MÉLANIE THIBEAULT

LA CATÉGORISATION GRAMMATICALE AUTOMATIQUE : ADAPTATION DU CATÉGORISEUR DE BRILL AU FRANÇAIS ET MODIFICATION DE L'APPROCHE

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

NOVEMBRE 2004

Résumé

La catégorisation grammaticale automatique est un domaine où il reste encore beaucoup à faire. De très bons catégoriseurs existent pour l'anglais, mais ceux dont dispose la communauté francophone sont beaucoup moins efficaces. Nous avons donc entraîné le catégoriseur de Brill pour le français pour ensuite en améliorer les résultats. Par ailleurs, quelle que soit la technique utilisée, certains problèmes restent irrésolus. Les mots inconnus sont toujours difficiles à catégoriser correctement. Nous avons tenté de trouver des solutions à ce problème. En somme, nous avons apporté une série de modifications à l'approche de Brill et évalué l'impact de celles-ci sur les performances. Les modifications apportées ont permis de faire passer les performances du traitement des mots inconnus français de 70,7% à 78,6%. Nous avons donc amélioré sensiblement les performances bien qu'il reste encore beaucoup de travail à faire avant que le traitement des mots inconnus français soit satisfaisant.

Avant propos

Nous tenons à remercier notre directeur de maîtrise, Jacques Ladouceur, pour son soutien et son enthousiasme constant ainsi que pour nous avoir permis d'utiliser son catégoriseur fondé sur l'approche de Brill. Nous tenons aussi à remercier Charlotte Jalbert et Martin Thibeault pour la correction du présent mémoire. Nous voulons aussi remercier l'INALF pour nous avoir permis d'utiliser leurs règles et leur dictionnaire pour catégoriser notre corpus d'entraînement. Nous remercions aussi Eric Brill pour avoir mis son logiciel à la disposition de la communauté scientifique. Finalement, nous remercions le fonds FQRSC et le CRSH pour leur soutien financier.

Table des matières

Liste des tableaux

Liste des figures

Introduction

Avec l'avènement de l'informatique au cours de la deuxième moitié du 20e siècle, de nouvelles opportunités se sont présentées et ce, dans divers domaines. Internet fait maintenant partie intégrante de notre quotidien et il contient une quantité phénoménale d'informations qui sont très utiles, voire indispensables. On peut y trouver, par exemple, une grande quantité d'ouvrages sur support électronique que les auteurs ont décidé de rendre publics. La masse d'information qu'on retrouve sur Internet est tellement grande et écrite en tant de langues que sans outils de traitement linguistique, il serait très difficile, voire impossible, d'y trouver ce que l'on cherche. Il existe maintenant une grande quantité de logiciels traitant les langues naturelles qui facilitent la consultation des documents électroniques, comme par exemple les outils de recherche sur Internet, les résumeurs automatiques et les traducteurs automatiques. Parmi ces logiciels, la majorité ne donne que des résultats partiellement satisfaisants. Le meilleur exemple est sans doute le logiciel de traduction automatique. Qui n'a pas déjà été dérouté par les résultats d'un tel logiciel ? Il est donc nécessaire de parfaire ces logiciels. Beaucoup de travail a déjà été accompli, mais dans certains cas il en reste tout autant à faire, sinon plus. En traitement automatique du langage, les premiers pas sont toujours beaucoup plus faciles à faire que les suivants.

Pour que l'ordinateur ait un jour pleinement accès au sens véhiculé dans les textes, il lui faut une représentation linguistique complète des énoncés. Une des étapes essentielles pour atteindre ce résultat est l'identification de la catégorie grammaticale à laquelle appartiennent les mots du texte. Dans le cadre de notre maîtrise, nous avons décidé de nous attaquer à ce problème spécifique. Pour le moment, de nombreux logiciels de catégorisation grammaticale automatique existent. Les plus connus sont ceux qui sont

issus des modèles de Markov, les catégoriseurs à base de règles et les catégoriseurs neuronaux. Tous produisent environ 3-4 % d'erreurs pour l'anglais. Puisqu'une page normale de texte contient environ 500 mots, il y a de 15 à 20 erreurs par page, soit presque une erreur par phrase.

Par ailleurs, les logiciels de catégorisation grammaticale automatique ne sont jamais utilisés seuls. En effet, la simple possibilité d'obtenir la catégorie grammaticale des mots d'un texte n'intéresse pas beaucoup de gens. Mais les catégoriseurs sont des outils essentiels dans une grande quantité d'autres traitements. Ils sont essentiels pour la traduction automatique, la correction grammaticale automatique, la génération automatique de résumés et le repérage d'information. Ils sont aussi utilisés en reconnaissance et en synthèse vocale. Si on veut, par exemple, traduire automatiquement un texte, il faut que les catégories grammaticales des mots du texte soient connues. On ne traduira pas *ferme* de la même façon s'il est un nom (anglais : *farm*) ou s'il est un verbe (anglais : *close*). En fait, la catégorisation grammaticale est nécessaire pour l'analyse syntaxique qui elle l'est pour l'analyse sémantique. Une erreur de catégorisation peut donc avoir une grande répercussion sur les autres niveaux d'analyse qui comportent eux aussi des lacunes. On se retrouve donc avec des résultats peu satisfaisants, comme c'est le cas en traduction automatique. Il est donc essentiel d'améliorer chacune des étapes du traitement automatique des langues naturelles.

D'autre part, les logiciels de traitement des langues naturelles sont normalement beaucoup plus efficaces pour traiter l'anglais que toute autre langue, principalement parce qu'il y a eu plus de recherche appliquée à l'anglais. Il en est de même pour la catégorisation grammaticale automatique. La nécessité d'améliorer la catégorisation grammaticale automatique est donc

plus grande en français. Les résultats obtenus avec l'anglais surpassent ceux obtenus avec le français de 2-3 %.

Comme nous l'avons déjà dit, les catégoriseurs les plus connus sont les catégoriseurs statistiques (comme celui de Markov), les catégoriseurs neuronaux et les catégoriseurs à base de règles. Appliqués à l'anglais, ils ont tous des performances similaires, soit à peu près 96-97 % de bons résultats. Dans le cadre de notre maîtrise, nous avons choisi d'améliorer un catégoriseur à base de règles, soit celui de Brill. Les règles utilisées par ce type de catégoriseurs sont similaires à celles que nous sommes habituée de manipuler comme linguiste. Avec les approches neuronales et statistiques, nos connaissances de la langue n'auraient pas été mises à profit. En fait, les catégoriseurs neuronaux et statistiques font très peu usage de connaissances linguistiques explicites.

L'approche de Brill atteint 97 % d'efficacité pour l'anglais, contre 94 % pour le français. Notre travail de recherche vise à réduire le plus possible l'écart entre les performances pour les deux langues. Mais, si le travail de conception d'un logiciel est long et fastidieux, le travail d'amélioration l'est tout autant, sinon plus. Les quelques pourcentages d'erreurs restants demanderont probablement beaucoup plus de travail que les 95 premiers. Par ailleurs, les résultats sont inférieurs en français principalement à cause du fait que l'approche a été développée à partir de textes anglais. La méthode de Brill se base sur le contexte et la morphologie pour créer ses règles. De ce côté, le français et l'anglais se ressemblent. Dans les deux langues, la position des mots dans la phrase n'est pas libre, comme ce serait le cas dans certaines langues à déclinaison, comme le russe ou le latin. De plus, la morphologie est somme toute assez réduite dans les deux langues, si on les compare avec des langues à déclinaison ou à des langues agglutinantes, comme le hongrois et l'inuktitut. Il y a donc fort à parier que l'approche

pourra atteindre des résultats aussi bons en français qu'en anglais. Ce qui ne serait pas le cas avec une langue agglutinante comme le hongrois. En effet, Megyesi 1998 a adapté l'approche de Brill pour le hongrois et n'a obtenu que des performances de 85 %.

D'autre part, un logiciel, par le biais de son dictionnaire, ne peut connaître tous les mots de la langue. Il y a toujours des mots rares qui ne seront pas connus : des sigles, des noms propres, des mots de spécialité, des néologismes, des mots étrangers, etc. Les mots inconnus sont donc inévitables. L'approche de Brill n'atteint un taux d'efficacité que d'environ 82 % pour les mots inconnus. Le traitement de ce type de mot représente un défi pour tous les types de traitement de la langue naturelle. La résolution de ce problème résoudra beaucoup de problèmes de traitement automatique de la langue naturelle et fera avancer les connaissances.

Nous nous sommes donc attaqué à l'adaptation du catégoriseur de Brill au français et à l'amélioration de la catégorisation des mots inconnus dans nos recherches. Pour y arriver, nous avons fait une évaluation contrôle et une typologie des erreurs. À la lumière de cette typologie, nous avons apporté des modifications à l'approche, soit par implémentation quand cela était possible, soit par simulation quand l'architecture du logiciel ne permettait pas de procéder par implémentation. Puis, nous avons évalué les résultats pour chaque modification sur deux textes.

Dans le premier chapitre, nous verrons d'abord le principal problème que doit résoudre l'ordinateur, soit l'ambiguïté. En traitement automatique des langues naturelles, l'ordinateur fait sans cesse face à ce problème. Dans le cadre de la catégorisation grammaticale automatique, le type d'ambiguïté à résoudre est l'ambiguïté grammaticale. Nous verrons ensuite les différents modèles utilisés pour catégoriser automatiquement un corpus. Ces modèles se divisent en deux grandes familles : les catégoriseurs supervisés (partant

de texte précatégorisé) et non supervisés (partant de texte brut). Ces deux grandes familles se subdivisent à nouveau en trois : catégoriseur à base de règles (comme celui de Brill), catégoriseur statistique ou catégoriseur neuronal. Tous les types de modèles ont à peu près la même efficacité et utilisent le contexte et la morphologie. La différence tient à la représentation de ces divers éléments. Puis nous regarderons de plus près les mots inconnus qui posent problème à tous ces modèles. Ensuite, nous verrons en détail le fonctionnement de l'approche de Brill qui se divise en deux parties, soit un module d'apprentissage de règles et un module de catégorisation. Au cours du deuxième chapitre, nous verrons la méthodologie utilisée. Nous verrons d'abord les objectifs poursuivis, suivra une description des corpus et des échantillons utilisés ainsi que les définitions des étiquettes grammaticales utilisées pour déterminer quelles sont les erreurs produites et évaluer les performances de l'approche. Nous verrons aussi comment l'évaluation a été effectuée. Dans le troisième chapitre, nous verrons une typologie des erreurs commises par le logiciel lors de la catégorisation ainsi que les modifications que nous proposons pour les enrayer. Les modifications proposées ont été vérifiées par implémentation quand cela était possible ou par simulation quand l'architecture du logiciel ne permettait pas de procéder par implémentation. Les modifications vérifiées par implémentation touchent la longueur des suffixes, la recherche d'infixes et la fréquence des mots utilisés pour formuler des règles. Les modifications vérifiées par simulation touchent le traitement des préfixes, des suffixes grammaticaux et lexicaux. La dernière section de ce chapitre présente une récapitulation des résultats obtenus.

Chapitre 1 État de la question et problématique

1.1 L'ambiguïté

En traitement automatique des langues naturelles, le principal problème à résoudre est l'ambiguïté. Il existe différents types d'ambiguïtés. D'abord, les mots peuvent être ambigus aux niveaux lexical et grammatical. Le mot *son* est ambigu lexicalement. Il peut désigner une céréale ou encore une impression auditive. Le mot *ferme,* quant à lui, est ambigu grammaticalement. Il peut appartenir à quatre catégories grammaticales différentes : nom, verbe, adjectif et adverbe. Le sens de ce mot sera très différent selon sa catégorie : nom = « bâtiment », verbe = « clore », adjectif = « dur » et adverbe = « fort ». Il existe aussi des ambiguïtés qui relèvent du niveau syntaxique. Une même phrase peut avoir plusieurs sens possibles. La phrase *l'homme voit la femme avec un télescope* peut signifier que l'homme voit la femme à l'aide d'un télescope ou encore que la femme qu'il voit a un télescope. Il existe encore des ambiguïtés pragmatiques. La phrase *peux-tu me passer le sel* peut signifier à la fois « passe-moi le sel » et « es-tu en mesure de le faire ». Dans le cadre de logiciels de traduction automatique, par exemple, toutes ces ambiguïtés doivent être levées.

Le logiciel que nous étudions s'attaque, à la levée des ambiguïtés grammaticales. Les catégoriseurs se servent normalement du contexte, c'est-à-dire de la proximité de certains mots (un mot précédé de *M.* est un nom propre) ou de certains types de mots (ex. : le mot *le* est un pronom quand il est suivi d'un verbe et un déterminant quand il est suivi d'un nom), et de la présence de certains morphèmes, ou suites de lettres, dans le mot (ex. : un mot se terminant par *-emment* est un adverbe). Toutefois, la tâche n'est pas aussi facile qu'elle en a l'air. L'utilisation du contexte est très efficace pour des langues à position de mots fixe, mais elle ne peut résoudre toutes les

ambiguïtés. Il faut alors se tourner vers l'utilisation de la morphologie. Toutefois, la seule présence d'un morphème n'est souvent pas suffisante pour déterminer la catégorie d'un mot. Un même morphème peut être associé à des mots de différentes catégories. C'est le cas de *-ique* qui peut être associé soit à un nom (ex. : la *mécanique*), soit à un adjectif (ex. : la pelle *mécanique*). De plus, ce ne sont pas tous les types de mots qui disposent d'une morphologie leur étant propre. Les sigles et les noms propres ne possèdent pas de morphèmes caractéristiques. Il n'y a donc pas d'indice, à part les majuscules, qui permette de les catégoriser. Nous verrons maintenant quelles sont les grandes approches utilisées pour résoudre l'ambiguïté grammaticale.

1.2 Les différents modèles de catégorisation grammaticale automatique

Les différents modèles utilisent tous les mêmes informations pour catégoriser les mots d'un texte : le contexte et la morphologie. Ce qui diffère, c'est la façon de représenter ces éléments.

Il existe deux grands types de catégoriseurs. Les premiers sont ceux qui appliquent des règles qui leur ont été fournies par des experts humains. Dans ce type de catégoriseurs, il y a très peu d'automatisation; c'est le concepteur qui dicte toutes les règles de catégorisation et qui fournit au besoin une liste de morphèmes. La conception d'un tel catégoriseur est fastidieuse, longue et coûteuse. De plus, les catégoriseurs ainsi conçus ne sont pas portables, c'est-à-dire qu'ils ne sont efficaces que pour une langue donnée et pour un domaine donné (ex. : le droit, la médecine, etc.). Le deuxième type, celui sur lequel nous nous attarderons ici, apprend de façon automatique comment catégoriser. Parmi les catégoriseurs de ce type, il existe deux grandes familles : les catégoriseurs supervisés qui apprennent à partir de corpus précatégorisés, et les catégoriseurs non supervisés qui

apprennent à partir de corpus bruts (description détaillée à la section (1.2.1)). Puis chaque grande famille peut être divisée en trois branches chacune : catégoriseur à base de règles, statistique et neuronal (décrits aux sections 1.2.2, 1.2.3 et 1.2.4). La figure ci-dessous est évidemment simplifiée, dans les faits, certains catégoriseurs grammaticaux combinent des éléments de plusieurs approches. Celui de Brill (décrit à la section 1.4) est de type mixte. C'est un catégoriseur supervisé, à base de règles, utilisant des données statistiques.

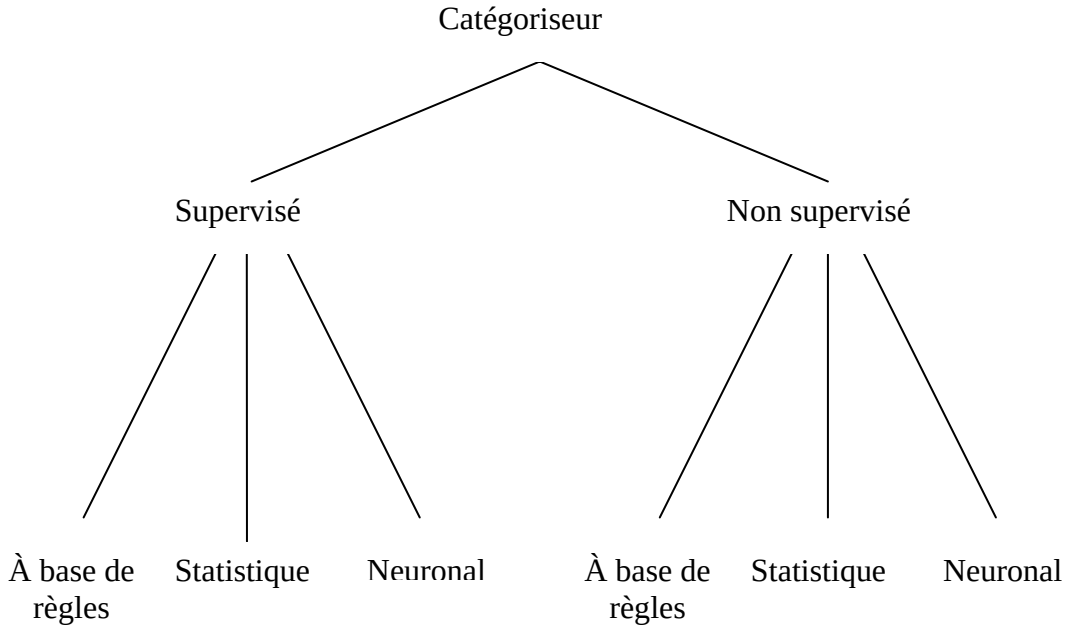

Figure 1 – **Les différents types de catégoriseurs**

Voyons maintenant un bref historique du développement de ces modèles. Les premiers modèles créés étaient supervisés et à base de règles (décrits à la section 1.2.2) et atteignaient 77 % d'efficacité pour l'anglais (Greene et Rubin 1971). Par la suite, des modèles statistiques (décrits à la section 1.2.3) ont été créés, dont le modèle de Markov qui est le plus répandu (voir Church 1985 et Kempe 1993). Les logiciels créés sur la base de ces modèles étaient

entraînés sur de très gros corpus précatégorisés et atteignaient 96 % d'efficacité pour l'anglais. Puis furent créés des modèles non supervisés (décrits à la section 1.2.1) pour éviter la catégorisation de gros corpus d'entraînement. Jelinek 1985 et Cutting et al. 1992 ont entraîné un catégoriseur de ce type et ont obtenu 96 % d'efficacité pour l'anglais. Puis les modèles neuronaux (décrits à la section 1.2.4) ont été développés (Lippman 1989). Ils atteignaient un niveau d'efficacité similaire pour l'anglais à celui obtenu avec les autres techniques. Nous verrons maintenant brièvement le fonctionnement de chacun de ces types de catégoriseurs ainsi que leurs avantages et leurs inconvénients.

1.2.1 Catégoriseurs supervisés et non supervisés

Cette première distinction en est une quant au degré d'automatisation du processus d'entraînement. Les catégoriseurs avec supervision sont entraînés sur des corpus préalablement catégorisés à la main (voir un exemple de corpus catégorisé en annexe B) qui servent de base à la création du matériel nécessaire à la catégorisation. Ce matériel comprend notamment des dictionnaires (voir en (1)), des listes de fréquence pour un mot et ses catégories possibles (voir en (2)), des statistiques sur la probabilité qu'une séquence d'étiquettes donnée apparaisse ou encore une liste de règles (voir en (3)). Le catégoriseur de Brill fait partie de ceux-ci (voir la section 1.4 pour une description détaillée du catégoriseur de Brill).

