la place de la linguistique au sein de la recherche en automatisation de tâches linguistiques; ce que certains appellent la linguistique informatique et d'autres, l'informatique linguistique. En particulier, nous nous sommes questionné sur sa place au sein de l'étiquetage automatique de texte.

Dans ce travail, nos connaissances en linguistique nous ont permis de cibler des contextes syntaxiques non ambigus pour les locuteurs humains, c'est-à-dire des contextes où ces locuteurs sont en mesure de reconnaître la nature des unités lexicales en question sans le moindre doute. Cela a permis de confronter le prototype et son modèle à ces situations et de comparer en quelque sorte, l'efficacité du modèle linguistique de tout locuteur au modèle mathématique.

Nos connaissances en linguistique nous ont aussi permis de jeter un regard critique sur l'adéquation des données linguistiques du corpus et du dictionnaire. En effet, les informations utilisées par le prototype ne sont pas mauvaises en soi du point de vue de la grammaire traditionnelle, mais il leur manque des précisions linguistiques pour que les trigrams reflètent certaines restrictions syntaxiques. Nous avons été à même de constater la sensibilité des trigrams quant à la nature des informations mises à leur disposition. Quelquefois, ce n'est pas le modèle des trigrams qui est en cause, mais plutôt, la description linguistique des étiquettes.

Nos recherches orientées par nos connaissances des faits linguistiques nous ont permis d'identifier trois avenues possibles d'amélioration de notre prototype, soit :

1. améliorer les informations linguistiques mises à la disposition du prototype;
2. reconnaître les noms propres;
3. utiliser le contexte droit.

Au-delà du prototype, notre réflexion nous a permis de proposer des directions de recherche qui nous apparaissent intéressantes.

D'une part, ajouter des propriétés linguistiques aux étiquettes attribuées aux mots dans le corpus et le dictionnaire permettrait, à notre avis, d'améliorer le taux de désambiguïsation, du moins, en ce qui concerne notre prototype.

D'autre part, cet ajout d'informations permettrait de définir de meilleurs génotypes, c'est-à-dire des génotypes au comportement syntaxique plus homogène.

Enfin, utiliser les génotypes, ou types d'ambiguïté, comme point central pour l'apprentissage des meilleures techniques de désambiguïsation pour chacun des génotypes possibles.

Rôles de la linguistique au sein de la recherche en automatisation linguistique

Au sortir de cet effort de réflexion et en regard des données observées au cours de ce travail, nous sommes plus que jamais convaincu que **le rôle de la linguistique au sein des recherches en automatisation de tâches linguistiques est de fournir des données**

d'apprentissage sous forme de corpus étiqueté. En effet, de notre point de vue de linguiste, les techniques d'apprentissage utilisées par les informaticiens pour formaliser les connaissances d'un corpus nous semblent suffisamment évoluées pour extraire adéquatement un nombre intéressant de modélisation en vue d'une automatisation. De plus, nous avons constaté le retard des recherches sur le français par rapport à celles effectuées sur l'anglais. Ce retard se situe plus particulièrement au niveau des données d'apprentissage c'est-à-dire, les corpus. Alors qu'il existe plusieurs corpus étiquetés manuellement de grande envergure pour l'anglais (le corpus Brown, le corpus LOB, le Corpus National Britannique), rien de tel n'existe en français. Nous avons eu de la difficulté à mettre la main sur le seul corpus qui existe à notre connaissance et de surcroît, nous n'en avons obtenu qu'un sous-ensemble. Il nous apparaît urgent de fournir aux équipes de recherche qui s'intéressent au français, des outils comme de grands corpus pour pouvoir non seulement comparer les recherches avec celles effectuées sur les autres langues dans le but d'étudier les universaux du langage et les différences spécifiques à chaque langue, mais aussi pour innover tant au niveau des connaissances linguistiques qu'informatiques. Il va donc de soi que l'apport le plus direct de la linguistique au domaine de l'automatisation de tâches linguistiques est d'abord de faire passer les connaissances accumulées ainsi que les nouvelles découvertes en linguistique dans des corpus à des fins d'apprentissage et d'analyse automatique. Il semble de toute façon que la quantité de données ne soit actuellement pas suffisante pour atteindre la limite des techniques d'apprentissage actuelles, une opinion que partagent Banko et Brill (2001).