1. Exemple de dictionnaire

donnant :participe présent
donne :verbe conjugué
donnent :verbe conjugué
donner :verbe infinitif
donné :participe passé
donnée :participe passé, nom
données :nom, participe passé
dont :relatif
dossier :nom
dote :verbe conjugué
doter :verbe infinitif
double :nom, adjectif
doublement :adverbe
doute :nom
dramatisations :nom
drapière :nom
drogue :nom
droit :nom

2. Exemple d'une liste de fréquence « mot – catégorie »

même :adjectif 63
y :pronom verbal 63
bien :adverbe 60
deux :cardinal 60
encore :adverbe 60
éducation :nom 60
leur :déterminant 58
où :relatif 58
psychologie :nom 58
droit :nom 57
en :pronom verbal 57
monde :nom 55
siècle :nom 54
dont :relatif 53
Mais :coordonant 51
donc :adverbe 50
nombre :nom 49
Dans :préposition 48
En :préposition 47
France :nom propre 47

3. Exemples de règles

- Si on peut enlever le suffixe *e* à un mot et que le mot qui en résulte se trouve au dictionnaire, c'est un adjectif.[1]

- Si le mot présente une ambiguïté déterminant /pronom verbal et que le mot suivant est catégorisé verbe conjugué, c'est un pronom verbal (ex. : je ***le*** juge).

Les catégoriseurs sans supervision quant à eux ne nécessitent pas de corpus préalablement catégorisés pour la phase d'entraînement. Ils utilisent plutôt une analyse distributionnelle afin de regrouper automatiquement les mots en

[1] Le mot présente une variation en genre, c'est donc un adjectif.

groupes ou classes de mots, c'est-à-dire en catégories grammaticales. En général, le regroupement des mots en classes se fait en fonction de la similarité des contextes. Plus précisément, le fait que des mots soient interchangeables dans des contextes formellement similaires détermine s'ils font partie d'une même catégorie ou classe. Par exemple, *classe* et *catégorie* peuvent être considérés comme appartenant à la même classe de mots puisqu'il est possible de les retrouver dans le même contexte (ex. : **une** *classe* **de mots**, **une** *catégorie* **de mots**) (voir Atwell 1983 pour plus de détails). En se basant sur ces classes de mots, on peut ensuite établir automatiquement les paramètres nécessaires à un catégoriseur statistique (décrit à la section 1.2.3) ou encore déduire les règles nécessaires à un catégoriseur à base de règles (décrit à la section 1.2.2) ou finalement les poids pour un catégoriseur neuronal (décrit à la section 1.2.4).

Le point fort des catégoriseurs non supervisés est qu'ils sont extrêmement portables. Les catégoriseurs supervisés quant à eux ont tendance à mieux performer si on les utilise pour catégoriser le même type de texte que ceux sur lesquels ils ont été entraînés. Toutefois, le catégoriseur de Brill échappe un peu à la norme. Il est en fait très portable vu le fait que le corpus nécessaire à l'entraînement n'a pas à être très gros (une centaine de pages). De plus, l'approche a été conçue dans une perspective multilingue, ce qui fait que le logiciel peut être entraîné en quelques heures pour tout type de texte et pour toute langue écrite avec un alphabet (arabe : Freeman 2003, hongrois : Megyesi 1999a et 1999b, suédois : Prütz 2002, etc.). Notons aussi qu'il n'y a pas si longtemps, les corpus catégorisés nécessaires à l'entraînement des catégoriseurs supervisés étaient plutôt rares, dispendieux et pas nécessairement disponibles pour toutes les langues et tous les genres de texte. Mais, depuis quelques années, plusieurs corpus ont été créés et rendus disponibles, il est donc de plus en plus facile de travailler dans le cadre d'une approche supervisée. Comme le dit Brill 1995b :

> There are a number of efforts worldwide to manually annotate large corpora with linguistic information, including parts of speech, phrase structure and predicate-argument structure (e.g., the Penn Treebank and the British National Corpus (Marcus, Santorini, and Marcinkiewicz 1993; Leech, Garside, and Bryant 1994)). A vast amount of on-line text is now available, and much more will become available in the future.

Les catégoriseurs non supervisés, quant à eux, ont le désavantage de regrouper les mots en classes trop générales, qui ne permettent pas les distinctions fines que l'on retrouve dans les catégories de mots telles que définies par les linguistes. En somme, les deux types d'approches ont leurs avantages et leurs inconvénients. L'une et l'autre ne seront pas utilisées dans le même type de situation. Un catégoriseur non supervisé pourra être utilisé pour le traitement d'une langue pour laquelle il n'y a pas de corpus catégorisé disponible.

1.2.2 Catégoriseurs à base de règles

Ce type de catégoriseurs crée et utilise des règles contextuelles afin d'identifier la catégorie des mots inconnus ou ambigus. Par exemple, une règle contextuelle peut dire qu'un mot X ambigu ou inconnu précédé d'un déterminant et d'un nom est un adjectif. Cette règle servira à désambiguïser un mot comme *or*, qui peut être un nom (de *l'or*) ou un adjectif (la couleur *or*), dans un contexte comme *la maison or*. En plus des informations contextuelles, plusieurs catégoriseurs exploitent la morphologie pour faciliter la désambiguïsation. Par exemple, une règle qui tient compte de la morphologie peut spécifier qu'un mot ambigu ou inconnu qui se termine par -*ique* et qui est précédé par un nom est un adjectif. Les mots se terminant en -*ique* sont majoritairement ambigus nom/adjectif (ex. : la *mécanique* : nom/un train *mécanique* : adjectif). Dans un contexte comme *un train mécanique*, cette règle permettrait de désambiguïser *mécanique*.

Le catégoriseur de Brill 1995b qui est supervisé et à base de règles obtient un taux d'efficacité de 96,7 %, ce qui est comparable aux résultats des autres approches. Les logiciels conçus selon ce modèle sont faciles à améliorer puisqu'ils nous donnent accès directement aux informations linguistiques par des règles. Les autres approches sont aussi efficaces, mais elles mettent les informations linguistiques dans des tables de statistiques ou des réseaux neuronaux qui sont opaques à la lecture. Il est difficile de les analyser, de les comprendre et de les améliorer. L'approche de Brill nous donne accès à une représentation très claire et directe des informations linguistiques sans pour autant être moins efficace. Ramshaw et Marcus 1994 disent de cette approche quelle est :

> […] able to survey a larger space of possible contextual factors than could be practically captured by a statistical model that required explicit probability estimates for every possible combination of factors. Brill's results on part-of-speech tagging show that the method can outperform the HMM [hidden Markov Model] techniques widely used for that task, while also providing more compact and perspicuous models.

1.2.3 Catégoriseurs statistiques

Ce terme est utilisé pour parler des catégoriseurs qui utilisent des fréquences, des calculs de probabilités, etc. La forme la plus simple du catégoriseur statistique catégorise les mots en utilisant leurs catégories la plus fréquente dans un texte de référence. Évidemment, le problème avec cette approche est que, bien que le mot soit bien catégorisé dans la majorité des cas, il y aura tout de même une grande quantité d'erreurs. On en voit un exemple en (4), où chacun des mots de la phrase a reçu son étiquette la plus fréquente, mais la séquence d'étiquettes produite est inadéquate.

4. Phrase catégorisée à partir de l'étiquette la plus fréquente

Il/pronom **le/déterminant juge/verbe** sévèrement/adverbe.

Le mot *le* est plus fréquemment déterminant que pronom, il a donc été catégorisé déterminant, mais ce n'est pas exact puisque le mot *juge* est un verbe, et a été catégorisé comme tel puisque c'est sa catégorie la plus fréquente.

Une autre approche est de calculer la probabilité qu'une certaine séquence d'étiquettes apparaisse. Cette approche est souvent appelée approche par « n-grams ». Les algorithmes les plus connus pour implémenter une approche de ce type sont les algorithmes de Viterbi 1967. Ces algorithmes font une recherche qui élimine les possibilités infructueuses en calculant pour chaque cas de figure la probabilité maximale de n (où n est le nombre d'étiquettes possibles du mot suivant).

Une approche encore plus complexe, connue sous le nom de Modèle de Markov, combine les deux approches précédentes (voir Rabiner 1989). Elle utilise à la fois la probabilité qu'une séquence d'étiquettes apparaisse et la fréquence des étiquettes pour un mot. Le modèle de Markov traite le texte comme des états et des transitions. Un état est un mot auquel on a associé une étiquette temporaire ou permanente et une transition est un mot avec son étiquette suivie d'un mot et son étiquette. Dans le modèle de Markov, on considère que le choix d'une étiquette pour un mot doit dépendre des n étiquettes précédentes. Les logiciels conçus selon ce modèle calculent les probabilités transitionnelles (probabilité que l'étiquette x soit suivie de l'étiquette y), selon différentes formules. La plus utilisée est : $prob(x|y) = nbr(x \& y) / nbr(x)$. Cette formule peut être décodée comme suit : la probabilité que x soit suivi de y est égale au nombre de fois où x suit effectivement y dans le corpus, divisé par le nombre de fois où x apparaît.

Le modèle de Markov peut être visible ou invisible (hidden). Il est dit invisible (caché) quand on ne peut déterminer la séquence d'états à travers lesquels il a passé en observant seulement le résultat. On voit en (5) le résultat du modèle visible et en (6) le résultat du modèle caché.

5. Résultat du modèle de Markov visible

Il/pronom le/**déterminant/pronom** juge/verbe sévèrement/adverbe.

6. Résultat du modèle de Markov invisible

Il/pronom le/**pronom** juge/verbe sévèrement/adverbe.

Ce type de catégoriseurs nécessite d'énormes corpus (plusieurs millions de mots). Ils sont difficiles à améliorer vu le fait que les informations linguistiques se retrouvent dans d'énormes tableaux statistiques et non sous forme de règles. Ce type de catégoriseurs permet d'obtenir environ 96-97 % d'efficacité.

1.2.4 Catégoriseurs neuronaux

Pour bien comprendre comment fonctionne un catégoriseur neuronal, il faut d'abord savoir comment fonctionnent les réseaux neuronaux en général. Nous verrons donc d'abord ce qu'est un réseau neuronal et ensuite son fonctionnement.

Un réseau neuronal est un ensemble de neurones artificiels (voir la Figure 2). Les neurones sont interconnectés par des liens. Ces liens ont tous un poids, ce qui permet de modeler l'influence des neurones entre eux. La somme des poids des liens qui vont vers un neurone, doit dépasser un certain seuil, qui

varie d'un neurone à l'autre, pour que celui-ci émette une valeur de sortie. Cette valeur de sortie va à son tour influencer d'autres neurones.

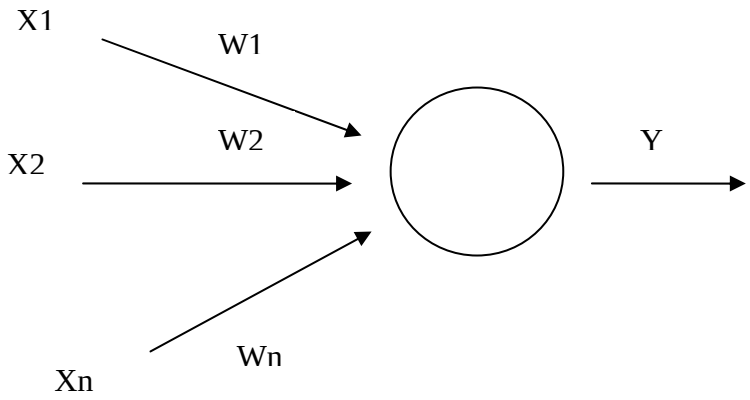

Figure 2 – **Neurone artificiel**

Un réseau neuronal comprend minimalement deux couches de neurones : des neurones d'entrée et des neurones de sortie (voir la figure ci-dessous). Les neurones d'entrée représentent les données fournies en entrée et les neurones de sortie le résultat.

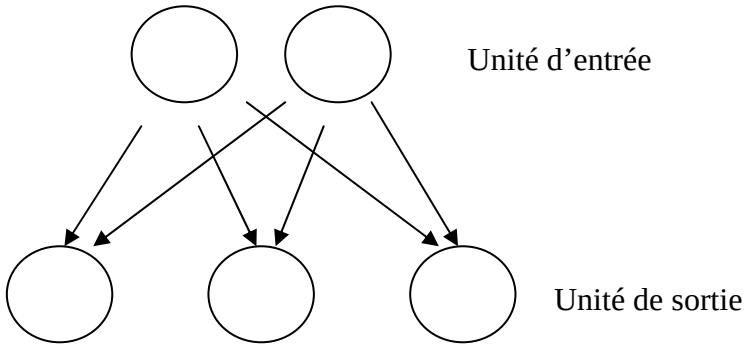

Figure 3 – **Réseau neuronal**

L'ensemble des neurones, ainsi interconnectés, peut accomplir des fonctions « intelligentes ». En effet, les réseaux neuronaux ont une grande capacité d'apprentissage. Pour l'apprentissage, des calculs sont effectués qui ont pour but de modifier le poids des entrées de chaque neurone. Les poids sont modifiés jusqu'à ce que le réseau neuronal donne en sortie le résultat désiré.

Les catégoriseurs neuronaux supervisés contiennent normalement un lexique et un réseau neuronal. Le réseau est d'abord entraîné à produire le résultat désiré, soit le corpus catégorisé servant d'exemple. Les étiquettes sont supprimées de ce texte et les mots du texte reçoivent l'étiquette qui apparaît le plus fréquemment pour un mot dans le corpus catégorisé. Puis le résultat obtenu est corrigé en modifiant le poids des entrées.

Dans un catégoriseur neuronal, les unités d'entrée sont : un neurone par catégorie possible pour le mot en traitement, des neurones qui représentent les mots précédents et leur catégorie (qui a déjà été assignée) et d'autres neurones qui représentent les mots suivants et leur catégories possibles (pas encore assignées) (voir la figure ci-dessous). Chaque unité d'entrée correspond donc à un mot avec une étiquette. Chaque unité de sortie correspond à un des éléments du jeu d'étiquettes.

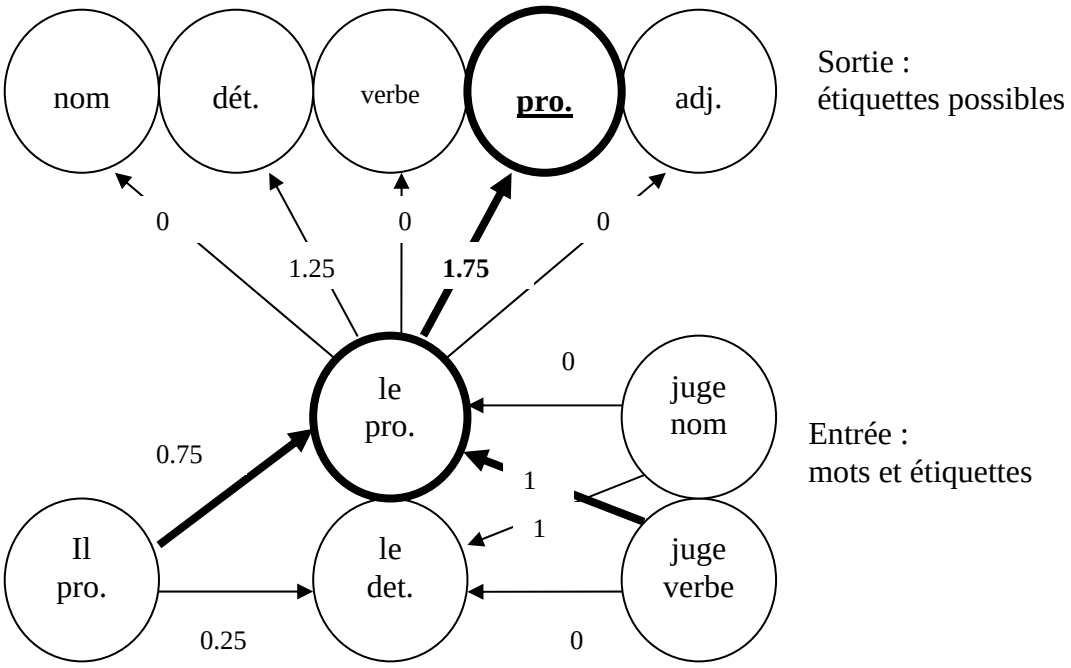

Figure 4 – **Catégoriseur neuronal**[2]

Le poids de chacun de ces éléments est ajusté jusqu'à ce que la combinaison idéale de poids, celle qui permettra la sélection du bon neurone de sortie, soit atteinte. La bonne unité de sortie (ou étiquette) sera donc sélectionnée en fonction des mots environnants et de leurs catégories et des liens plus ou moins déterminants qu'elle entretient avec ses divers éléments. Une fois le réseau neuronal entraîné, les bons poids sont fixés pour différents types de contextes et le catégoriseur qui en résulte peut catégoriser de nouveaux textes.

Le plus grand avantage des catégoriseurs neuronaux est qu'ils ne nécessitent pas de très gros corpus d'entraînement. Marques et Lopes 1996 rapportent une efficacité de 96 % pour le portugais pour un catégoriseur neuronal entraîné sur 30 pages de texte seulement. Toutefois, il est difficile de

[2] Les poids donnés dans cette figure ne sont là que pour les besoins de l'exemple. Dans les faits, les poids attribués par un réseau neuronal sont l'objet de calculs statistiques plus poussés que ceux que nous avons effectués ici.

perfectionner les logiciels obtenus puisque les réseaux neuronaux ne nous donnent pas accès à des représentations symboliques, à des tableaux de statistiques ou encore à des règles, mais uniquement à un réseau. Bien comprendre toutes les interactions entre les divers éléments est donc extrêmement difficile.

1.3 Les mots inconnus

Un problème important pour toutes les approches discutées jusqu'à maintenant est celui des mots inconnus. La principale raison pour laquelle les mots inconnus sont difficiles à traiter est leur très grande ambiguïté. Théoriquement, un mot inconnu peut appartenir à toutes les catégories, contrairement aux mots connus qui ne peuvent appartenir qu'à quelques-unes d'entre elles. Certains catégoriseurs à base de règles peuvent étiqueter les mots inconnus. Vasilakopoulos 2003, par exemple, rapporte un taux d'efficacité de 87 %, pour l'anglais. Pour ce qui est des catégoriseurs statistiques, leur taux efficacité est d'environ 85 % pour l'anglais. Dans le cas des catégoriseurs neuronaux, Nakagawa et al. 2001 rapportent une efficacité de 88,5 % pour l'anglais. Encore une fois, les résultats sont donc sensiblement les mêmes, peu importe l'approche utilisée.

Évidemment, plus le corpus d'entraînement est gros, moins il y a de mots inconnus. On pourrait donc penser qu'il suffit d'utiliser le plus gros corpus d'entraînement possible afin de minimiser ou d'éliminer le problème. Ce n'est toutefois pas aussi simple. Il y a une énorme quantité de noms propres, de mots étrangers, de mots rares, de mots de spécialité (termes), de sigles, de néologismes et de mots mal orthographiés qui ne peuvent pas tous se retrouver dans un corpus donné. De plus, le plus gros désavantage des approches supervisées vient du fait qu'elles nécessitent de gros corpus d'entraînement qui sont coûteux et qui ne sont pas disponibles pour toutes

les langues et tous les domaines. L'amélioration du traitement des mots inconnus rendrait les approches supervisées beaucoup plus polyvalentes.

Pour déterminer la nature d'un mot inconnu, les différents modèles de catégorisation grammaticale automatique utilisent la morphologie et le contexte. Mais, la tâche n'est pas simple. Les règles utilisant le contexte sont très efficaces pour traiter les mots connus, chaque mot connu ne pouvant appartenir qu'à un nombre restreint de catégories. Mais il est difficile de tirer profit du contexte pour catégoriser les mots inconnus puisqu'ils peuvent théoriquement appartenir à toutes les catégories. Il est difficile de définir tous les contextes dans lesquels peut apparaître un type de mot et différents types de mots peuvent apparaître dans des contextes similaires. L'utilisation de la morphologie est donc beaucoup plus efficace pour le traitement des mots inconnus car, elle permet de restreindre le nombre de catégories possibles pour un mot donné. L'utilisation du contexte est donc reléguée au second plan.

Il est à noter que les mots inconnus sont majoritairement des mots pleins puisque les mots outils font partie de listes finies. Mais, la seule présence d'un morphème n'est souvent pas suffisante pour étiqueter un mot plein. Les substantifs, les adjectifs, les verbes et les participes sont souvent constitués de morphèmes ambigus, c'est-à-dire de morphèmes qui peuvent s'associer à des mots de différentes natures. C'est le cas des préfixes (ex : *re-construire* : verbe, *re-construit* : participe, *re-constructible* : adjectif, etc.). Les suffixes lexicaux peuvent eux-aussi être associés à des mots de différentes natures (ex : *mécan-ique* : nom/adjectif). Les suffixes grammaticaux aussi (ex : *étudiant(es)* : nom, *aimable(s)* : adjectif). Toutefois, ces trois types d'affixes ont chacun des caractéristiques propres desquelles on peut tirer profit. Les préfixes ne modifient jamais la catégorie du mot auquel ils s'adjoignent (ex : *construire* : verbe, *re-construire* : verbe). Donc, leur présence ou leur absence

n'a pas d'impact direct sur la catégorie des mots. Les suffixes grammaticaux (genre, nombre, temps, personne et mode) ne modifient jamais la catégorie des mots. Donc, leur présence ou leur absence n'a pas d'impact direct sur la catégorie des mots. C'est pourquoi on peut donner à *étudiant, étudiants, étudiante et étudiantes* la même catégorie et il en est de même pour *joue*, *jouerais*, *jouerai*, etc. Les suffixes lexicaux quant à eux modifient la catégorie des mots auxquels ils s'adjoignent. Les mots formés à partir de certains d'entre eux sont ambigus (ex : *-ique* peut donner lieu à un nom ou à un adjectif) et d'autres non (ex. : *-emment* : adverbe). Un même suffixe lexical, donnera toujours lieu à des mots de mêmes catégories. Puisque les suffixes lexicaux modifient la catégorie des mots auxquels ils s'adjoignent, la catégorie de ceux-ci est prévisible à partir du suffixe. C'est-à-dire qu'un mot se terminant par *-emment* sera toujours un adverbe et un mot se terminant par *-ique* un nom ou un adjectif. On peut donc associer des catégories grammaticales aux suffixes lexicaux.

Enfin, les sigles, les noms propres, les mots étrangers, les abréviations et certains adverbes ne se terminant pas en *-ment* (ex. : surtout) ne possèdent pas de morphèmes caractéristiques. Il n'y a donc pas d'indice, à part les majuscules dans le cas des noms propres et des sigles, qui permette de les catégoriser. Il faut donc tirer profit du contexte dans ces cas.

1.4 Le catégoriseur de Brill

Le logiciel de Brill atteint une performance d'environ 97 % pour l'anglais contre environ 94 % pour le français. L'approche est de type supervisée à base de règles, mais elle utilise aussi des informations statistiques, soit la fréquence. Pour créer son approche, Brill (1993, 1994, 1995b) s'est inspiré des écrits de Zellig Harris qui a formulé plusieurs règles pour l'analyse de langues inconnues. Toutefois, Brill n'a retenu que peu des règles formulées

par Harris, puisque l'analyse de la langue que peut faire un linguiste diffère nécessairement de celle que peut faire un ordinateur. Mais le catégoriseur de Brill fait tout de même une analyse linguistique d'une langue qu'il ne connaît pas et ne se base que sur les faits de surface, tout comme le conseillait Zellig Harris.

Nous verrons, dans les sections suivantes le détail du fonctionnement de l'approche de Brill. En bref (voir la figure ci-dessous), il faut en entrée un corpus préalablement catégorisé. À partir de ce texte, un dictionnaire est créé ainsi que deux types de règles : contextuelles et lexicales. Les règles contextuelles vont servir à catégoriser les mots connus ambigus selon le contexte. Les règles lexicales vont servir à catégoriser les mots inconnus d'après leur morphologie et un contexte très restreint. Pour créer ces règles, le logiciel utilise une copie du corpus duquel les catégories ont été enlevées. Pour la création des règles contextuelles, les mots de ce corpus recevront dans un premier temps leur catégorie la plus fréquente tandis que pour la création des règles lexicales, les mots seront d'abord identifiés comme nom ou nom propre dépendant s'ils commencent par une majuscule ou une minuscule. Ensuite, le logiciel tente de corriger le résultat ainsi obtenu en créant des règles. Il compare sa catégorisation avec celle du corpus précatégorisé. À la première erreur trouvée, il crée plusieurs règles. Puis, il applique chacune de ces règles à tout le texte pour voir si les règles créées sont généralisables. La règle qui réduit le plus le nombre d'erreurs est enregistrée. Puis l'entraînement continu pour corriger la deuxième erreur.

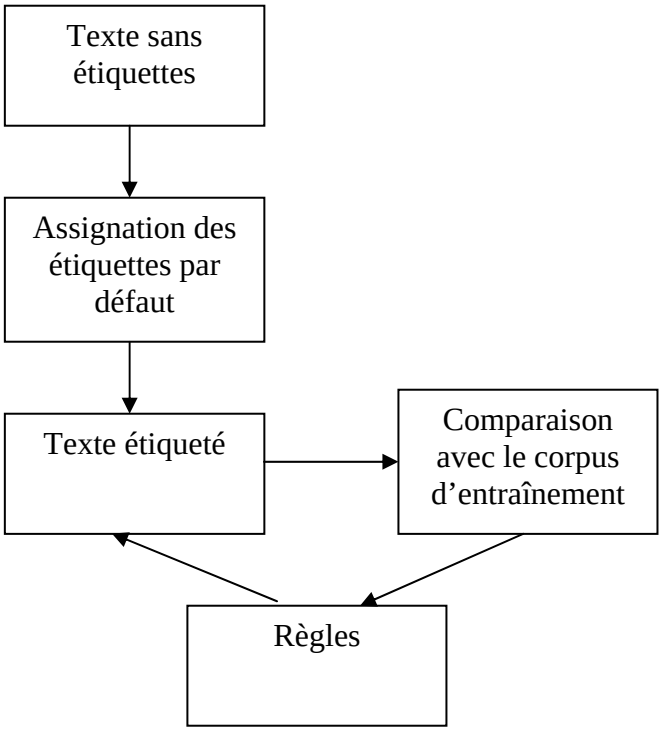

Figure 5 – **Fonctionnement des modules d'apprentissage de l'approche de Brill**

1.4.1 Les modules d'apprentissage

Pour la phase d'apprentissage, il faut un texte précatégorisé (voir l'exemple en (7)) d'une centaine de pages.

7. Phrase catégorisée[3]

La/DTN:sg pédagogie/SBC:sg est/ECJ:sg une/DTN:sg oeuvre/SBC:sg de/PREP coordination/SBC:sg et/COO de/PREP rapports/SBC:pl ;/;

Un dictionnaire est créé à partir des mots qui se trouvent dans le texte et des différentes catégories qui leur sont attachées. Il place la catégorie qui se retrouve le plus souvent pour un même mot en tête de liste dans le dictionnaire (voir en (8)). Les autres catégories possibles sont placées à la suite de celle-ci sans ordre précis.

8. Contenu du dictionnaire[4]

donnant VNCNT
donne VCJ:sg
donnent VCJ:pl
donner VNCFF
donné PAR:sg
donnée PAR:sg SBC:sg
données SBC:pl PAR:pl
dont REL
dossier SBC:sg
dote VCJ:sg
doter VNCFF
double SBC:sg ADJ:sg
doublement SBC:sg
doute SBC:sg
dramatisations SBC:pl
drapière ADJ:sg
drogue SBC:sg
droit SBC:sg

Par la suite, des règles contextuelles sont créées (voir en (9)) qui serviront à déterminer la bonne catégorie pour un mot ambigu présent au dictionnaire d'après le contexte.

---

[3] Voir la liste des étiquettes utilisées en annexe A
[4] Un même mot écrit en majuscules ou en minuscules sera l'objet de deux entrées dans le dictionnaire.

9. Exemples de règles contextuelles

- DTN:sg PRV:sg NEXTTAG VCJ:sg
  - Si le mot est catégorisé déterminant singulier (sg.), qu'il présente une ambiguïté déterminant sg./pronom verbal sg. et que le mot suivant est catégorisé verbe conjugué sg., changer son étiquette pour pronom verbal sg. (ex. : je ***le*** juge).

- SBC:sg ADJ:sg PREVTAG SBC:sg
  - Si le mot est catégorisé substantif sg., qu'il présente une ambiguïté substantif sg./adjectif sg. et que le mot précédent est catégorisé substantif sg., changer son étiquette pour adjectif sg. (ex. : la maison ***or***).

- PRV:sg PRO:sg WDPREVTAG PREP elle
  - Si le mot est catégorisé pronom verbal sg., qu'il présente une ambiguïté pronom verbal sg. / pronom sg. et que le mot précédent est *elle* ou une préposition, changer son étiquette pour pronom sg. (ex. : on parle de ***lui***).

Pour ce faire, il crée d'abord une version non étiquetée du corpus catégorisé ou de référence et donne à tous les mots de ce texte leur catégorie la plus fréquente, soit la catégorie en tête de liste dans le dictionnaire (voir en (10)).

10. Catégorisation de départ

Hier/ADV ,/, j'/PRV:sg ai/ACJ:sg rencontré/VPAR:sg le/DTN:sg **juge/VCJ:sg** ./.

Puis il compare le résultat obtenu avec le corpus de référence et tente de corriger les erreurs en créant des règles, à partir d'une liste de modèles de règles (voir en (11)). Le domaine d'une règle est la phrase, et dans cette phrase, un contexte de trois mots avant et de trois mots après est considéré. Les mots d'une autre phrase ne seront pas considérés même s'ils apparaissent à moins de trois mots de distance.

11. Modèles de règles contextuelles

Changer l'étiquette de Z à W (W étant une catégorie possible pour ce mot dans le dictionnaire) si :

PREVBIGRAM	Le bigramme[5] précédent est X.
NEXTBIGRAM	Le bigramme suivant est X.
NEXT1OR2OR3TAG	1 des 3 mots suivants est catégorisé X.
NEXT1OR2TAG	1 des 2 mots suivants est catégorisé X.
NEXTTAG	Le mot suivant est catégorisé X.
NEXT2TAG	Le 2e mot suivant est catégorisé X.
SURROUNDTAG	Le mot précédent est catégorisé X et le mot suivant est catégorisé Y.
PREV1OR2OR3TAG	1 des 3 mots précédents est catégorisé X.
PREV1OR2TAG	1 des 2 mots précédents est catégorisé X.
PREVTAG	Le mot précédent est catégorisé X.
PREV2TAG	Le 2e mot précédent est catégorisé X.
NEXT1OR2WD	1 des deux mots suivants est X.
NEXT2WD	Le 2e mot suivant est X.
NEXTWD	Le mot suivant est X.
CURWD	Le mot est X.
PREV1OR2WD	1 des deux mots précédents est X.
PREV2WD	Le 2e mot précédent est X.
PREVWD	Le mot précédent est X.
WDAND2TAGAFT	Le 2e mot suivant est catégorisé X et le mot suivant est Y.
WDAND2TAGBFR	Le 2e mot précédent est catégorisé X et le mot précédent est Y.
WDNEXTTAG	Le mot suivant est X et le 2e mot suivant est catégorisé Y.
WDPREVTAG	Le mot précédent est X et le 2e mot précédent est catégorisé Y.

Donc, à la première erreur rencontrée, des règles sont créées. Si le mot en traitement est *juge* catégorisé verbe conjugué singulier plutôt que nom singulier dans la phrase *j'ai rencontré le juge*, une tentative sera faite pour instancier les variables contenues dans les modèles de règles à partir du contexte. La règle suivante sera donc créée : changer l'étiquette de verbe

[5] Deux mots se retrouvant côte à côte.

conjugué singulier à nom singulier (nom singulier étant une catégorie possible pour ce mot dans le dictionnaire) si le mot précédent est catégorisé déterminant singulier. D'autres règles seront aussi créées comme par exemple : changer l'étiquette de verbe conjugué singulier à nom singulier (nom singulier étant une catégorie possible pour ce mot dans le dictionnaire) si le 2e mot précédent est catégorisé participe passé singulier. Puis les règles créées seront appliquées à tout le texte, pour déterminer si elles sont des généralisations valables. Les résultats obtenus sont comparés avec le corpus de référence. La règle qui a permis d'améliorer le plus les résultats est enregistrée en tête de liste et les autres sont écartées (voir la figure ci-dessous). L'apprentissage se poursuit pour corriger la 2e erreur à partir du texte résultant de l'application de la 1ère règle trouvée. L'apprentissage se termine quand il n'est plus possible de créer de règle permettant de corriger un nombre d'erreur donné. Ce nombre est fixé à trois par défaut.

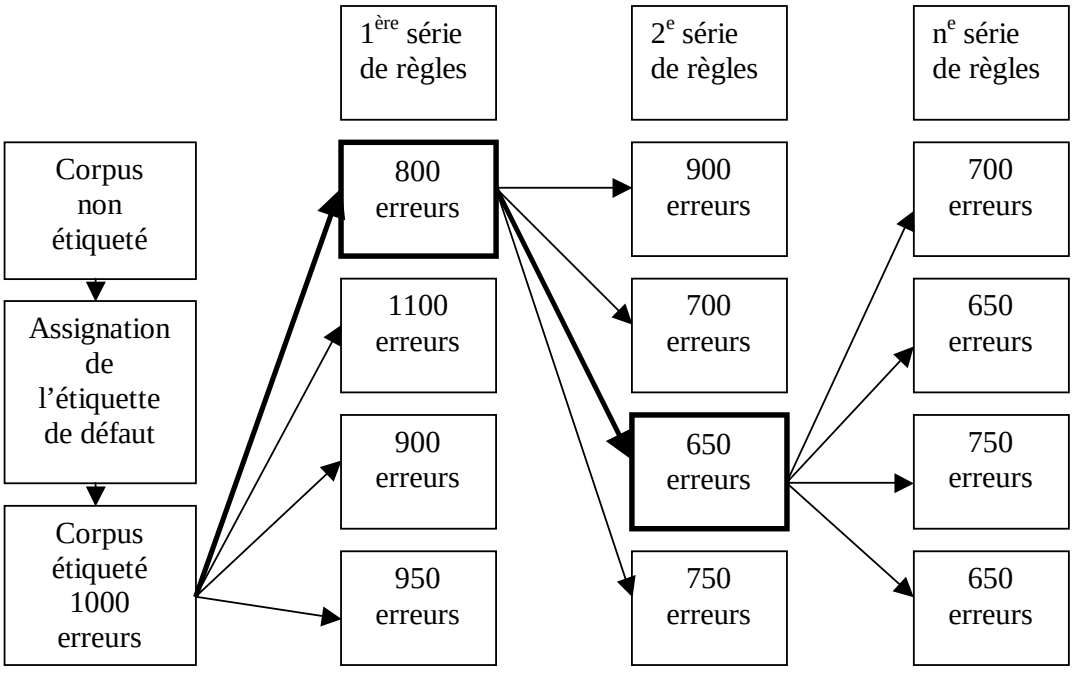

Figure 6 – **Processus de sélection des règles**

La première règle trouvée sera celle qui sera appliquée en premier lors de la catégorisation de nouveaux textes. Au fur et à mesure de l'apprentissage des règles, il subsiste de moins en moins d'erreurs. Les dernières règles trouvées sont donc peu productives, elles ne permettent pas de corriger beaucoup d'erreurs, tandis que les premières règles trouvées permettent d'en corriger beaucoup.

Des règles lexicales sont aussi créées (voir en (12)). Elles serviront à catégoriser les mots absents du dictionnaire.

12. Exemples de règles lexicales

- SBC:sg e fdeletesuf 1 ADJ:sg
  - Si le mot est catégorisé substantif sg. et qu'en enlevant le suffixe *e* on obtient un mot existant, changer son étiquette pour adjectif sg.[6]

- SBC:pl nous fgoodright VCJ:pl
  - Si le mot est catégorisé substantif pluriel (pl.) et qu'il est à droite de *nous*, changer son étiquette pour verbe conjugué pl.

- SBC:pl ais fhassuf 3 VCJ:sg
  - Si le mot est catégorisé substantif pl. et qu'il comporte le suffixe *-ais*, changer son étiquette pour verbe conjugué sg.

Pour créer ces règles, le logiciel utilise des données concernant deux corpus : le corpus de référence et un corpus non catégorisé[7] aussi gros que possible. Il faut au logiciel la liste des mots contenus dans le texte catégorisé avec la fréquence pour chacune des catégories qui leur sont associés (voir en (13)). Cette liste servira à déterminer la bonne catégorie pour un mot. Il faut aussi la liste des bigrammes contenus dans le corpus non étiqueté en ordre inverse

[6] Le mot présente une variation en genre, c'est donc un adjectif.
[7] Ce corpus peut comprendre ou non le corpus de référence.

de fréquence (voir en (14)) qui servira à créer des règles lexicales utilisant le contexte. Puis il faut une liste de tous les mots figurant dans le corpus non étiqueté en ordre inverse de fréquence (voir en (15)). Cette liste servira à vérifier des opérations morphologiques comme l'ajout et la suppression de préfixes et de suffixes.

13. Extrait de la liste « mot - catégorie – fréquence » tiré du corpus catégorisé

même ADJ:sg 63	où REL 58	Mais COO 51
y PRV:++ 63	droit SBC:sg 57	donc ADV 50
se PRV:++ 62	en PRV:++ 57	C' PRV:sg 49
bien ADV 60	été EPAR:sg 57	nombre SBC:sg 49
deux CAR 60	monde SBC:sg 55	Dans PREP 48
encore ADV 60	siècle SBC:sg 54	En PREP 47
éducation SBC:sg 60	dont REL 53	France SBP:sg 47
leur DTN:sg 58		

14. Extrait de la liste « bigramme » tiré du corpus non catégorisé

cas allemand	cas graves	cas particuliers
cas apparentés	cas général	cas précis
cas avait	cas historique	cas psychopathologique
cas c'	cas homogènes	cas qu'
cas certains	cas ici	cas quant
cas cité	cas il	

15. Extrait de la liste « mot » tiré du corpus non catégorisé

La	s'	ont
qu'	L'	être
n'	se	elle
Les	avec	cette
comme	peut	entre
Le	son	l'
des	Il	ce
aux	mais	

À partir de ces données, des règles seront formulées permettant d'assigner la bonne catégorie aux mots du corpus. Tous les mots du corpus de référence reçoivent comme catégorie nom singulier s'ils commencent par une minuscule et nom propre singulier s'ils commencent par une majuscule. Puis le résultat obtenu est comparé avec la liste mot-catégorie-fréquence et, pour corriger les erreurs, des règles sont créées à partir d'une liste de modèles de règles (voir en (16)). Les règles créées seront appliquées à tous les mots du texte. Les résultats obtenus sont comparés avec la catégorisation du corpus de référence. La règle qui a permis d'améliorer le plus les résultats est enregistrée avec son score, les autres sont écartées et l'apprentissage se poursuit, pour corriger les autres erreurs. Une fois l'apprentissage terminé les règles seront placées en ordre inverse de score. La règle la plus productive en premier et la règle la moins productive en dernier.

Notons que lors de l'apprentissage des règles lexicales, tous les mots du texte sont utilisés. Or, dans les textes qui devront être catégorisés par la suite, ce ne sont pas tous les mots du texte qui seront traités par les règles lexicales, mais seulement ceux qui ne sont pas dans le dictionnaire.