Enfin, tout en rattrapant ce retard, à notre avis, un autre rôle de la linguistique est de guider les modèles d'apprentissage automatique et d'évaluer la validité de leur modélisation. Avec tous ces efforts, nous arriverons peut-être un jour à doter un ordinateur de la même aptitude que les êtres humains à discerner la nature des mots.

8. ANNEXE

Conversion des étiquettes du dictionnaire français vers celles du corpus Paris VII.

Étiquettes du dictionnaire	Étiquettes du corpus Paris VII
Abr	
Adj:Card	A-card-ms A-card-mp A-card-fs A-card-fp
Adj:Fem+PL	A-qual-fp
Adj:Fem+SG	A-qual-fs
Adj:Mas+PL	A-qual-mp
Adj:Mas+SG	A-qual-mg
Adv	
Con	
Det	
Det:Fem+PL	
Det:Fem+SG	
Det:Mas+PL	
Det:Mas+SG	
Fem+PL+DemPro+Prox	
Fem+SG+DemPro+Prox	
Int	
InvGen+PL+P1+PProRefl	
Mas+SG+DemPro+Prox	
Mas+SGDemPro2	
Neg	
NNom+Fem+SG+P3+PProRefl	
NNom+InvGen+SG+P1+PProRefl	
NNom+InvGen+SG+P3+PProRefl	
NNom+Mas+SG+P3+PProRefl	
Nom:Fem+PL	NCfp
Nom:Fem+SG	NCfs
Nom:Mas+PL	NCmp
Nom:Mas+SG	NCms
Nom+InvGen+SG+P3+PC	
Ono	
Pre	
Pro	
Pro:3Fem+PL	
Pro:3Fem+SG	
Pro:3Mas+PL	
Pro:3Mas+SG	
Pro:Fem+PL	
Pro:Fem+SG	
Pro:InvGen+PL	
Pro:InvGen+SG	
Pro:Mas+PL	
Pro:Mas+SG	

Pro:PL+P1	
Pro:PL+P2	
Pro:PL+P3	
Pro:SG+P1	
Pro:SG+P2	
Pro:SG+P3	
Pro:SG+P3PL+P3	
QPro	
Ver:CPre+PL+P1	VC1p
Ver:CPre+PL+P2	VC2p
Ver:CPre+PL+P3	VC3p
Ver:CPre+SG+P1	VC1s
Ver:CPre+SG+P2	VC2s
Ver:CPre+SG+P3	VC3s
Ver:IFut+PL+P1	VF1p
Ver:IFut+PL+P2	VF2p
Ver:IFut+PL+P3	VF3p
Ver:IFut+SG+P1	VF1s
Ver:IFut+SG+P2	VF2s
Ver:IFut+SG+P3	VF3s
Ver:IImp+PL+P1	VI1p
Ver:IImp+PL+P2	VI2p
Ver:IImp+PL+P3	VI3p
Ver:IImp+SG+P1	VI1s
Ver:IImp+SG+P2	VI2s
Ver:IImp+SG+P3	VI3s
Ver:Imp+PL+P1	VY1p
Ver:Imp+PL+P2	VY2p
Ver:Imp+SG+P2	VY2s
Ver:ImPre+PL+P1	VY1p
Ver:ImPre+PL+P2	VY2p
Ver:ImPre+SG+P2	VY2s
Ver:Inf	VW
Ver:IPre+PL+P1	VP1p
Ver:IPre+PL+P2	VP2p
Ver:IPre+PL+P3	VP3p
Ver:IPre+SG+P1	VP1s
Ver:IPre+SG+P2	VP2s
Ver:IPre+SG+P3	VP3s
Ver:IPSim+PL+P1	VJ1p
Ver:IPSim+PL+P2	VJ2p
Ver:IPSim+PL+P3	VJ3p
Ver:IPSim+SG+P1	VJ1s
Ver:IPSim+SG+P2	VJ2s
Ver:IPSim+SG+P3	VJ3s
Ver:PPas	VK
Ver:PPas+Fem+PL	VKfp
Ver:PPas+Fem+SG	VKfs
Ver:PPas+Mas+PL	VKmp