16. Modèles de règles lexicales

Changer l'étiquette de W à X si :

deletepref	Enlever le préfixe Y donne un mot existant ($Y \leq 4$)[8].
fdeletepref	Le mot est catégorisé Z et enlever le préfixe Y donne un mot existant[9] ($Y \leq 4$).
deletesuf	Enlever le suffixe Y donne un mot existant ($Y \leq 4$).
fdeletesuf	Le mot est catégorisé Z et enlever le suffixe Y donne un mot existant ($Y \leq 4$).
addsuf	Ajouter le suffixe Y donne un mot existant ($Y \leq 4$).
faddsuf	Le mot est catégorisé Z et ajouter le suffixe Y donne un mot existant ($Y \leq 4$).
addpref	Ajouter le préfixe Y donne un mot existant ($Y \leq 4$).
faddpref	Le mot est catégorisé Z et ajouter le préfixe Y donne un mot existant ($Y \leq 4$).
haspref	Les premiers caractères du mot sont Y ($Y \leq 4$).
fhaspref	Le mot est catégorisé Z et les premiers caractères du mot sont Y ($Y \leq 4$).
hassuf	Les derniers caractères du mot sont Y ($Y \leq 4$).
fhassuf	Le mot est catégorisé Z et les derniers caractères du mot sont Y ($Y \leq 4$).
char	Le caractère Y apparaît dans le mot.
fchar	Le mot est catégorisé Z et le caractère Y apparaît dans le mot.
goodright	Le mot Z apparaît immédiatement à droite dans la liste des bigrammes.
fgoodright	Le mot est catégorisé Y et le mot Z apparaît immédiatement à droite dans la liste des bigrammes.
goodleft	Le mot Z apparaît immédiatement à gauche dans la liste des bigrammes.
fgoodleft	Le mot est catégorisé Y et le mot Z apparaît immédiatement à gauche dans la liste des bigrammes.

Prenons le mot *rediriger*. Les variables contenues dans les modèles de règles seront instanciées à partir des lettres contenues dans le mot. La catégorie de *rediriger*, verbe, est connue à partir de la liste tirée du corpus catégorisé. Des

[8] ($Y \leq 4$) signifie que le morphème considéré sera d'au maximum quatre lettres.
[9] Lors de l'apprentissage, la recherche des mots existants se fait dans la liste de mots tirée du grand corpus. Lors de la catégorisation de nouveaux textes, la recherche se fait dans le dictionnaire.

règles permettant de modifier la catégorie de défaut, nom, pour verbe seront donc créées. Une tentative d'instanciation de la première règle sera effectuée : enlever le préfixe Y donne un mot ($Y \leq 4$). Tour à tour seront enlevées la première lettre du mot, puis la deuxième, puis la troisième et la quatrième. À chaque fois, le logiciel vérifie si le reliquat existe dans la liste de mot tiré du corpus non catégorisé. *-ediriger* ne sera pas trouvé, ni *-iriger*, ni *-riger*, mais *-diriger* sera trouvé. Les variables du premier modèle de règle seront donc instanciées, ce qui donnera : changer l'étiquette de nom à verbe si enlever le préfixe *re-* donne un mot. Puis il en sera de même pour tous les modèles de règles. D'autres règles seront ainsi créées, comme par exemple : changer l'étiquette de nom à verbe si enlever le suffixe *-r* donne un mot (*redirige*). Puis chacune des règles créées sera appliquée à tous les mots, une évaluation de la performance de chacune des règles sera effectuée, et la règle la plus efficace, celle qui permet d'éliminer le plus d'erreurs, sera enregistrée. L'apprentissage continuera à partir du corpus résultant de l'application de la première règle trouvée. Une fois l'apprentissage terminé, tout ce qui est nécessaire pour catégoriser de nouveaux textes est disponible.

Il est à noter que l'approche, dans son état actuel, n'est pas en mesure de résoudre l'ambiguïté de certains affixes. Le suffixe *-er*, par exemple, peut être associé à un nom ou à un verbe à l'infinitif. Si les deux règles suivantes sont créées : un mot se terminant par *-er* est un nom ; un mot se terminant par *-er* est un verbe à l'infinitif. Celle qui identifie les mots se terminant par *-er* comme des verbes à l'infinitif sera conservée vu le nombre et la fréquence des verbes à l'infinitif se terminant par *-er*. L'autre règle sera écartée parce qu'elle cause plus d'erreurs qu'elle n'en résout. En effet, cette règle aurait pour effet de catégoriser tous les mots se terminant par *-er* comme des substantifs. Or, la majorité d'entre eux sont des verbes à l'infinitif.

1.4.2 Le module de catégorisation

À partir du dictionnaire et des règles contextuelles et lexicales créés, un autre module catégorisera de nouveaux textes. On donne au catégoriseur un texte non catégorisé en entrée. Dans un premier temps, il le découpe en mots et en phrases (voir en (17)). Seront considérés comme des mots les caractères apparaissant entre deux séparateurs (' .,;:_+=-#|\!"/£$¢%¤?¬&¦*()[]«»<>'"”), sauf *aujourd'hui* qui est toujours considéré comme un seul mot et non deux. Le logiciel détermine qu'il y a fin de phrase quand il rencontre un point (.), un point d'exclamation (!) ou un point d'interrogation (?) suivi d'une majuscule. Les mots seront tous séparés par des espaces et les phrases par des retours de chariot.

17. Découpage du texte

> entrée : L'essentiel à retenir sur les lagunes du nord-ouest de l'Adriatique est qu'elles sont une immensité d'eau saumâtre, de sable et de boue.
>
> sortie : L' essentiel à retenir sur les lagunes du nord - ouest de l' Adriatique est qu' elles sont une immensité d' eau saumâtre , de sable et de boue .¶

Une fois le découpage fait, la catégorisation du texte commence. Tout comme lors de la phase d'apprentissage, chacun des mots du texte reçoit une catégorie de défaut, soit la catégorie en tête de liste dans le dictionnaire pour les mots connus et NN (nom commun) pour les mots inconnus commençant par une minuscule et NNP[10] (nom propre) pour les mots inconnus commençant par une majuscule. Les signes de ponctuations ne reçoivent pas d'étiquette grammaticale, ils sont simplement dédoublés. On voit en (18) le

[10] Les étiquettes NN et NNP sont spécifiques aux mots inconnus, c'est-à-dire que l'étiquette normalement utilisée pour nom commun est SBC :sg/pl et pour nom propre est SBP :sg/pl.

résultat de l'étiquetage initial. Les mots inconnus sont encadrés et les mots ambigus sont surlignés.

18. Étiquetage initial

> entrée : L' essentiel à retenir sur les lagunes du nord - ouest de l' Adriatique est qu' elles sont une immensité d' eau saumâtre , de sable et de boue .¶
>
> sortie : L'/DTN :sg essentiel/ADJ :sg à/PREP retenir/VNCFF sur/PREP les/DTN :pl lagunes/NN du/DTC :sg nord/SBC :sg -/- ouest/SBC :sg de/PREP l'/DTN :sg[11] Adriatique/NNP est/ECJ :sg[12] qu'/SUB elles/PRV :pl sont/ECJ :pl une/DTN :sg immensité/SBC :sg d'/PREP eau/SBC :sg saumâtre/NN ,/, de/PREP sable/SBC :sg et/COO de/PREP boue/SBC :sg ./.¶

Dans cette phrase, les mots *lagunes*, *Adriatique* et *saumâtre* sont inconnus. Les autres mots sont connus, mais certains sont ambigus : *essentiel* (adjectif singulier, substantif singulier), *l'* (déterminant sg., pronom verbal sg., particule attachée), *qu'* (subordonnant, pronom singulier), *elles* (pronom verbal pluriel, pronom pluriel), *une* (déterminant sg., pronom sg.) et *d'* (préposition, déterminant sg.).

Les mots connus ambigus sont traités par les règles contextuelles, alors que les mots inconnus sont traités par les règles lexicales. Les mots sont traités un à un et pour chacun des mots, l'application de chacune des règles (lexicales ou contextuelles) sera tentée. La règle ayant le meilleur score sera la première à être appliquée puisqu'elle est la plus générale. La dernière règle appliquée, celle ayant le plus faible score, corrigera peu d'erreurs. Si la règle

---

[11] *L'* et *l'* sont traités comme deux mots distincts. *L'* n'est pas ambigu parce qu'il ne se retrouvait qu'avec une seule catégorie dans le corpus d'entraînement alors que *l'* est ambigu parque ce qu'on le retrouvait avec plusieurs catégories dans le corpus d'entraînement.
[12] Le mot *est* aurait été ambigu s'il s'était retrouvé avec la catégorie substantif singulier (point cardinal) dans le corpus d'entraînement.

peut être appliquée au mot, la catégorie du mot est modifiée. Puis, que le mot ait changé de catégorie ou non, on passe à la deuxième règle la plus productive et ainsi de suite, jusqu'à la fin des règles. Ainsi, un même mot peut changer plusieurs fois de catégorie pendant la phase d'assignation de catégorie.

Le premier mot traité sera le mot *essentiel* qui est ambigu adjectif singulier/substantif singulier. Le logiciel va tenter d'appliquer toutes les règles contextuelles, une à une, à ce mot. On peut voir les premières règles contextuelles en (19).

19. Premières règles contextuelles

- PRO:sg DTN:sg NEXT1OR2TAG SBC:sg
  - Si le mot est catégorisé pronom sg., qu'il présente une ambiguïté pronom sg./déterminant sg. et qu'il est suivi d'un substantif sg. à un ou deux mots de distance, changer sont étiquette pour déterminant sg.
- SBC:sg ADJ:sg PREVTAG SBC:sg
  - Si le mot est catégorisé substantif sg., qu'il présente une ambiguïté substantif sg./adjectif sg.[13] et qu'il est précédé d'un substantif sg, changer son étiquette pour adjectif sg.
- DTN:sg PUL NEXT1OR2WD on
  - Si le mot est catégorisé déterminant sg., qu'il présente une ambiguïté déterminant sg/particule attachée et qu'il est suivi de *on* à un ou deux mots de distance, changer sont étiquette pour particule attachée.
- ADJ:sg SBC:sg PREVTAG DTN:sg
  - Si le mot est catégorisé adjectif sg., qu'il présente une ambiguïté adjectif sg./substantif sg. et qu'il est précédé d'un déterminant sg., changer son étiquette pour substantif sg.

[13] l'ambiguïté nom/adj. et adj./nom ne sont pas équivalentes.

Le logiciel va tenter d'appliquer la première règle à *essentiel*. Ça ne fonctionnera pas puisque *essentiel* n'est pas catégorisé pronom singulier, mais bien adjectif singulier, pour le moment. Puis, il en sera de même pour la deuxième et la troisième règle. La quatrième règle qui spécifie qu'un mot catégorisé adjectif singulier (c'est le cas), qui présente une ambiguïté adjectif singulier/substantif singulier (c'est aussi le cas), est un substantif singulier s'il est précédé d'un déterminant singulier (c'est encore le cas) pourra être appliquée. *Essentiel* changera donc de catégorie pour devenir un substantif singulier (voir en (20)). Le logiciel tentera d'appliquer toutes les autres règles contextuelles, mais sans succès. *Essentiel* demeurera donc un substantif singulier.

20. Traitement du premier mot ambigu

entrée : L'/DTN :sg essentiel/ADJ :sg à/PREP retenir/VNCFF sur/PREP les/DTN :pl lagunes/NN du/DTC :sg nord/SBC :sg -/- ouest/SBC :sg de/PREP l'/DTN :sg Adriatique/NNP est/ECJ :sg qu'/SUB elles/PRV :pl sont/ECJ :pl une/DTN :sg immensité/SBC :sg d'/PREP eau/SBC :sg saumâtre/NN ,/, de/PREP sable/SBC :sg et/COO de/PREP boue/SBC :sg ./.¶

sortie : L'/DTN :sg essentiel/SBC :sg à/PREP retenir/VNCFF sur/PREP les/DTN :pl lagunes/NN du/DTC :sg nord/SBC :sg -/- ouest/SBC :sg de/PREP l'/DTN :sg Adriatique/NNP est/ECJ :sg qu'/SUB elles/PRV :pl sont/ECJ :pl une/DTN :sg immensité/SBC :sg d'/PREP eau/SBC :sg saumâtre/NN ,/, de/PREP sable/SBC :sg et/COO de/PREP boue/SBC :sg ./.¶

Le deuxième mot traité sera le mot inconnu *lagunes*. Le logiciel tentera d'appliquer toutes les règles lexicales à ce mot. Regardons de plus près les premières règles lexicales.

21. Premières règles lexicales

- e char SBC:sg
  - Si le mot contient la lettre *e*, changer sont étiquette pour substantif sg.
- s hassuf 1 SBC:pl
  - Si le mot comporte le suffixe *-s*, changer son étiquette pour substantif pl.
- e addsuf 1 PAR:sg
  - Si on peut ajouter le suffixe *-e* au mot et le retrouver au dictionnaire, changer son étiquette pour participe passé sg.
- er hassuf 2 VNCFF
  - Si le mot comporte le suffixe *-er*, changer son étiquette pour verbe à l'infinitif.
- SBC:sg ment faddsuf 4 ADJ:sg
  - Si le mot est catégorisé substantif sg. et qu'on peut lui ajouter le suffixe *-ment* et le retrouver au dictionnaire, changer son étiquette pour adjectif sg.

Il est à noter qu'une règle prise en isolation peut sembler fautive sans pour autant l'être, les règles étant appliquées en série. Le logiciel tentera d'appliquer tour à tour les règles présentées en (21). La première d'entre elles qui spécifie qu'un mot contenant la lettre *e* est un substantif singulier sera appliquée puisque le mot en traitement contient bel et bien un *e*. Le statut de substantif singulier de *lagunes* est donc confirmé temporairement. Puis, la deuxième règle, qui spécifie qu'un mot ayant le suffixe *-s* est un substantif pluriel, sera appliquée puisque *lagunes* se termine par un *-s*. Le mot perdra donc son étiquette de substantif singulier pour devenir un substantif pluriel. Le logiciel tentera ensuite d'appliquer la troisième règle qui spécifie qu'un mot auquel on peut ajouter le suffixe *-e* et le retrouver au dictionnaire est un participe passé singulier. Toutefois, cette règle ne sera pas appliquée puisque le mot *lagunes-e* ne sera pas trouvé au dictionnaire. Il en est de même pour

toutes les autres règles dans ce cas-ci. On voit le résultat du traitement du mot *lagunes* en (22).

22. Traitement du deuxième mot

entrée : L'/DTN :sg essentiel/SBC :sg à/PREP retenir/VNCFF sur/PREP les/DTN :pl lagunes/NN du/DTC :sg nord/SBC :sg -/- ouest/SBC :sg de/PREP l'/DTN :sg Adriatique/NNP est/ECJ :sg qu'/SUB elles/PRV :pl sont/ECJ :pl une/DTN :sg immensité/SBC :sg d'/PREP eau/SBC :sg saumâtre/NN ,/, de/PREP sable/SBC :sg et/COO de/PREP boue/SBC :sg ./.¶

sortie : L'/DTN :sg essentiel/SBC :sg à/PREP retenir/VNCFF sur/PREP les/DTN :pl lagunes/SBC :pl du/DTC :sg nord/SBC :sg -/- ouest/SBC :sg de/PREP l'/DTN :sg Adriatique/NNP est/ECJ :sg qu'/SUB elles/PRV :pl sont/ECJ :pl une/DTN :sg immensité/SBC :sg d'/PREP eau/SBC :sg saumâtre/NN ,/, de/PREP sable/SBC :sg et/COO de/PREP boue/SBC :sg ./.¶

Le logiciel traitera ainsi tour à tour tous les autres mots ambigus ou inconnus.

Chapitre 2 Méthodologie

2.1 Objectifs

Les résultats d'abord obtenus par Brill pour l'anglais étaient d'environ 90% d'efficacité globale et de 70 % d'efficacité pour les règles lexicales. Plus tard (Brill 1994), l'ajout de modèles de règles et l'utilisation de plus gros corpus d'entraînement (plusieurs milliers de pages) ont permis d'améliorer la performace globale jusqu'à environ 97 % et la performance des règles lexicales jusqu'à 82 %. L'utilisation de jeux d'étiquettes restreints et l'ajout de dictionnaires indépendants améliorent aussi les performances du logiciel. L'INALF (voir Lecompte 1998) a entraîné le catégoriseur de Brill pour le français. Les résultats obtenus tournent autour de 94 % d'efficacité, soit 3 % de moins que les résultats obtenus pour l'anglais. Cette diminution s'explique par le fait qu'ils ont simplement repris telle quelle l'approche de Brill sans procéder à une adaptation pour le français. Pour ce qui est des mots inconnus, l'INALf a contourné le problème en utilisant un dictionnaire énorme (400 000 mots). De notre côté, comme il a déjà été démontré qu'un dictionnaire étendu, un jeu d'étiquettes restreint ou l'utilisation de gros corpus améliorent les résultats, nous n'avons pas testé ces variables. Notre objectif est d'améliorer le traitement des mots inconnus et d'adapter l'approche au français. En fait, nous voulons trouver une façon pour réduire au minimum l'écart entre les performances pour l'anglais et pour le français. Nous voulons aussi voir dans quelle mesure l'approche de Brill peut être améliorée ou non. Nous nous sommes penchée sur les modèles de règles lexicales afin de mieux les adapter au français et d'améliorer le traitement des mots inconnus. Nous avons donc apporté quelques modifications à l'approche afin d'en améliorer les résultats.

2.2 Version du logiciel utilisée

Le catégoriseur de Brill est disponible gratuitement pour des fins de recherche. Toutefois, nous n'avons pas utilisé directement le logiciel de Brill puisque certains modules ont été conçus pour fonctionner sous Unix. Notre directeur ayant développé un logiciel de catégorisation grammaticale automatique et d'apprentissage des règles contextuelles selon l'approche de Brill, nous avons utilisé ces deux modules. Nous avons utilisé un module programmé par Brill, soit le module d'apprentissage des règles lexicales et c'est uniquement ce module que nous avons modifié.

2.3 Corpus et échantillons

Pour entraîner le catégoriseur, il faut du texte français étiqueté grammaticalement qui ne comporte pas d'erreur (voir un extrait du corpus en annexe B) et du texte non étiqueté. Comme nous ne disposions pas de corpus étiquetés, nous avons, avec la permission de l'INALF, catégorisé un texte en utilisant leurs règles, leur dictionnaire et notre module de catégorisation. Nous avons ensuite corrigé minutieusement le résultat à la main. Notre corpus, a été construit à partir d'extraits d'ouvrages de la collection *Que sais-je ?* On y traite de sujets très variés appartenant à différents domaines. Cette variété de sujets permettra au catégoriseur d'étiqueter, par la suite, des textes portant sur autant de sujets. Notre corpus d'entraînement a été créé à partir d'une méthode d'échantillonnage par quotas. Nous avons identifié 15 domaines traités dans la collection (bioastronomie, communication, démographie, économie, ethnographie, ethnologie, géolinguistique, histoire, informatique, minéralogie, psychopathologie, pédagogie, psychopathologie de la violence, psychosomatique et sociologie), puis nous avons prélevé 10 pages par domaine. Les dix pages par domaine ont été prélevées de façon à ce que

toutes les parties d'un texte (introduction, développement, conclusion) se retrouvent dans le corpus puisque certains mots comme *introduction* et *conclusion* apparaissent normalement seulement en début et en fin de texte. Nous avons choisi de limiter ce corpus à 150 pages (44 573 mots) pour en accélérer la correction manuelle bien qu'un plus gros corpus aurait assuré des résultats légèrement supérieurs (voir Brill 1995b pour une discussion sur l'effet de la taille du corpus). De plus, il faut le plus de texte non étiqueté possible. Nous avons donc utilisé tout le reste des *Que sais-je ?* dont nous disposions en format électronique. Nous avons pris la partie inutilisée des 15 *Que sais-je ?* mentionnés ci-dessus et cinq autres *Que sais-je ?* entiers, portant sur la sociologie de la famille, les légendes urbaines, la religion, le temps et Venise.

Par ailleurs, pour évaluer les performances, il faut du texte français non étiqueté. Nous avons évalué les performances sur deux textes différents. Le premier texte est constitué d'une partie des 15 *Que sais-je ?* sélectionnés[14]. Cela nous a permis d'évaluer les performances du logiciel sur le type de texte pour lequel il a été entraîné. Nous avons sélectionné une page par domaine, pour un total de 15 pages. Le deuxième texte est de cinq pages et a été prélevé des cinq *Que sais-je ?* qui ne faisaient pas partie du corpus d'entraînement catégorisé. Ce texte permettra de constater les performances du catégoriseur sur un autre type de texte que celui sur lequel il a été entraîné et aussi de vérifier nos premières évaluations. Nous avons décidé d'évaluer les performances de l'approche sur 20 pages car l'évaluation devait être rapide puisque plusieurs évaluations devaient avoir lieu. Dans ces 20 pages, nous avons retrouvé toutes les parties du discours, et ce, dans des contextes très variés.

---

[14] Ces pages ne se retrouvent pas dans les corpus d'entraînement.

2.4 Critères d'analyse

Pour évaluer les performances du catégoriseur, nous avons compté le nombre d'erreurs contenues dans les textes d'évaluation catégorisés par le logiciel. Pour déterminer s'il y avait erreur ou non, nous avons choisi une définition pour chacune des parties du discours. Si l'étiquette grammaticale apposée à un mot ne correspondait pas à sa définition, c'est qu'il y avait erreur.

Puisque notre corpus d'entraînement, avant d'être corrigé à la main, a d'abord été catégorisé en utilisant les règles et le dictionnaire de l'INALF, nous avons utilisé les mêmes étiquettes et nous avons généralement suivi les mêmes critères de catégorisation qu'eux. Toutefois, les définitions données par l'INALF portent souvent sur des détails et non sur la catégorie même. Par exemple, ils ne donneront pas de définition générale à l'étiquette adjectif, mais vont expliquer comment ils ont traité tel ou tel cas particulier. Étant donné que leurs étiquettes correspondent en grande partie aux catégories reconnues en grammaire traditionnelle, nous avons eu recours à des ouvrages de référence en grammaire traditionnelle, soit *Le Bon Usage* (1969) et le *Petit Robert électronique* (1996). Nous avons privilégié celui des deux ouvrages de référence qui donnait la définition la plus exclusive possible. C'est-à-dire que nous n'avons pas retenu les définitions qui permettaient de classer les mots dans deux catégories à la fois. Tout comme l'INALF (voir Lecompte 1998) nous avons considéré les étiquettes en discours plutôt qu'en langue :

> Un de nos problèmes a été de définir à quel niveau nous allions coder le texte : Langue ? ou Discours ? Un codage en « langue » va offrir un éventail de possibilités, en quelque sorte poser les ambiguïtés, hors contexte. Un codage en « discours » va affecter une étiquette de PdeD [partie du discours] selon l'emploi en contexte dans le message écrit considéré. Après de multiples tentatives, nous avons finalement opté pour un codage « en discours », « en contexte ». Et ceci est dû en grande partie au fait que nous voulons jouer le jeu du « taggueur » de Brill, qui se place dans cette optique.

Tout ce qui se trouve entre deux séparateurs (' .,;:_+=-#|\!"/£$¢%¤?¬&¦*()[]«»<>''") est un mot et reçoit une étiquette sauf pour ce qui est de *aujourd'hui* qui est considéré comme un seul mot. Quelques mots complexes revenant souvent dans notre corpus d'entraînement ont été l'objet d'une attention spéciale lors de la correction. Les deux composantes des mots *peut-être, rendez-vous et faire-valoir*[15] ont reçu la même étiquette, soit respectivement adverbe, substantif et substantif.

Voici les définitions des catégories utilisées (voir la liste des étiquettes en annexe A)[16].

SBC :sg/pl (substantif) : Le substantif (singulier ou pluriel) est une « Unité du lexique […] qui peut se combiner avec divers morphèmes exprimant des modalités particulières (articles; pronoms démonstratifs, possessifs; marques du genre et du nombre, etc.) et qui correspond sémantiquement à une substance (être ou classe d'êtres, choses, notions). » (*Le petit Robert électronique* 1996)

---

[15] Il aurait été possible d'accorder un traitement spécial aux mots complexes comme l'ont fait l'INALF, mais cela ne faisait pas partie de nos objectifs.

[16] Pour faciliter la correction du corpus, nous n'avons retenu que trois étiquettes pour les participes, contrairement à l'INALF qui en retenait cinq.

SBP :sg/pl (substantif propre) : Le substantif propre (singulier ou pluriel) est celui « […] qui ne peut s'appliquer qu'à un seul être ou objet ou à une catégorie d'êtres ou d'objets pris en particulier; il individualise l'être, l'objet ou la catégorie qu'il désigne […]. » (*Le Bon Usage* 1969)

ADJ :sg/pl (adjectif) : L'adjectif (singulier ou pluriel) est un « Mot susceptible d'être adjoint directement (épithète) ou indirectement (attribut) au substantif avec lequel il s'accorde, pour exprimer une qualité (qualificatif) […]. » (*Le petit Robert électronique* 1996) Les mots considérés comme des adjectifs déterminatifs par la grammaire traditionnelle sont classés ailleurs ici. Certains sont classés comme des pronoms (ex. : interrogatif : qui) sur la base du fait qu'ils sont en lien avec le verbe et non avec le nom. D'autres sont classés comme des déterminants (ex. : possessif : ma, démonstratif : cette, indéfini : certain) sur la base du fait qu'ils peuvent être substitués à d'autres déterminants comme les articles (ex. : ***mon*** *livre est beau* = ***le*** *livre est beau*). Enfin, d'autres sont classés comme des relatifs (ex. : qui) sur la base du fait qu'ils sont en lien à la fois avec un nom et un verbe et non seulement avec un nom. Les adjectifs numéraux cardinaux sont classés comme des cardinaux, car ils ne sont pas substituables aux adjectifs, c'est-à-dire qu'ils ont un comportement syntaxique différent des autres adjectifs. Les adjectifs numéraux ordinaux sont classés comme des adjectifs puisqu'ils peuvent s'y substituer. Les adjectifs comparatifs, comme *meilleur*, sont considérés comme des adjectifs puisqu'ils ont à peu près le même comportement syntaxique que ceux-ci, c'est-à-dire qu'on peut substituer, dans bien des cas, un adjectif et un comparatif (ex. : *mon* ***meilleur*** *ami* = *mon* ***bon*** *ami*). Les superlatifs, comme *mieux*, sont classés comme des substantifs puisqu'ils sont normalement précédés d'un déterminant tout comme les substantifs (ex. : ***le mieux*** *c'est d'y aller*). Les adjectifs substantivés sont classés comme des substantifs puisqu'ils se comportent syntaxiquement comme des substantifs.

Les adjectifs adverbialisés sont classés comme des adverbes vu leur comportement syntaxique (ex. : *il crachait **noir***).

VCJ :sg/pl (verbe conjugué), **ACJ :sg/pl** (avoir conjugué), **ECJ :sg/pl** (être conjugué), **VNCFF** (verbe infinitif), **ANCFF** (avoir infinitif), **ENCFF** (être infinitif) : Le verbe conjugué (singulier ou pluriel) ou à l'infinitif est un « Mot qui exprime une action, un état, un devenir, et qui présente un système complexe de formes […]. » (*Le petit Robert électronique* 1996) Les verbes conjugués sont ceux qui ont une personne, un temps et un mode ; les autres sont à l'infinitif. Lecompte 1998 explique qu' « Une différenciation est faite, en ce qui concerne les verbes auxiliaires de temps et de voix, et les verbes pleins. Les verbes porteurs de modalités (« modaux ») ne sont pas distingués des autres et sont considérés comme verbes pleins. » Ainsi, les verbes avoir et être ont été distingués des autres verbes sur la base du fait qu'ils peuvent remplir des fonctions syntaxiques différentes à savoir être des auxiliaires dans les formes composées des verbes.

VNCNT (verbe au participe présent), **ANCNT** (avoir au participe présent), **ENCNT** (être au participe présent), **PAR :sg/pl** (verbe au participe passé), **APAR :sg/pl** (avoir au participe passé), **EPAR :sg/pl** (être au participe passé) : Le participe est une « Forme modale impersonnelle qui « participe » de l'adjectif (peut s'accorder en genre et en nombre) et du verbe (peut exprimer temps et voix et régir un complément). » (*Le petit Robert électronique* 1996) Les participes présents ont un sens actif et n'ont pas de sujet. Ils se terminent par *-ant* et sont invariables. Lecompte 1998 dit que « […] les adjectifs qualificatifs en *-ant* ne sont pas étiquetés ANT [participe présent] mais ADJ. » Les participes passés ont un sens passif et sont utilisés dans les formes composées des verbes et comme épithètes. De plus, ils ont un genre et un nombre et se terminent majoritairement par *-é*, *-i* et *-u*. Les participes

d'avoir et d'être sont distingués de ceux des autres verbes puisqu'ils servent à former les formes passives des verbes.

ADV (adverbe) : L'adverbe est un « Mot invariable ajoutant une détermination à un verbe, un adjectif, un adverbe, ou une phrase. » (*Le petit Robert électronique* 1996) Lecompte 1998 dit qu' « Il n'y a pas de codage particulier en ce qui concerne les adverbes prédéterminants. Exemple : *presque*/ADV tous les mois... A noter que "*tout*" / "*toute*" / "*toutes*" / "*tous*" en position de prédéterminants sont codés comme des déterminants et non comme des adverbes. [...] Les particules déictiques *ci* et *là* sont traitées comme des adverbes [...]. »

PREP (préposition) : La préposition est un « Mot grammatical, invariable, introduisant un complément (d'un substantif, d'un verbe, d'un adjectif, d'un adverbe) en marquant le rapport qui unit ce complément au mot complété. » (*Le petit Robert électronique* 1996) Lecompte 1998 dit que « Pour les prépositions plus complexes (= les Locutions), [...] l'un des éléments est une particule non indépendante : *afin*/PUL de/PREP - *quant*/PUL à/PREP [...] Les prépositions déictiques (*voici, voilà, revoici, revoilà*) ne reçoivent pas d'étiquette particulière. [...] "*comme*" n'est jamais une préposition, mais un subordonnant (introduisant une conjonctive normale ou tronquée), ou un adverbe exclamatif. »

DTN :sg/pl (déterminant), **DTC :sg/pl** (déterminant contracté) : Les déterminants contractés ou non (singulier ou pluriel) : « [...] sont les introducteurs de syntagmes nominaux [...]. » (Lecompte 1998) Les déterminants non contractés sont formés d'un seul mot alors que les déterminants contractés contiennent une préposition et un article (ex. : *du*, *des*, *au*, *aux*). Cette classe comprend des articles définis (ex: *le*, *la*, *les*), des articles indéfinis (ex: *un*, *une*, *des*, *certains*, *tout*), des articles partitifs (ex: *du*, *des*, *de*), des articles possessifs (ex: *son*, *sa*, *ses*, *leurs*), des articles

démonstratifs (ex: *ces*, *cette*, *cet*, *ce*) et des articles quantifieurs (ex: *plusieurs*, *quelques*, *aucune*). *De* et *des* ne sont jamais classés comme des articles partitifs, ils sont toujours respectivement préposition et déterminant contracté. Le *l'* de *si* ***l'****on mangeait* est classé comme une particule attachée puisqu'elle est indissociable de *on*. Pour bien différencier les déterminants indéfinis des adjectifs, nous avons suivi les règles de Lecompte 1998 :

- Placés entre un déterminant et un nom : adjectifs (ex. : un *certain*/ADJ:sg monsieur X).
- Postposés au substantif : adjectifs (ex. : il est d'un âge *certain*/ADJ:sg).
- Antéposés au substantif, et en l'absence de tout autre déterminant : déterminants (ex. : *certains*/DTN:pl jours, il fait gris).
- Antéposés à un déterminant non contracté, "*tout*" et ses variantes deviennent déterminants (ex. : *presque*/ADV *tous*/DTN:pl *leurs*/DTN:pl gamins en veulent).

PRV :sg/pl/++ (pronom verbal), **PRO :sg/pl/++** (pronom) : Le pronom supporté ou non par le verbe (singulier, pluriel ou nombre indéterminé) est un « Mot grammatical qui sert à représenter un nom de sens précis déjà employé à un autre endroit du contexte ou qui joue le rôle d'un nom absent, généralement avec une nuance d'indétermination. » (*Le petit Robert électronique* 1996) Les pronoms verbaux sont les pronoms personnels sujets (ex. : *je*, *tu*, *il*, *elle*, *nous*, *vous*, *ils*, *elles*, *ce*, *on*), compléments directs (ex. : *me*, *te*, *se*, *leur*, *les*, *la*, *le*), réfléchis (ex. : *se*, *me*, *le*). Les pronoms non verbaux sont tous les autres pronoms, non liés au verbe, comme les pronoms personnels (ex. : *moi*, *toi*, *soi*, *lui*, *vous*, *nous)*, les pronoms démonstratifs (ex. : *ceci*, *cela*, *ceux*, *ce*), les pronoms possessifs[17] (ex. : *sien*, *sienne*), les

[17] Ils sont à différencier des adjectifs possessifs. Ils sont pronoms seulement s'ils sont précédés d'un article défini.

pronoms indéfinis (ex. : *plusieurs*, *maint*, *chacun*) et les pronoms interrogatifs (ex. : *qui*, *que*). Le genre indéterminé (++) est donné à *en* et *y*, et à *s'* et *se* devant un verbe non conjugué. Les relatifs sont exclus de cette catégorie puisqu'ils entrent en relation à la fois avec un nom et un verbe.

COO (coordonnant) : Le coordonnant ou conjonction de coordination est une « Partie du discours qui sert à joindre deux mots ou groupes de mots [...] qui, entre des mots ou des propositions de même fonction, marquent l'union (ex. : et), l'opposition (ex. : mais, [...]), l'alternative ou la négation (ex. : ni, ou) [...]. » (*Le petit Robert électronique* 1996) Lecompte 1998 note que « *donc* est considéré comme un adverbe. *Sinon* est préposition ou adverbe [...] chaque élément de coordonnant discontinu (*soit... soit*) reçoit un code de coordonnant normal. »

REL (relatif) : Les relatifs servent à « [...] rapporter, à joindre à un nom ou à un pronom qu'ils représentent [antécédent] une proposition subordonnée dite relative, qui explique ou détermine ce nom ou ce pronom. » (*Le Bon Usage* 1969) *Qui, que, quoi, dont, où, lequel, duquel* et *auquel* sont des relatifs. Lecompte 1998 note qu'ils « [...] se distinguent des Subordonnants en ce qu'ils ont toujours une fonction dans la Subordonnée qu'ils introduisent. »

SUB (subordonnant) : Le subordonnant ou conjonction de subordination sert à « [...] relier une proposition subordonnée à la proposition dont elle dépend. » (*Le Bon Usage* 1969) Lecompte 1998 ajoute que « La subordination est une relation asymétrique de dépendance entre une proposition dite "subordonnée" et une proposition dite "principale". Les subordonnées sont le plus souvent introduites par des marqueurs de subordination. Parmi ceux-ci, les conjonctions de subordination permettent d'introduire des subordonnées complétives (c.o.d. ou c.o.i.) ou circonstancielles, complètes ou incomplètes. »

CAR (cardinaux) : Les cardinaux sont des nombres écrits en chiffres ou en lettres. Ils « indiquent le nombre précis des êtres ou des objets désignés par le nom : Deux livres, Trois hommes. » (*Le Bon Usage* 1969) contrairement aux ordinaux qui « indiquent l'ordre, le rang des êtres ou des choses […]. » Lecompte 1998 explique qu' « Il n'y a pas de subdivision des cardinaux selon leur fonction. Qu'ils soient Déterminants, Pronoms, ou Noms, ils sont tous étiquetés CAR. […] La distinction entre déterminant indéfini et cardinal, à propos de "*un*" et "*une*", n'est pas posée, car insoluble au niveau où nous travaillons, avec l'outil que nous utilisons. Ils sont systématiquement considérés comme déterminants ou pronoms, jamais comme cardinal. »

ABR (abréviation) : L'abréviation est un mot auquel on a retranché des lettres (ex. : av/ABR J/ABR C/ABR). Les lettres d'un sigle contenant des points sont chacune considérées comme une abréviation.

FGW (mot étranger) : Les mots étrangers sont des mots qui n'appartiennent pas à la langue française. Les emprunts qui font l'objet d'une entrée dans *Le Petit Robert électronique* 1996 ont été considérés comme des mots appartenant à la langue française (à moins qu'ils n'apparaissent dans une phrase écrite dans une autre langue que le français) tandis que les autres ont été classés comme mots étrangers. Les mots constituants certaines locutions latines (ex. : *a priori*) ont été classés comme des mots étrangers puisqu'ils ont un comportement syntaxique anormal, c'est-à-dire qu'ils ne peuvent apparaître seul.

INJ (interjection) : L'interjection est un « Mot invariable pouvant être employé isolément pour traduire une attitude affective du sujet parlant. » (*Le petit Robert électronique* 1996) Les interjections n'entretiennent pas de lien syntaxique avec les autres éléments du discours.

PFX (préfixe) : Les préfixes sont des éléments placés entre deux séparateurs[18] (ex. : franco/PFX -/- danois/ADJ:sg), mais ne pouvant apparaître seul, n'étant pas des morphèmes libres.

SYM (symbole) : Les symboles sont des caractères simples ne faisant pas partie des signes de ponctuation ou de l'alphabet du français (ex. : &, =, %, etc.).

PUL (particule) : La particule est un « Petit mot invariable, élément de composition ou de liaison » (*Le petit Robert électronique* 1996). Lecompte 1998 dit que « Cette étiquette concerne des éléments qui n'ont pas d'existence autonome […] ». Par exemple, le *s'* dans *s'agit* n'a pas de sens propre et est indissociable du verbe. Il ne s'agit pas d'un pronom réflexif. Il en est de même pour le *t* dans *a-t-il*. C'est un élément non autonome, indissociable du verbe. La particule se distingue du préfixe en ce qu'elle apparaît toujours lié au même mot.

2.5 Évaluation

Pour être en mesure d'évaluer l'impact des modifications effectuées, un entraînement contrôle a été effectué. Ce qui veut dire que nous avons pris le logiciel tel quel, sans modification, que nous l'avons entraîné à partir de notre corpus d'entraînement pour qu'il produise un dictionnaire, des règles contextuelles et des règles lexicales. Puis, nous avons fait catégoriser nos corpus d'évaluation par le logiciel en utilisant ces règles et ce dictionnaire. Les résultats de cet entraînement ont servi de base de comparaison. De plus, à partir de l'évaluation contrôle du texte de 15 pages, une analyse et une typologie des erreurs a été effectué pour déterminer quelles sont les modifications devant être apportées à l'approche.

---

[18] (' .,;:_+=-# | \!"/£$¢%¤?¬&¦*()[]«»<>'")

Certaines de ces modifications ont été implémentées dans le logiciel. Chaque modification a été apportée séparément en partant du module initial. Après chaque modification, le logiciel a été entraîné à nouveau et ses performances évaluées. L'évaluation a toujours été faite sur le même texte. Ensuite nous avons compté le nombre d'erreurs produites. Nous avons considéré qu'il y avait une erreur si la catégorie apposée à un mot n'était pas conforme à la définition donnée pour cette catégorie. Puis nous avons combiné les modifications s'étant avérées efficaces et nous avons réentraîné et réévalué le logiciel sur deux textes cette fois.