Ver:PPas+Mas+SG	VKms
Ver:PPre	VG
Ver:PPre+Fem+PL	VG
Ver:PPre+Fem+SG	VG
Ver:PPre+Mas+PL	VG
Ver:PPre+Mas+SG	VG
Ver:SImp+PL+P1	VT1p
Ver:SImp+PL+P2	VT2p
Ver:SImp+PL+P3	VT3p
Ver:SImp+SG+P1	VT1s
Ver:SImp+SG+P2	VT2s
Ver:SImp+SG+P3	VT3s
Ver:SPre+PL+P1	VS1p
Ver:SPre+PL+P2	VS2p
Ver:SPre+PL+P3	VS3p
Ver:SPre+SG+P1	VS1s
Ver:SPre+SG+P2	VS2s
Ver:SPre+SG+P3	VS3s

9. BIBLIOGRAPHIE

ABEILLÉ, Anne, Lionel CLÉMENT et François TOUSSENEL (2003). Building a Treebank for French. In *Building and using Parsed Corpora*, Anne Abeillé (ed.). Language and Speech series, Kluwer, Dordrecht.

ABEILLÉ, Anne, Lionel CLÉMENT & Rodrigo REYÈS (1998). "TALANA annotated corpus: the first results", *Proceedings of the First Conference on Linguistic Resources*, Grenade, pp. 992-999.

ABNEY, Steven (1996). Statistical Methods and Linguistics. In Judith Klavans and Philip Resnik (eds.). MIT Press, Cambridge.

ANDERSSON, Annette Östling (1987). *L'identification automatique des lexèmes du français contemporain.* Almqvist & Wiksell International: Stockholm.

ATWELL, Eric (1987). Constituent-likelihood Grammar. In Garside, R. Leech, G. & G. Sampson (Eds.). *The Computational Analysis of English: A Corpus-Based Approach*. Longman: Londres.

BANKO, Michele & Eric BRILL (2001). Scaling to Very Very Large Corpora for Natural Language Disambiguation. In Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics and the 10th Conference of the European Chapter of the Association for Computational Linguistics, Toulouse.

BESCHERELLE conjugaison (Le) (1997). Paris, Hatier.

BOUTIN, Jean-Luc (1995). « La reconnaissance automatique de l'adverbe : perspective morphologique ». Mémoire de maîtrise, Québec, Université Laval.

BRILL, Eric (1995). Transformation-Based Error-Driven Learning and Natural Language Processing : A Case Study in Part-of-Speech Tagging. *Computational Linguistics*, Décembre.

BROUSSEAU, Anne-Marie & Yves ROBERGE (2000). Syntaxe et sémantique du français. Montréal : Éditions Fides.

CARADEC, Robert & Gilles SAADA (1982). « Définition de la classe syntaxique d'une forme lexicale à partir de sa terminaison graphique ». *Lingvisticae Investigationes* VI : 2, p.271-281.

CARRÉ, René, Jean-François Dégremont, Maurice Gross, Jean-Marie Pierrel et Gérard Sabah (1991). Langage humain et machine. Presses du CNRS : Paris.

CHANDIOUX, John (1998). « Les promesses de la traduction automatique », *Terminogramme*, no 84-85, mars 1998, p. 23-25.

CHANOD, Jean-Pierre & Pasi TAPANAINEN (1995a). Tagging French – comparing a statistical and a constraint-based method, *Proceedings of the Seventh Conference of the European Chapter of the Association for Computational Linguistics* (EACL'95). Association for Computational Linguistics, Dublin, pp. 149-156.

----- (1995b). Creating a tagset, lexicon and guesser for a French Tagger, Proceedings of the ACL *SIGDAT workshop*: *From Texts To Tags: Issues In Multilingual Language Analysis*, University College Dublin, Ireland, pp. 58-64.

CHARNIAK, Eugene, Curtis HENDRICKSON, Neil JACOBSON & Mike PERKOWITZ (1993). Equations for Part-of-Speech Tagging. In *Proceedings of the eleventh national conference on artificial intelligence*, Washington, DC, pp 784-789.

CHURCH, Kenneth W.(1988). A Stochastic Parts Program and Noun Phrase Parser for Unrestricted Text. In *Second Conference on Applied Natural Language Processing*, Austin, Texas, pp. 136-143.

CLÉMENT, Lionel (2001). Construction et exploitation d'un corpus syntaxiquement annoté pour le français, Thèse de doctorat, Paris VII.