Certaines modifications ne pouvaient pas être vérifiées par implémentation parce que l'architecture du logiciel utilisé ne le permet pas, ce qui veut dire que le module aurait dû être reprogrammé en entier. Nous avons donc procédé par simulation. Pour ce faire, nous avons établi la liste des mots pouvant être affectés par une modification donnée dans nos deux textes d'évaluation. Puis, nous avons appliqué la modification manuellement à chaque mot. Nous avons ensuite comptabilisé le nombre d'erreurs éliminées et le nombre d'erreurs ajoutées par cette modification pour déterminer si la modification était efficace.

Chapitre 3 Expérimentation et analyse des résultats

3.1 Typologie des erreurs

Comme nous l'avons déjà mentionné, l'approche de Brill permet d'atteindre environ 97 % de bons résultats pour l'anglais (Brill 1995b), et environ 94 % pour le français. Dans le cadre de nos recherches nous avons choisi d'utiliser un corpus d'entraînement de 150 pages (44 573 mots), pour en faciliter la manipulation, ce qui permet d'obtenir une performance de 92,8 % pour le français. Brill 1995b rapporte une efficacité de 97,2 % pour un corpus de 600 000 mots contre une efficacité de 96,7 % pour un corpus de 64 000 mots. D'autres facteurs influencent la performance du logiciel : le nombre d'étiquettes choisies et leur type, l'utilisation ou non de dictionnaires supplémentaires et le degré de spécialisation du texte à catégoriser. Les modifications que nous proposons aux sections 3.2 et 3.3 pourront être reprises avec un plus gros corpus de référence.

Le texte ayant servi à établir la typologie des erreurs était de 15 pages et contenait 8602 mots dont 7044 (81,9 %) étaient connus et 1558 (18,1 %) inconnus. Des mots connus, 4966/7044 (70,5 %) étaient non ambigus. Le corpus contenait aussi 1258 signes de ponctuation que nous n'avons pas considérés dans notre évaluation. Le logiciel a produit 617 erreurs (/8602) en catégorisant ce texte pour un taux d'efficacité de 92,8 % (voir le Tableau 1), comme cela était prévu compte tenu de la taille du corpus d'entraînement.

Tableau 1 – **Nombre d'erreurs produites par le logiciel**

	Erreurs		Erreurs/nombre de mots		Erreur/nombre de mots ambigus	
Mots connus	153 /617	24,8 %	153 /7044	2,2 %	153 /2078	7,4 %
Mots inconnus	464 /617	75,2 %	464 /1558	29,8 %	464 /1558	29,8 %
Total	617 /617	100 %	617 /8602	7,2 %	617 /3636	17 %

La performance pour les mots inconnus est de 70,2 % (1094 mots bien catégorisés / 1558 mots inconnus), tandis que la performance pour les mots connus est de 97,8 % (6891 mots bien catégorisés / 7044 mots connus) ou de 92,6 % (1925 mots bien catégorisés / 2078 mots connus ambigus) si on ne considère que les mots connus ambigus. Les résultats obtenus pour les mots inconnus (70,2 %) sont similaires à ceux d'abord obtenus par Brill (1993).

3.1.1 Erreurs portant sur les mots connus

Les erreurs portant sur des mots connus peuvent être attribuées à diverses causes (voir le Tableau 2).

Tableau 2 - **Typologie des erreurs portant sur les mots connus**

Ambiguïté trop rare dans le corpus	58/153	37,9 %
Catégorie absente du dictionnaire	54/153	35,3 %
Mot adjacent mal catégorisé	14/153	9,2 %
Surpuissance de certaines règles	8/153	5,2 %
Mauvaise catégorie dans le dictionnaire	6/153	3,9 %
Élément déterminant trop loin	5/153	3,2 %
Suite de mots ambigus	4/153	2,6 %
Mauvais découpage en mot	2/153	1,3 %
Faute de frappe ou de numérisation	2/153	1,3 %

Certains mots (37,9 %) sont mal catégorisés parce que l'ambiguïté à résoudre était trop rare dans le corpus d'entraînement. Ce type d'erreurs touche des mots présents au dictionnaire avec la catégorie appropriée. Le logiciel n'a pas créé de règle permettant d'assigner la bonne catégorie à ces mots bien qu'il existe un modèle de règle adéquat. En effet, le logiciel ne retient que les règles qui permettent de corriger plus de trois erreurs. Le contexte permettrait une bonne catégorisation, mais l'ambiguïté ne se présentait que trop rarement dans le corpus d'entraînement. On en voit un exemple en (23).

23. Exemple d'erreur due au fait que l'ambiguïté à résoudre était trop rare dans le corpus d'entraînement

L'/DTN:sg autre/ADJ:sg constituant/VNCNT > SBC :sg organique/ADJ:sg des/DTC:pl nucléotides/ADJ:pl ,/, un/DTN:sg sucre/SBC:sg à/PREP cinq/CAR atomes/SBC:pl de/PREP carbone/ADJ:sg ,/, fut/ECJ:sg découvert/ADJ:sg en/PREP 1911/CAR ./.

Si on regarde de plus près l'exemple en (23) on constate que *constituant* qui est catégorisé participe présent plutôt que substantif est précédé d'un déterminant et d'un adjectif et qu'il est suivi d'un adjectif. Dans ce contexte, il est impossible de retrouver un participe présent. Le contexte aurait donc permis un étiquetage adéquat. Le logiciel aurait pu créer par exemple une règle indiquant qu'un mot ambigu participe présent/substantif précédé d'un déterminant et d'un adjectif est un substantif. Toutefois, il ne l'a pas fait puisque, dans le corpus d'entraînement, il n'y avait qu'un seul mot (*constituant*) qui avait l'ambiguïté participe présent/substantif et ce mot

n'apparaissait que trois fois[19]. Donc, la faible fréquence de cette ambiguïté a fait en sorte qu'aucune règle ne soit créée.

Certaines erreurs sont dues au fait que la bonne catégorie pour un mot n'était pas présente dans le dictionnaire. Ce qui veut dire que le mot apparaissait dans le corpus d'entraînement, c'est donc un mot connu, mais il n'apparaissait pas avec la catégorie voulue. Ce type d'erreurs représente 35,3 % des erreurs et pourrait être éliminé en augmentant la taille du corpus d'entraînement ou en enrichissant le dictionnaire. On voit en (24) des exemples de ce type d'erreurs. Le mot et la catégorie qui lui a été assignée sont encadrés. La bonne catégorie suit et est soulignée.

24. Exemples d'erreurs dues au fait que la bonne catégorie était absente du dictionnaire[20]

La/DTN:sg découverte/SBC:sg de/PREP la/DTN:sg structure/SBC:sg de/PREP l'/DTN:sg ADN/SBC:sg marque/SBC:sg >VCJ :sg le/DTN:sg développement/SBC:sg de/PREP ce/PRO:sg que/SUB l'/PUL on/PRV:sg nomme/VCJ:sg à/PREP présent/SBC:sg la/DTN:sg biologie/SBC:sg moléculaire/ADJ:sg ./.

[…] car/COO tous/DTN:pl les/DTN:pl systèmes/SBC:pl vivants/ADJ:pl ,/, de/PREP l'/DTN:sg être/ENCFF >SBC :sg monocellulaire/ADJ:sg le/DTN:sg plus/ADV simple/ADJ:sg ,/, à/PREP l'/DTN:sg être/ENCFF >SBC :sg le/DTN:sg plus/ADV "/" évolué/PAR:sg "/" […].

[19] L'ambiguïté substantif sg./ participe présent est traitée séparément.
[20] Il y aura, des les exemples présentés, d'autres erreurs que celles qui sont encadrées. Dans chacun des cas, nous avons vérifié que ces erreurs n'en causaient pas d'autres et surtout qu'elles n'étaient pas la cause de l'erreur encadrée.

[…] (/(ou/COO mesure/SBC:sg d'/PREP incertitude/SBC:sg de/PREP l'/DTN:sg arrangement/SBC:sg des/DTC:pl éléments/SBC:pl du/DTC:sg système/SBC:sg physique/ADJ:sg ,/, degré/PAR:sg qui/REL va/VCJ:sg croissant/ADJ:sg >VNCNT jusqu'/PREP au/DTC:sg désordre/SBC:sg et/COO signifie/VCJ:sg par/PREP là/ADV la/DTN:sg dissolution/SBC:sg du/DTC:sg système/SBC:sg)/) […].

La première erreur présentée en (24) porte sur le mot *marque*. Celui-ci est catégorisé substantif bien qu'il s'agisse d'un verbe conjugué. Le logiciel n'a pas pu lui donner la catégorie verbe conjugué puisque celle-ci n'était pas inscrite comme une possibilité pour ce mot; elle n'apparaissait pas dans le corpus d'entraînement.

Certains mots (9,2 %) sont mal catégorisés à cause qu'un des mots adjacents était mal catégorisé. C'est-à-dire que le mot est présent au dictionnaire avec comme catégorie possible la bonne catégorie. L'application (ou la non-application) d'une règle contextuelle aurait permis d'apposer la bonne étiquette si le mot adjacent avait été bien catégorisé. La majorité de ces erreurs porte sur des mots outils. Ce type d'erreurs diminuera au fur et à mesure que le catégoriseur sera amélioré. On en voit des exemples en (25) où le mot mal catégorisé est encadré et le mot adjacent ciblé comme la cause est surligné.

25. Exemples d'erreurs dues au fait qu'un mot adjacent est mal catégorisé

Le/DTN:sg biochimiste/ADJ:sg allemand/SBC:sg > ADJ :sg Albrecht/SBP:sg Kossel/SBC:sg ,/, se/PRV:sg lança/SBC:sg alors/ADV dans/PREP l'/DTN:sg étude/SBC:sg systématique/ADJ:sg de/PREP la/DTN:sg structure/SBC:sg des/DTC:pl acides/ADJ:pl nucléiques/ADJ:pl ./.

Le/DTN:sg Parlement/SBC:sg ,/, à/PREP l'/PRV:sg >DTN :sg inverse/VCJ:sg ,/, considère/SBC:sg que/SUB la/DTN:sg décision/SBC:sg budgétaire/ADJ:sg qu'/SUB il/PRV:sg a/ACJ:sg arrêtée/PAR:sg quant/PUL à/PREP l'/DTN:sg inscription/SBC:sg […].

[…] tarifer/VNCFF le/DTN:sg voyage/SBC:sg selon/PREP la/DTN:sg réduction/SBC:sg dont/REL bénéficie/VCJ:sg le/PRV:sg >DTN :sg passager/VNCFF ,/, le/DTN:sg parcours/SBC:sg souhaité/PAR:sg et/COO l'/DTN:sg heure/SBC:sg de/PREP la/DTN:sg journée/PAR:sg […].

Dans le premier exemple en (25) on voit que *allemand* est catégorisé substantif plutôt qu'adjectif. Cette erreur est due au fait que le mot précédent, *biochimiste*, est inconnu et qu'il a été mal catégorisé, c'est un substantif et non un adjectif. S'il avait été bien catégorisé, la règle voulant qu'un mot présentement catégorisé substantif et précédé d'un substantif devienne un adjectif aurait pu s'appliquer et *allemand* aurait été bien catégorisé.

D'autres erreurs viennent du fait que certaines règles sont trop puissantes (5,2 %). Ce qui veut dire qu'elles s'appliquent là où elles ne le devraient pas. On en voit un exemple en (26).

26. Exemple d'erreur due à la trop grande puissance de certaines règles

On/PRV:sg ignore/SBC:sg le/PRV:sg > DTN :sg mécanisme/SBC:sg biologique/ADJ:sg qui/REL conduit/VCJ:sg à/PREP cette/DTN:sg remarquable/ADJ:sg constance/SBC:sg ./.

Dans l'exemple présenté en (26), on voit le mot *le* catégorisé pronom verbal plutôt que déterminant. Pourtant, il est suivi d'un substantif, ce qui aurait

dû permettre d'éviter cette erreur. Mais il existe une règle, selon laquelle un mot catégorisé déterminant, précédé par *on* de un ou deux mots devient un pronom verbal. Cette règle s'applique ici puisque *on* le précède de deux mots. La règle est beaucoup trop puissante bien qu'elle puisse donner souvent de bons résultats.

Puis des erreurs viennent du fait qu'il y avait dans le dictionnaire une mauvaise catégorie pour un mot. Il subsistait donc dans le corpus d'entraînement quelques erreurs de catégorisation. Il est pratiquement inévitable que des erreurs échappent au correcteur humain. Le pourcentage d'erreurs dues à la présence d'une mauvaise catégorie dans le dictionnaire représente 3,9 %. On en voit un exemple en (27).

27. Exemple d'erreur due au fait qu'il y avait une mauvaise catégorie dans le dictionnaire

Enfin/ADJ:sg >ADV ,/, le/DTN:sg modèle/SBC:sg de/PREP la/DTN:sg double/ADJ:sg hélice/SBC:sg de/PREP l'/DTN:sg acide/SBC:sg désoxyribonucléique/ADJ:sg fut/ECJ:sg établi/PAR:sg par/PREP le/DTN:sg physicien/SBC:sg anglais/ADJ:sg Francis/SBP:sg Harry/SBP:sg Compton/SBP:sg Crick/SBC:sg et/COO son/DTN:sg collaborateur/SBC:sg américain/ADJ:sg biochimiste/ADJ:sg James/SBP:sg Dewey/SBP:sg Watson/SBP:sg dans/PREP les/DTN:pl années/SBC:pl 1950/CAR ./.

Dans l'exemple présenté en (27), le mot *Enfin* est catégorisé adjectif singulier plutôt qu'adverbe puisque la seule catégorie qui était disponible pour ce mot dans le dictionnaire était adjectif. La catégorie adjectif n'aurait pas dû se retrouver dans le dictionnaire, c'est adverbe qui aurait dû s'y retrouver. Il y avait donc une erreur de catégorisation dans le corpus d'entraînement.

Ensuite, certaines erreurs sont dues au fait que le contexte dont tient compte le logiciel pour créer ses règles est trop restreint (trois mots de chaque côté). On voit un exemple de ce type d'erreurs en (28).

28. Exemples d'erreurs dues au contexte trop restreint dont tient compte le logiciel pour créer ses règles

> En/PREP effet/SBC:sg ,/, dans/PREP la/DTN:sg société/SBC:sg internationale/ADJ:sg ,/, on/PRV:sg ne/ADV peut/VCJ:sg pas/ADV forcer/VNCFF les/DTN:pl États/SBC:pl à/PREP s'/PRV:++ engager/VNCFF dans/PREP un/DTN:sg processus/SBC:sg juridique/ADJ:sg dont/REL ils/PRV:pl ne/ADV veulent/VCJ:pl pas/ADV ,/, soit/ECJ:sg > COO parce/PUL qu'/SUB il/PRV:sg est/ECJ:sg inacceptable/ADJ:sg pour/PREP eux/PRO:pl dans/PREP l'/DTN:sg absolu/SBC:sg ,/, soit/ECJ:sg > COO parce/PUL qu'/SUB il/PRV:sg intervient/VCJ:sg trop/ADV tôt/ADV ./.

Dans notre corpus d'évaluation, ces erreurs portent toutes sur le mot *soit* utilisé comme coordonnant. Le meilleur indice pour déterminer que *soit* est un coordonnant est la présence d'un autre *soit*. Mais la distance entre deux *soit* est rarement de moins de trois mots.

Certaines erreurs (2,6 %) viennent du fait que plusieurs mots ambigus nom/adjectif se suivent. On en voit un exemple en (29).

29. Exemples d'erreurs dues au fait que plusieurs mots ambigus nom/adjectif se suivent

Ritter/SBP:sg est/ECJ:sg en/PREP revanche/SBC:sg attentif/ADJ:sg à/PREP l'/DTN:sg impact/SBC:sg des/DTC:pl nouvelles/SBC:pl > ADJ :pl techniques/ADJ:pl > SBC :pl et/COO de/PREP la/DTN:sg révolution/SBC:sg industrielle/ADJ:sg :/: il/PRV:sg est/ECJ:sg un/PRO:sg des/DTC:pl premiers/ADJ:pl à/PREP signaler/VNCFF le/DTN:sg rapetissement/SBC:sg du/DTC:sg monde/SBC:sg que/SUB provoque/VCJ:sg la/DTN:sg navigation/SBC:sg à/PREP vapeur/SBC:sg ,/, et/COO à/PREP s'/PRV:++ interroger/VNCFF sur/PREP sa/DTN:sg signification/SBC:sg ./.

Cela est particulièrement problématique en français puisque certains adjectifs sont prénominaux, d'autres sont postnominaux (ex. : un gros ballon rouge) et d'autres encore peuvent se placer ou bien avant le nom ou bien après le nom. Pour le moment, le logiciel n'arrive pas à créer une règle pouvant traiter correctement tous les mots ambigus nom/adjectif puisque la position de l'adjectif par rapport au nom varie.

Puis il y a deux erreurs qui sont dues à des fautes de frappe ou de numérisation[21], ce qui représente 1,3 % des erreurs. On voit ces erreurs en (30).

[21] Les *Que Sais-Je ?* utilisés ont été numérisés.

30. Erreurs dues à une faute de frappe ou de numérisation

Tous/DTN:pl les/DTN:pl gouvernements/SBC:pl doivent/VCJ:pl avoir/ANCFF recours/SBC:sg aux/DTC:pl prêteurs/SBC:pl privés/PAR:pl et/COO aux/DTC:pl maisons/SBC:pl marchandes/ADJ:pl internationales/ADJ:pl pour/PREP assurer/VNCFF les/PRV:pl >le/DTN :sg transfert/SBC:sg d'/PREP argent/SBC:sg

l/COO >1/CAR L'/DTN:sg acquisition/SBC:sg des/DTC:pl métaux/ADJ:pl précieux/ADJ:sg est/ECJ:sg le/DTN:sg but/SBC:sg principal/SBC:sg de/PREP l'/DTN:sg activité/SBC:sg économique/ADJ:sg

Dans le premier exemple présenté en (30) le mot *les* aurait dû être orthographié *le* (ou *transfert* orthographié *transferts*). Il n'a donc pas pu être bien catégorisé. Dans le deuxième exemple, le caractère numérique *1* a été reconnu lors de la numérisation des textes comme un *L*. Il n'a donc pas pu être catégorisé comme un cardinal.

Deux erreurs (1,3 %) sont dues à un mauvais découpage en mots. Certains traits d'union devraient être considérés comme faisant partie d'un mot tandis que d'autres devraient être considérés comme des séparateurs de mots[22]. Notre logiciel les considère tous comme des séparateurs de mots. On voit ces erreurs en (31).

[22] Il est possible de dresser une liste des mots ne devant pas être découpés. C'est d'Ailleurs ce qu'a fait l'INALF.

31. Erreurs dues à un mauvais découpage en mot

[…] autant/ADV que/SUB faire/VNCFF se/PRV:sg peut/VCJ:sg ,/, la/DTN:sg quantité/SBC:sg de/PREP bien/ADV -/- être/ADV >bien-être/SBC :sg (/(c'/PRV:sg est/ECJ:sg le/DTN:sg mouvement/SBC:sg de/PREP libération/SBC:sg animale/ADJ:sg cher/ADJ:sg aux/DTC:pl Anglo/PFX -/- Saxons/SBC:pl)/) ./.

3.1.2 Erreurs portant sur les mots inconnus

Les erreurs portant sur les mots inconnus représentent 75,2 % des erreurs. Les erreurs de ce type sont donc très nombreuses et les mots inconnus sont de nature très diverse. Il n'est pas possible ici de dire que X % des erreurs viennent de telle ou telle cause puisqu'un même mot inconnu peut avoir été mal catégorisé pour plusieurs raisons. L'erreur peut porter à la fois sur les préfixes, les infixes, les suffixes et le contexte. Nous allons donc simplement faire la liste des causes d'erreurs.

Premièrement, seulement 3 % de ces erreurs portent sur des mots outils. On en voit un exemple en (32).

32. Exemple d'erreur portant sur un mot outil inconnu

[…] :/: par/PREP exemple/SBC:sg ,/, l'/DTN:sg ouragan/SBC:sg n'/ADV est/ECJ:sg pas/ADV habité/SBC:sg par/PREP un/DTN:sg esprit/SBC:sg ,/, mais/COO il/PRV:sg est/ECJ:sg vivant/SBC:sg en/PREP soi/SBC:sg > PRO :sg ./.

Dans cet exemple, on voit que le mot *soi* est catégorisé substantif plutôt que pronom. Dans ce cas-ci, la règle qui est intervenue est : un mot qui porte comme étiquette de défaut substantif auquel on peut ajouter le suffixe *-s* (*soi-s*) est un substantif singulier. Évidemment cette règle ne peut donner de bons résultats sur des mots outils.

Dans le Tableau 3, on voit le rappel et la précision pour chaque type de mot outil. La précision correspond au nombre d'erreurs pour une étiquette assignée par le logiciel. Le rappel correspond au nombre d'erreurs pour une étiquette qui aurait dû être assignée par le logiciel. 0 % représente donc la meilleure performance.

Tableau 3 **- Précision et rappel des mots outils inconnus**

Catégorie	Nombre	Précision erreur/total donné		Rappel erreur/total voulu	
Déterminant	singulier				
	pluriel	0/1	0 %	0/1	0 %
Déterminant contracté	singulier	0/3	0 %	0/3	0 %
	pluriel				
Pronom verbal	singulier				
	pluriel				
Pronom	singulier			2/2	100 %
	pluriel			1/1	100 %
Préposition		0/1	0 %	3/4	75 %
Coordonnant		1/1	100 %		
Subordonnant				1/1	100 %
Relatif				1/1	100 %
Particule attachée		0/1	0 %	0/1	0 %
Préfixe		2/6	33,3 %	10/14	71,4 %
Abréviation				7/7	100 %
Total		3/13	23,1 %	25/35	71,4 %

La performance du logiciel sur ce type de mots inconnus n'est pas très bonne. En effet, la précision est de 23,1 % et le rappel de 71,4 %. Ce n'est toutefois pas un grave problème dans le cadre de cette approche, puisque les mots outils font partie de listes finies. Avec un corpus d'entraînement plus gros, ou l'ajout d'un dictionnaire externe, ils seraient tous connus. De plus, il serait difficile d'arriver à catégoriser les mots outils sur la base de leur

morphologie, c'est-à-dire à l'aide de règles lexicales, puisqu'ils ne comportent souvent qu'une morphologie grammaticale.

Dans le cas des mots pleins, la situation est plus compliquée. Les règles utilisent leur morphologie et les mots adjacents pour les catégoriser. Toutefois cela ne fonctionne pas très bien (voir le Tableau 4). La précision est de seulement 29,8 % et le rappel de 28,8 % (la meilleure performance étant représentée par 0 %).

Tableau 4 - **Précision et rappel des mots pleins inconnus**

Catégorie	Nombre	Précision erreur/total donné		Rappel erreur/total voulu	
Nom propre	singulier	18/97	18,6 %	30/109	27,5 %
	pluriel				
Nom commun	singulier	205/521	39,3 %	71/387	18,3 %
	pluriel	65/228	28,5 %	65/228	28,5 %
Adjectif	singulier	42/164	25,6 %	67/189	35,4 %
	pluriel	59/141	41,8 %	38/120	31,7 %
Verbe conjugué	singulier	8/43	18,6 %	63/98	64,3 %
	pluriel	9/61	14,8 %	3/55	5,5 %
Participe passé	singulier	20/75	26,7 %	30/85	35,3 %
	pluriel	18/50	36 %	19/51	37,3 %
Verbe infinitif		8/79	10,1 %	7/78	9 %
Participe présent		7/18	38,9 %	1/12	8,3 %
Adverbe		2/31	6,5 %	24/53	45,3 %
Cardinaux		0/37	0 %	5/42	11,9 %
Mot étranger				16/16	100 %
Total		461/1545	29,8 %	439/1523	28,8 %

Les résultats sont tels car la plupart des morphèmes sont ambigus, c'est-à-dire qu'ils ne se retrouvent pas exclusivement associés à une classe de mots, et le contexte dont tient compte le logiciel est très réduit (un mot avant et un mot après).

Certains des mots qu'il a à traiter n'ont pas de morphologie particulière. C'est le cas des noms propres et des sigles. Les noms propres et les sigles sont normalement caractérisés par le fait qu'ils commencent par une majuscule et, sur la base de cette caractéristique, le logiciel leur donne la catégorie nom propre dans un premier temps. Toutefois puisque ce ne sont pas tous les mots commençant par une majuscule qui sont des noms propres, il est prévu que leur catégorie peut changer. Il arrive qu'un nom propre contienne une suite de caractères pouvant correspondre à un affixe et le logiciel modifiera sa catégorie sur la base de la présence de cette suite de lettres, souvent à tort. On voit un exemple d'erreur de ce type en (33).

33. Exemple d'erreur portant sur des mots sans morphologie particulière

Le/DTN:sg biochimiste/ADJ:sg allemand/SBC:sg
Albrecht/SBP:sg Kossel/SBC:sg > SBP :sg ,/, se/PRV:sg
lança/SBC:sg alors/ADV dans/PREP l'/DTN:sg étude/SBC:sg
systématique/ADJ:sg de/PREP la/DTN:sg structure/SBC:sg
des/DTC:pl acides/ADJ:pl nucléiques/ADJ:pl ./.

Dans ce cas-ci, la règle suivante est intervenue : un mot contenant la lettre -*e*- est un substantif singulier. Cette règle, qui est appliquée en tout premier, semble au premier coup d'œil trop puissante. Toutefois, il y a plusieurs règles qui viendront changer les catégories de mots à sa suite. Ce qui se produit avec les noms propres, c'est qu'il n'y a aucun indice morphologique qui permettra de leur redonner la catégorie nom propre.

Pour ce qui est des mots étrangers, bien qu'ils comportent une morphologie particulière dans la plupart des cas, il n'y a pas assez d'exemples dans le corpus d'entraînement pour créer des règles sur la base de leur morphologie. Seulement trois règles ont été créées pour les mots étrangers. Une disant qu'un mot qui se termine par -*y* est un mot étranger, l'autre disant qu'un

mot auquel on peut ajouter le préfixe *perm-* est un mot étranger, et la dernière disant qu'un mot auquel on peut ajouter le suffixe *-fert* est un mot étranger. On voit des exemples d'erreurs portant sur des mots étrangers inconnus en (34).

34. Exemples d'erreurs portant sur des mots étrangers inconnus

> Le/DTN:sg naturisme/SBC:sg d'/PREP Andrew/SBP:sg Lang/SBP:sg ,/, développé/PAR:sg dans/PREP deux/CAR ouvrages/SBC:pl ,/, The/SBC:sg > FGW making/SBC:sg > FGW of/FGW Religion/SBC:sg > FGW ,/, 1898/CAR ,/, et/COO Magic/SBP:sg > FGW and/FGW Religion/SBC:sg > FGW ,/, 1901/CAR ,/, est/ECJ:sg plus/ADV nuancé/SBC:sg et/COO plus/ADV proche/SBC:sg du/DTC:sg réel/ADJ:sg ./.

Dans cette phrase, *The* a été catégorisé substantif puisqu'il contient un *-e-*, *making* substantif puisqu'il contient un *-i-*, *Religion* substantif puisqu'il se termine par *-ion* et *Magic* nom propre puisqu'il commence par *M-*. On voit aussi que *of* et *and*, qui sont des mots connus, sont étiquetés mots étrangers. La seule présence d'un mot étranger connu à côté d'un mot inconnu devrait suffire à catégoriser celui-ci comme un mot étranger. Mais le logiciel ne considère qu'un mot avant et un mot après pour créer ses règles et il ne considère pas la catégorie des mots adjacents, mais seulement le mot lui-même.

Le logiciel pourrait considérer le fait que plusieurs mots inconnus se suivent. Il est très rare qu'une suite de mots inconnus ne soit pas une suite de noms propres (ex : Phoebus/SBP:sg Aaron/SBP:sg Théodore/SBP:sg Levene/SBP:sg) ou de mots étrangers (ex : The/FGW making/FGW of/FGW Religion/FGW). Notre corpus contient 149 suites de deux mots inconnus, 16 suites de trois mots inconnus et neuf suites de quatre mots inconnus. Parmi les suites de deux mots inconnus, 16/149 (10,7 %) seulement sont des

suites de noms propres ou de mots étrangers. Pour les suites de trois mots, 4/16 (25 %) en sont. Pour les suites de quatre mots, 6/9 (66,7 %) en sont. On constate donc que plus il y a de mots inconnus regroupés ensemble, plus ils risquent d'être des noms propres ou des mots étrangers. De plus, dans cinq des neuf cas, il y avait aussi un mot adjacent au groupe qui était connu et avait été reconnu comme un nom propre ou un mot étranger. Il faudrait donc tirer profit de cette caractéristique.

Plusieurs erreurs portent sur les noms communs, les adjectifs, les verbes conjugués ou non, les participes présent et passé et certains adverbes. Les mots appartenant à ces catégories comportent une morphologie riche, mais souvent ambiguë. Premièrement, certaines erreurs sont dues à un mauvais traitement des suffixes lexicaux. Les mots formés à partir de certains d'entre eux sont ambigus (ex. : *-er* : nom ou verbe à l'infinitif). Or, le logiciel ne retient que la catégorie la plus fréquente pour un même suffixe (ex. : *-er* : verbe à l'infinitif). Il y a donc plusieurs erreurs causées par ces règles. On en voit un exemple en (35).

35. Exemple d'erreur causée par un suffixe lexical ambigu

> […] payer/VNCFF une/DTN:sg communication/SBC:sg téléphonique/ADJ:sg effectuées/PAR:pl par/PREP le/PRV:sg passager/VNCFF > SBC :sg pendant/PREP le/DTN:sg voyage/SBC:sg à/PREP partir/VNCFF d'/PREP un/DTN:sg téléphone/SBC:sg public/ADJ:sg situé/SBC:sg dans/PREP l'/DTN:sg autobus/PAR:pl ./.

La règle qui est intervenue dans ce cas-ci est : un mot ayant le suffixe *-er* est un verbe à l'infinitif. Le suffixe *-er* peut faire partie d'un nom, mais il fait plus souvent partie d'un verbe à l'infinitif. Les modèles de règles ne permettent pas d'alternative. Si le logiciel crée une règle spécifiant qu'un mot se terminant par *-er* est un substantif, cette règle sera écartée au profit de celle

qui spécifie qu'un tel mot est un verbe à l'infinitif car cette dernière permet de corriger plus d'erreurs que l'autre. Le logiciel a donc assigné la catégorie la plus fréquente pour les mots se terminant par ce suffixe à *passager*, soit verbe infinitif.

Deuxièmement, certaines erreurs viennent d'un traitement inadéquat des préfixes. Les préfixes aussi sont ambigus. Le préfixe *re-*, par exemple, peut faire partie d'un nom (*reconstruction*), d'un adjectif (*reconstructible*), d'un verbe conjugué ou non (*reconstruire*), d'un participe (*reconstruit*), etc. Le logiciel donnera aux mots comportant ou pouvant comporter un préfixe donné la catégorie la plus fréquemment associée aux mots comportant ce préfixe. On voit une erreur de ce type en (36).

36. Exemple d'erreur causée par un préfixe ambigu

> Il/PRV:sg y/PRV:++ aurait/ACJ:sg donc/ADV une/DTN:sg entropie/SBC:sg croissante/ADJ:sg qui/REL guetterait/VCJ:sg les/DTN:pl informations/SBC:pl transmises/SBC:pl > PAR :pl,/, au/DTC:sg cas/SBC:sg où/REL celles/PRO:pl -/- ci/ADV ne/ADV seraient/ECJ:pl pas/ADV freinées/PAR:pl par/PREP la/DTN:sg redondance/SBC:sg ./.

Dans cet exemple, le mot *transmises* est catégorisé substantif plutôt que participe passé. Le mot a changé de catégorie plusieurs fois pendant le traitement. Le logiciel lui a d'abord donné l'étiquette de défaut substantif singulier, puis une règle a confirmé son statut de substantif puisqu'il contient la lettre *-e-*, ensuite il est devenu substantif pluriel puisqu'il se termine par un *-s*, puis il est devenu adjectif pluriel puisqu'on peut lui supprimer le suffixe *-es* (*transmis*) et enfin il est redevenu substantif pluriel puisqu'on peut lui adjoindre le préfixe *re-* (*retransmises*). Il est sûr que la règle portant sur le préfixe ne permettra pas de traiter tous les mots contenant le préfixe *re-*. Encore une fois, les modèles de règles ne permettent

pas d'alternative. Des règles permettant d'assigner une autre catégorie que substantif pluriel aux mots commençant par *re-* seraient écartées au profit de celle qui les identifie comme substantif pluriel. Ce cas de figure étant le plus fréquent, le score de cette règle est supérieur à celui des autres règles concurrentes.

L'erreur que nous venons de voir fait aussi ressortir un autre défaut des règles lexicales : leur traitement des morphèmes grammaticaux de genre et de nombre. La présence d'un *-e-* serait un indice qu'il s'agit d'un nom singulier, la présence d'un *-s* qu'il s'agit d'un nom pluriel et la présence de *-es* qu'il s'agit d'un adjectif pluriel. Le logiciel tente en effet de trouver à quelle catégorie de mots associer un suffixe, c'est-à-dire qu'il donne aux mots la catégorie la plus fréquemment associée aux mots comportant tel suffixe. Or, les suffixes grammaticaux peuvent être associés à presque toutes les catégories de mots. Il est donc vain de tenter de déterminer la catégorie d'un mot sur la base du fait qu'il comporte ou peut comporter le suffixe *-s*, *-e*, ou *-es*.

Des erreurs surviennent du fait que le logiciel ne considère que des suffixes d'au maximum 4 lettres. En français, il y a beaucoup de suffixes de cinq et de six lettres *(-euses*, *-emment*, *-aires*, etc.). Un suffixe comme *-emment* n'est pas ambigu, c'est-à-dire qu'il sera toujours associé à un adverbe. L'erreur en (37) serait évitée si le logiciel considérait des suffixes de plus de quatre lettres.

37. Erreur due au fait que le logiciel ne considère que des suffixes d'au maximum quatre lettres

Banquiers/SBC:pl et/COO prêteurs/SBC:pl se/PRV:pl
voient/VCJ:pl fréquemment/SBC:sg >ADV abandonner/VNCFF
la/DTN:sg perception/SBC:sg des/DTC:pl impôts/SBC:pl
et/COO taxes/SBC:pl publiques/ADJ:pl ./.

Dans l'exemple ci-dessus, le mot *fréquemment* a été catégorisé substantif singulier plutôt qu'adverbe. La première règle à être intervenue a catégorisé *fréquemment* substantif singulier puisqu'il contient un *e*. Puis, une deuxième règle a changé cette catégorie pour verbe conjugué pluriel puisqu'il se termine par *-nt*. Enfin une autre règle a changé sa catégorie pour substantif singulier puisqu'il se termine par *-ment*. Les mots se terminant par *-ment* sont en effet majoritairement des substantifs. Toutefois, les mots se terminant par *-emment* ne sont jamais des substantifs.

D'autres erreurs sont dues au fait que le logiciel recherche des infixes. Or, il n'y en a pas en français. Le logiciel trouve tout de même des lettres qui apparaissent souvent à l'intérieur des mots d'une catégorie donnée et crée des règles indiquant qu'un mot contenant telle ou telle lettre appartient à cette catégorie. On voit un exemple de ce type d'erreur en (45).

38. Erreur due au fait que le logiciel recherche des infixes

Le/DTN:sg biochimiste/ADJ:sg allemand/SBC:sg
Albrecht/SBP:sg Kossel/SBC:sg >SBP :sg,/, se/PRV:sg
lança/SBC:sg alors/ADV dans/PREP l'/DTN:sg étude/SBC:sg
systématique/ADJ:sg de/PREP la/DTN:sg structure/SBC:sg
des/DTC:pl acides/ADJ:pl nucléiques/ADJ:pl ./.

Dans cet exemple, le mot *Kossel* a été catégorisé substantif commun plutôt que substantif propre. Dans un premier temps, le mot avait reçu la catégorie substantif propre puisqu'il commence par une majuscule. Puis une règle a changé sa catégorie pour substantif commun puisqu'il contient un *e*. S'il est vrai que beaucoup de nom commun contiennent un *e*, ce n'est toutefois pas un bon indice pour déterminer la catégorie d'un mot.

Certaines erreurs sont dues au fait que le logiciel tient compte de tous les mots du corpus de référence pour créer ses règles lexicales. Or, les règles

lexicales ne sont pas conçues pour catégoriser la totalité d'un texte, mais seulement les mots qui ne figurent pas au dictionnaire. Il est donc peu probable que les règles lexicales aient à traiter des mots très fréquents, dont les mots outils. Or, les règles qui obtiendront les meilleures scores et qui seront donc conservées seront celles qui traitent les mots de haute fréquence. Or, les mots de haute fréquence n'ont pas les mêmes caractéristiques morphologiques que les mots de basse fréquence. Les mots de spécialité, par exemple, ont une morphologie lexicale riche (ex. : *programmatique*) tandis que les mots outils n'en ont pas. On voit une erreur de ce type en (39).

39. Erreur due au fait que le logiciel tient compte de tous les mots du corpus de référence

> Par/PREP rapport/SBC:sg aux/DTC:pl enjeux/SBC:pl évoqués/PAR:pl ci/ADV -/- dessus/ADV ,/, service/SBC:sg rendu/PAR:sg au/DTC:sg citadin/SBC:sg ,/, économie/SBC:sg de/PREP système/SBC:sg ,/, réorganisation/SBC:sg territoriale/ADJ:sg ,/, la/DTN:sg carte/SBC:sg à/PREP mémoire/SBC:sg représente/VCJ:sg un/DTN:sg atout/ADV technologique/ADJ:sg très/ADV sérieux/ADJ:sg./.

Dans cet exemple, le mot *atout* a été catégorisé adverbe plutôt que substantif singulier. La règle à être intervenue spécifie qu'un mot auquel on peut retrancher le suffixe *-tout* est un adverbe. On peut bel et bien retrancher *-tout* de *atout*, et on trouvera le mot *a* au dictionnaire. Cette règle a été créée à cause de la présence, dans le corpus de référence, des adverbes *partout* et *surtout* qui ont une fréquence de 11 et de 29 respectivement. Toutefois, *-tout* n'est pas un morphème productif ; il ne donnera pas lieu à des mots de spécialité ou à des néologismes. Or, c'est, entre autre, ce type de mot que doivent traiter les règles lexicales et pas des mots comme *partout* et *surtout*.

Les cardinaux écrits en chiffres ne sont pas problématiques puisqu'on peut les catégoriser sur la base du fait qu'ils contiennent des caractères numériques. Mais les cardinaux écrits en lettres sont mal catégorisés. Comme ils font partie de listes finies, l'élargissement du corpus d'entraînement ou du dictionnaire règlerait le problème. On voit un exemple de catégorisation de cardinaux écrits en lettres et en chiffres en (40).

40. Exemples de catégorisation de cardinaux écrits en lettres et en chiffres

> […] l'/DTN:sg espérance/SBC:sg de/PREP vie/SBC:sg à/PREP la/DTN:sg naissance/SBC:sg des/DTC:pl femmes/SBC:pl est/ECJ:sg supérieure/ADJ:sg de/PREP huit/VCJ:sg > CAR ans/SBC:pl à/PREP celle/PRO:sg des/DTC:pl hommes/SBC:pl (/(80/CAR ,/, 3/CAR ans/SBC:pl au/DTC:sg lieu/SBC:sg de/PREP 72/CAR ,/, 0/CAR ans/SBC:pl en/PREP 1987/CAR)/) ./.