CUTTING, Doug, KUPIEC, Julian, PEDERSEN, Jan & Penelope SIBUN (1992). A Practical Part-of-Speech Tagger. In *Proceedings of ANLP-92*.

DUBOIS, Jean (1969). Grammaire structurale du français : la phrase et les transformations. Paris : Larousse.

DUBOIS, Jean, GUESPIN Louis, GIACOMO, Mathée, MARCELLESI Christiane, MARCELLESI, Jean-Baptiste et Jean-Pierre MÉVEL (1994). *Dictionnaire de linguistique et des sciences du langage*. Paris : Larousse.

FRANCIS, Nelson W. (1980). A Tagged Corpus – Problems and Prospects. In *Studies in English Linguistics*. London : Longman, p. 192-209.

FUCHS, Catherine, Laurence DANLOS, Anne LACHERET-DUJOUR, Daniel LUZZATI et Bernard VICTORRI (1993). *Linguistique et traitement automatique des langues*. Hachette : Baume-les-Dames.

GARSIDE, Roger (1987). The CLAWS Word-Tagging System. In Garside, R. Leech, G. & G. Sampson (Eds.). *The Computational Analysis of English: A Corpus-Based Approach*. Longman: Londres.

GARSIDE, Roger, Geoffrey LEECH et Anthony McENERY, éd. (1997). *Corpus Annotation : Linguistic Information from Computer Text Corpora* (Collectif). Longman : New-York.

GREENE, Barbara B. & Gerald M. RUBIN (1971). Automatic Grammatical Tagging of English. Providence, Rhode Island: Brown University, 153p.

GRÉVISSE, Maurice (1986). *Le bon usage* (12ème édition). Duculot : Paris.

GUILLET, Alain (1990). « Reconnaissance des formes verbales avec un dictionnaire minimal ». *Langue Française*, no 87. Paris : Larousse.

HABERT, Benoît, NAZARENKO, Adeline & André SALEM (1997) *Les linguistiques de corpus*. Armand Colin : Paris.

HARRIS, Zelig S. (1964). *String Analysis of Sentence Structure*. Deuxième édition. Mouton & Co. : The Hague.

JELINEK, Frederick (1985). Markov Source Modeling of Text Generation. In *The Impact of Processing Techniques on Communications*, Skwirzynski, J.K. (ed.). Nato Asi Series, Series E: Applied Sciences, no 91, Martinus Nijhoff Publishers: Dordrecht.

JENSEN, Karen & George HEIDORN (1982). *The fitted parse: 100% parsing capability in a syntactic grammar of English*. San José, CA: IBM Research Division. Computer science research report, RC9729.

JOHANSSON, Stig & Mette-Cathrine JAHR (1982). Grammatical Tagging of the LOB Corpus : Predicting Word Class from Word Endings. In *Computer Corpora in English Language Research*, (Stig Johansson ed.). Norwegian Computing Centre for the Humanities: Bergen.

KOSSEIM, Leila et Thierry POIBEAU (2001). Extraction de noms propres à partir de textes variés : problématiques et enjeux. In *Actes de la 8ème conférence annuelle sur le Traitement Automatique des Langues Naturelles (TALN-2001)*, Tours (France), juillet 2001, p. 365-363.

KLEIN, Sheldon & Robert F. SIMMONS (1963). A Computational Approach to Grammatical Coding of English Words. In *Journal of the Association for Computing Machinery*, vol. 10, p. 334-347.

LABESSE, Henri (1985). « Un analyseur syntaxique du français ». *Revue québécoise de linguistique* (Numéro thématique : Linguistique et informatique), vol. 14, n° 2, p. 103-117.

LADOUCEUR, Jacques (1988). *Une analyse automatique en syntaxe textuelle*. Recherche en linguistique appliquée à l'informatique, K-5. Centre International de Recherche sur le Bilinguisme (CIRB) : Québec.

LECOMTE, Josette (1998). Le catégoriseur Brill14-JL5/WinBrill-0.3, Rapport technique, Inalf.

MAEGAARD, Bente et Ebbe SPANG-HANSSEN (1978). *La segmentation automatique du français écrit*. Éditions Jean-Fayard : Paris.

MANNING G. Christopher et Hinrich SCHÜTZE (2001). Foundations of Statistical Natural Language Processing, MIT Press, Massachussetts, Cambridge.

MARSHALL, Ian (1983). Choice of Grammatical Word-Class without Global Syntactic Analysis: Tagging Words in the LOB Corpus. *Computers and the Humainities*, vol. 17, p. 139-150.

----- (1987). Tag Selection Using Probabilistic Methods. In Garside, R. Leech, G. & G. Sampson (Eds.). *The Computational Analysis of English: A Corpus-Based Approach*. Longman: Londres.

MIKHEEV, Andrei, Marc MOENS and Claire GROVER, Named Entity Recognition without Gazetteers, In *Proceedings of EACL'99*, Bergen, Norway, 1999, pp. 1-8.

MORIN, Jean-Yves (1985). « Théorie syntaxique et théorie du parsage : quelques réflexions ». *Revue québécoise de linguistique* (Numéro thématique : Linguistique et informatique), vol. 14, n° 2, p. 9 - 48.

QIAO, Hong Liang & Renjie TONG (1998). Design and Implementation of the AGTS Probabilistic Tagger. *ICAME Journal*, no 22.

RABINER, L. R. & B. H. JUANG (1986). An Introduction to Hidden Markov Models. *IEEE ASSP Magazine*, Janvier.

RASTIER, François, Marc CAVAZZA et Anne ABEILLÉ (1994). Sémantique pour l'analyse. Paris : Masson.

REYES, R. (1997). Un étiqueteur du français inspiré du taggeur du Brill, Rapport de stage, UFRL, Paris VII.

ROSNER, Michael et Roderick JOHNSON (éds.) (1992). Computational linguistics and formal semantics. Cambridge University Press: Cambridge.

SALTON, Gerard et Michael J. MCGILL (1983). *Introduction to Modern Information Retreival*. McGraw-Hill : New York.

SAMPSON, Geoffrey (1987). Probabilistic Models of Analysis. In Garside, R. Leech, G. & G. Sampson (Eds.). *The Computational Analysis of English: A Corpus-Based Approach*. Longman: Londres.

SAMUELSSON, Christer & Atro VOUTILAINEN (1997). Comparing a Linguistic and a Stochastic Tagger. *Proceedings of 35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics*, ACL, Madrid, Espagne.

SCHMID, Helmut (1994). Part-of-Speech Tagging with Neural Networks. In *Proceedings of the International Conference on Computational Linguistics*, p. 172-176, Kyoto, Japon.

SILBERZTEIN, Max (1993). Dictionnaires électroniques et analyse automatique de textes. Masson : Paris.

SIMMONS, Robert F., KLEIN, Sheldon & Keren MCCONLOGUE (1962). Toward the Synthesis of Human Language Behavior. *Computers in Behavioral Science*, juillet 1962, p. 402-407.

STARETS, Moshé (2000). Théories syntaxiques du français contemporain. Québec : Les presses de l'Université Laval.

TELLIER, Christine (1995). Éléments de syntaxe du français. Montréal : Les presses de l'Université de Montréal.

TZOUKERMANN, Evelyne & Dragomir R. RADEV (1996). Using Word Class for Part-of-Speech Disambiguation. *Proceedings of the Fourth Workshop on Very Large Corpora WVLC'96*, Copenhagen, Denmark.

TZOUKERMANN, Evelyne, Dragomir R. RADEV & William A. GALE (1997). "Tagging French without Lexical Probabilities – Combining Linguistic Knowledge and Statistical Learning". in *Natural Language Processing using Very Large Corpora*, Eds Amstrong, Susan and Kenneth Church and Pierre Isabelle and, Evelyne Tzoukermann and David Yarowsky, Kluwer.

----- (1995). Combining Linguistic Knowledge and Statistical Learning in French Part-of-Speech Tagging, *Proceedings of the EACL Workshop on Very Large Corpora WVLC'95*, Dublin, Republic of Ireland.

VÉRONIS, Jean (2000). « Annotation automatique de corpus : panorama et état de la technique ». In J.-M. PIERREL (Ed.), *Ingénierie des langues*. Paris : Éditions Hermès. p.111-129.

VOLK, Martin & Simon CLEMATIDE: *Learn-Filter-Apply-Forget. Mixed Approaches to Named Entity Recognition.* In: *Proc. of 6th International Workshop on Applications of Natural Language for Information Systems.* Madrid: 2001.

WEHRLI, Éric (1997). L'analyse syntaxique des langues naturelles : problèmes et méthodes. Masson : Paris.