Dans cette phrase, le premier cardinal a été catégorisé verbe conjugué puisqu'il se termine par *-it* et le deuxième est bien catégorisé puisqu'il contient des caractères numériques.

Les meilleures performances sont obtenues pour les néologismes et les mots de spécialité puisqu'ils sont normalement constitués de morphèmes très productifs et au comportement régulier. On en voit des exemples en (41).

41. Exemples de bonne catégorisation de mots de spécialité

[…] montra/VCJ:sg par/PREP la/DTN:sg même/ADJ:sg occasion/SBC:sg l'/DTN:sg existence/SBC:sg de/PREP deux/CAR types/SBC:pl d'/PREP acides/ADJ:pl nucléiques/ADJ:pl :/: l'/DTN:sg acide/SBC:sg ribonucléique/ADJ:sg (/(ARN/SBC:sg)/) et/COO l'/DTN:sg acide/SBC:sg désoxyribonucléique/ADJ:sg (/(ADN/SBC:sg)/) ./.

3.2 Modifications vérifiées par implémentation

Comme nous l'avons déjà dit, la performance pour les mots inconnus est de 70,2 % (1094 /1558 mots inconnus), tandis que la performance pour les mots connus est de 97,8 % (6891 /7044 mots connus). Nous avons donc travaillé sur l'amélioration des règles lexicales uniquement. Par ailleurs, nous nous sommes concentrée sur le cas des mots pleins puisque les mots outils font partie de listes finies.

3.2.1 Les suffixes

La première modification effectuée portait sur le nombre de caractères des suffixes. Ce nombre était fixé à quatre pour l'anglais, mais les mots français sont souvent plus longs que les mots anglais, et les suffixes aussi (ex. : *-euses*, *-emment*, *-eraient*, *-ables*). Nous avons donc modifié la longueur maximale des suffixes. Des tests ont été effectués avec une longueur maximale de cinq, six et sept lettres. On voit dans le Tableau 5 que l'augmentation du nombre de caractères pouvant constituer un suffixe a amélioré un peu les résultats. Sans modification, c'est-à-dire avec des suffixes d'au maximum quatre lettres, le logiciel produisait 29,8 % d'erreurs portant sur des mots inconnus. En augmentant le nombre de lettres à cinq le

taux d'erreurs passe à 29,1 %, une diminution de 0,7 %. Puis avec des suffixes ayant au maximum six ou sept[23] lettres, les résultats passent à 29 %, une amélioration de 0,8 %. Cette modification a donc éliminé 2,6 % des erreurs (12 / 464).

Tableau 5 - **Modification de la longueur des suffixes**

Sans modification	Suffixe de 5 lettres	Suffixe de 6 lettres	Suffixe de 7 lettres
464 / 1558 29,8 %	453 / 1558 29,1 %	452 / 1558 29 %	452 / 1558 29 %

Avec la possibilité de considérer les suites de six lettres comme des suffixes, les règles qu'on voit en (42) ont été créées.

42. Exemples de règles lexicales produites avec la possibilité que les préfixes aient six lettres

a. iques hassuf 5 ADJ:pl
   Si le mot comporte le suffixe *-iques*, changer son étiquette pour adjectif pl.

b. aient hassuf 5 VCJ:pl
   Si le mot comporte le suffixe *-aient*, changer son étiquette pour verbe conjugué pl.

c. PAR:sg ement faddsuf 5 ADJ:sg
   Si le mot est catégorisé participe passé sg. et qu'on peut lui ajouter le suffixe *-ement*, changer son étiquette pour adjectif sg.

d. euses hassuf 5 ADJ:pl
   Si le mot comporte le suffixe *-euses*, changer son étiquette pour adjectif pl.

[23] Le logiciel n'a pas trouvé de suffixes de 7 lettres. Il n'est donc pas utile de considérer des suffixes de sept lettres.

e. SBC:sg mment fhassuf 5 ADV
   Si le mot est catégorisé substantif sg. et qu'il comporte le suffixe *-mment*, changer son étiquette pour adverbe.

f. illes hassuf 5 SBC:pl
   Si le mot comporte le suffixe *-illes*, changer son étiquette pour substantif pl.

g. naires deletesuf 6 SBC:pl
   Si on peut retrancher le suffixe *-naires* du mot, changer son étiquette pour substantif pl.

h. rieurs hassuf 6 ADJ:pl
   Si le mot comporte le suffixe *-rieurs*, changer son étiquette pour adjectif pl.

i. nement hassuf 6 SBC:sg
   Si le mot comporte le suffixe *-nement*, changer son étiquette pour substantif sg.

j. ville hassuf 5 SBP:sg
   Si le mot comporte le suffixe *-ville*, changer son étiquette pour nom propre sg.[24]

k. ieurs addsuf 5 ADV
   Si on peut ajouter le suffixe *-ieurs* au mot, changer son étiquette pour adverbe.

l. uième addsuf 5 CAR
   Si on peut ajouter le suffixe *-uième* au mot, changer son étiquette pour cardinal.

m. quels deletesuf 5 REL
   Si on peut retrancher le suffixe *-quels* au mot, changer son étiquette pour relatif.

On voit en (43) quelques erreurs qui se produisaient avant la modification. Ces erreurs ont été éliminées respectivement par l'ajout des règles 42e), 42h), 42e) et 42i).

[24] Bidonville serait la seule exception à cette règle.

43. Erreurs ayant été éliminées par la possibilité d'avoir des suffixes de six lettres

Sur/PREP un/DTN:sg ensemble/SBC:sg suffisamment/SBC:sg >ADV long/ADJ:sg ,/, une/DTN:sg stabilité/SBC:sg statistique/ADJ:sg des/DTC:pl choix/SBC:pl ultérieurs/SBC:pl >ADJ :pl se/PRV:pl produit/SBC:sg (/(processus/SBC:sg ergodique/ADJ:sg)/) ./.

Banquiers/SBC:pl et/COO prêteurs/SBC:pl se/PRV:pl voient/VCJ:pl fréquemment/SBC:sg >ADV abandonner/VNCFF la/DTN:sg perception/SBC:sg des/DTC:pl impôts/SBC:pl et/COO taxes/SBC:pl publiques/ADJ:pl ./.

Le/DTN:sg fait/SBC:sg n'/ADV est/ECJ:sg aucunement/SBC:sg > ADV établi/PAR:sg ./.

Nous avons aussi fait le test sur les préfixes, mais cela n'a pas amélioré les performances. En effet, il est difficile de trouver des préfixes de plus de quatre caractères en français et aucun n'a été trouvé.

3.2.2 Les infixes

La deuxième modification fut l'élimination de la recherche d'infixes. On retrouve quelques infixes dans les verbes anglais (ex. : t**a**ke / t**oo**k), mais il n'y a pas d'infixes en français. Nous avons donc éliminé la possibilité de créer des règles impliquant des infixes. On voit dans le Tableau 6 que l'élimination de la recherche d'infixes pour la création des règles lexicales a amélioré les performances de 0,3 %. Cette modification a donc éliminé 1,1 % (5 / 473) des erreurs.

Tableau 6 – **Élimination de la recherche d'infixes**

Sans modification	Sans infixe
464 / 1558 29,8 %	459 / 1558 29,5 %

En éliminant la recherche d'infixes, des règles comme celles qu'on voit en (44) n'ont plus été créées.

44. Exemples de règles éliminées par l'élimination de la possibilité de rechercher des infixes

a. e char SBC:sg
   Si le mot contient la lettre *e*, changer son étiquette pour substantif sg.

b. NN i fchar SBC:sg
   Si le mot est catégorisé NN[25] et qu'il contient la lettre *i*, changer son étiquette pour substantif sg.

c. 1 char CAR
   Si le mot contient le caractère *1*, changer son étiquette pour cardinal.

d. NNP a fchar SBP:sg
   Si le mot est catégorisé NNP[26] et qu'il contient la lettre *a*, changer son étiquette pour substantif propre sg.

e. I char SBC:sg
   Si le mot contient le caractère *I*, changer son étiquette pour substantif sg.

[25] L'étiquette NN est donnée comme première étiquette aux mots inconnus qui commencent par une minuscule.
[26] L'étiquette NNP est donnée comme première étiquette aux mots inconnus qui commencent par une majuscule.

f. B char SBP:sg
   Si le mot contient le caractère *B*, changer son étiquette pour nom propre sg.

g. NNP S fchar SBC:sg
   Si le mot est catégorisé NNP et qu'il contient le caractère *S*, changer son étiquette pour substantif sg.

h. X char ADJ:sg
   Si le mot contient le caractère *X*, changer son étiquette pour adjectif sg.

i. VNCFF S fchar SBP:sg
   Si le mot est catégorisé verbe infinitif et qu'il contient le caractère *S*, changer son étiquette pour nom propre sg.

j. à char ADV
   Si le mot contient le caractère *à*, changer son étiquette pour adverbe.

k. ' char PRV:sg
   Si le mot contient une apostrophe, changer son étiquette pour pronom verbal sg.

l. PAR:pl h fchar SBC:pl
   Si le mot est catégorisé participe passé pl. et qu'il contient le caractère *h*, changer son étiquette pour substantif pl.

m. SBC:sg w fchar ADJ:sg
   Si le mot est catégorisé substantif sg. et qu'il contient le caractère *w*, changer son étiquette pour adjectif sg.

On voit en (45) quelques erreurs ayant été éliminées par l'élimination de la règle 44a).

45. Exemples d'erreurs ayant été éliminées par l'élimination de la recherche d'infixes

Le/DTN:sg biochimiste/ADJ:sg allemand/SBC:sg
Albrecht/SBP:sg Kossel/SBC:sg >SBP :sg,/, se/PRV:sg
lança/SBC:sg alors/ADV dans/PREP l'/DTN:sg étude/SBC:sg
systématique/ADJ:sg de/PREP la/DTN:sg structure/SBC:sg
des/DTC:pl acides/ADJ:pl nucléiques/ADJ:pl ./.

Support/SBC:sg universel/SBC:sg >ADJ :sg de/PREP l'/DTN:sg
information/SBC:sg biologique/ADJ:sg ,/, la/DTN:sg
macromolécule/SBC:sg d'/PREP ADN/SBC:sg est/ECJ:sg
capable/ADJ:sg de/PREP se/PRV:++ dédoubler/VNCFF ;/;

Sur/PREP un/DTN:sg ensemble/SBC:sg suffisamment/SBC:sg
>ADV[27] long/ADJ:sg ,/, une/DTN:sg stabilité/SBC:sg
statistique/ADJ:sg des/DTC:pl choix/SBC:pl ultérieurs/SBC:pl
se/PRV:pl produit/SBC:sg (/(processus/SBC:sg
ergodique/ADJ:sg)/) ./.

3.2.3 La fréquence

La troisième modification fut l'élimination de la prise en compte des mots de haute fréquence pour la génération des règles lexicales. Nous avons apporté cette modification pour plusieurs raisons. Premièrement, les règles lexicales sont d'abord conçues pour catégoriser les mots pleins, pas les mots outils. Les mots outils font partie de listes finies, ils peuvent donc tous être dans le dictionnaire. Les mots outils étant de haute fréquence, nous avons éliminé les mots de haute fréquence. Deuxièmement, les mots courants sont souvent constitués de morphèmes peu productifs (ex. : les verbes irréguliers, enfant,

[27] Cette erreur a aussi été corrigée par la modification précédente. Toutefois, ici c'est bel et bien l'élimination de la recherche d'infixe qui a réglé le problème. Pour être en mesure de bien identifier les liens de cause à effet entre les modifications et les améliorations, nous avons apporté chaque modification séparément en repartant à chaque fois du module initial.

avant). Il est donc inutile d'observer leur forme pour tenter d'en déduire des règles (voir Baayen et Sproat 1996 pour une discussion en profondeur sur ce phénomène). Troisièmement, les mots inconnus ont plus de chance de se comporter comme des mots rares que comme des mots fréquents. Ils ont une fréquence de zéro dans le corpus d'entraînement; il est donc plus probable qu'ils s'apparentent aux mots ayant une basse fréquence qu'une haute fréquence. Nous avons donc fait plusieurs tests pour trouver quelle serait la fréquence maximale optimale. On voit dans le Tableau 7 que de considérer uniquement les mots de notre corpus d'entraînement ayant une fréquence de 20 ou moins a amélioré les performances de 1 %. Cette modification a donc éliminé 5,6 % (26 / 464) des erreurs[28].

Tableau 7 - **Élimination des mots de haute fréquence**

Sans modification	Fréquence maximale de 30	Fréquence maximale de 20	Fréquence maximale de 10	Fréquence maximale de 1
464/ 1558 29,8 %	450 / 1558 28,9 %	448 / 1558 28,8 %	450 / 1558 28,9 %	455 / 1558 29,2 %

Des règles inadéquates ont donc été éliminées. On peut en voir en (46).

46. Exemples de règles lexicales éliminées par l'élimination des mots de haute fréquence pour la création des règles lexicales

a. ADJ:sg f fdeletepref 1 SBC:sg
   Si le mot est catégorisé adjectif sg. et qu'on peut lui soustraire le préfixe *f-*, changer son étiquette pour substantif sg.
b. ADV sus faddsuf 3 DTC:pl
   Si le mot est catégorisé adverbe et qu'on peut lui ajouter le suffixe *-sus*, changer son étiquette pour déterminant contracté pl.

[28] Il faudrait étudier des corpus de différentes tailles pour créer une formule mathématique qui permettrait de déterminer automatiquement la fréquence optimale à utiliser.

c. auc addpref 3 DTN:sg
   Si on peut ajouter le préfixe *auc-* au mot, changer son étiquette pour déterminant sg.
d. ce deletepref 2 PRO:sg
   Si on peut soustraire le préfixe *ce-* au mot, changer son étiquette pour pronom sg.
e. cun addsuf 3 DTC:sg
   Si on peut ajouter le suffixe *-cun* au mot, changer son étiquette pour déterminant contracté sg.
f. DTN:pl vi faddpref 2 PRO:pl
   Si le mot est catégorisé déterminant pl. et qu'on peut lui ajouter le préfixe *vi-*, changer son étiquette pour pronom pl.
g. fi addpref 2 COO
   Si on peut ajouter le préfixe *fi-* au mot, changer son étiquette pour coordonnant.
h. PAR:sg ions faddsuf 4 PREP
   Si le mot est catégorisé participe passé sg. et qu'on peut lui ajouter le suffixe *-ions* , changer son étiquette pour préposition.
i. pris addsuf 4 PREP
   Si on peut ajouter le suffixe *-pris* au mot, changer son étiquette pour préposition.
j. PRO:sg t fchar ADV
   Si le mot est catégorisé pronom sg. et qu'il contient le caractère *t*, changer son étiquette pour adverbe.
k. SBC:pl mi faddpref 2 DTN:pl
   Si le mot est catégorisé substantif pl. et qu'on peut lui ajouter le préfixe *mi-*, changer son étiquette pour déterminant pl.
l. SBC:sg fon faddpref 3 DTC:sg
   Si le mot est catégorisé substantif sg. et qu'on peut lui ajouter le préfixe *fon-*, changer son étiquette pour déterminant contracté sg.
m. SBC:sg venu faddsuf 4 PREP
   Si le mot est catégorisé substantif sg. et qu'on peut lui ajouter le suffixe -*venu*, changer son étiquette pour préposition.

D'autre part, de nouvelles règles ont été créées, on peut en voir en (47).

47. Exemples de nouvelles règles créées suite à l'élimination des mots de haute fréquence pour la création des règles lexicales

a. ADJ:sg it fhassuf 2 VCJ:sg
   Si le mot est catégorisé adjectif sg. et qu'il a pour suffixe *-it*, changer son étiquette pour verbe conjugué sg.
b. ADJ:sg nt faddsuf 2 VCJ:sg
   Si le mot est catégorisé adjectif sg. et qu'on peut lui ajouter le suffixe *-nt*, changer son étiquette pour verbe conjugué sg.
c. age addsuf 3 SBC:sg
   Si on peut ajouter le suffixe *-age* au mot, changer son étiquette pour substantif sg.
d. ée addsuf 2 SBC:sg
   Si on peut ajouter le suffixe *-ée* au mot, changer son étiquette pour substantif sg.
e. ier addsuf 3 SBC:sg
   Si on peut ajouter le suffixe *-ier* au mot, changer son étiquette pour substantif sg.
f. ille hassuf 4 SBC:sg
   Si le mot comporte le suffixe *-ille*, changer son étiquette pour substantif sg.
g. int hassuf 3 VCJ:sg
   Si le mot comporte le suffixe *-int*, changer son étiquette pour verbe conjugué sg.
h. it addsuf 2 VCJ:sg
   Si on peut ajouter le suffixe *-it* au mot, changer son étiquette pour verbe conjugué sg.
i. ître hassuf 4 VNCFF
   Si le mot comporte le suffixe *-ître*, changer son étiquette pour verbe infinitif.
j. le hassuf 2 ADJ:sg
   Si le mot comporte le suffixe *-le*, changer son étiquette pour adjectif sg.
k. SBC:sg s fhassuf 1 SBC:pl
   Si le mot est catégorisé substantif sg. et qu'il comporte le suffixe *-s*, changer son étiquette pour substantif pl.
l. SBC:sg ut fdeletesuf 2 VCJ:sg
   Si le mot est catégorisé substantif sg. et qu'on peut lui soustraire le suffixe *-ut*, changer son étiquette pour verbe conjugué sg.
m. tte addsuf 3 SBC:sg
   Si on peut ajouter le suffixe *-tte* au mot, changer son étiquette pour substantif sg.

On voit en (48) des erreurs ayant été éliminées par cette modification. La première erreur a été éliminée par l'ajout de la règle 47i) et la deuxième par l'ajout de la règle 47j).

48. Exemples d'erreurs ayant été éliminées par l'élimination des mots de haute fréquence pour la création des règles lexicales

[…] elles/PRV:pl ont/ACJ:pl la/DTN:sg propriété/SBC:sg d'/PREP accroître/SBC:sg >VNCFF de/PREP façon/SBC:sg extrêmement/ADV sélective/ADJ:sg la/DTN:sg vitesse/SBC:sg de/PREP chacune/PRO:sg des/DTC:pl millions/SBC:pl de/PREP réactions/SBC:pl biochimiques/ADJ:pl qui/REL sont/ECJ:pl mises/PAR:pl en/PREP jeu/SBC:sg à/PREP chaque/DTN:sg instant/VNCNT dans/PREP les/DTN:pl organismes/SBC:pl vivants/ADJ:pl ./.

Certaines/DTN:pl épidémies/SBC:pl peuvent/VCJ:pl ,/, de/PREP même/ADJ:sg ,/, provoquer/VNCFF une/DTN:sg élévation/SBC:sg anormale/SBC:sg >ADJ :sg de/PREP l'/DTN:sg indice/SBC:sg de/PREP masculinité/SBC:sg à/PREP la/DTN:sg naissance/SBC:sg ;/;

3.2.4 Combinaison

Nous avons ensuite combiné les modifications s'étant avérées efficaces. Nous avons donc généré des règles lexicales en utilisant une fréquence maximale de 20, des préfixes de six lettres et aucun infixe. On voit dans le Tableau 8 que les performances ont été améliorées de 1,8 % par rapport à notre entraînement contrôle. Les modifications combinées diminuent donc le nombre d'erreurs de 5,8 % (27 / 464).

Tableau 8 - **Combinaison des modifications**

Sans modification	Fréquence maximale de 20, préfixe de 6 lettres et sans infixe
464 / 1558 29,8 %	437 / 1558 28 %

Finalement, nous avons voulu vérifier si nous obtenions des résultats similaires sur d'autres textes. Nous avons donc évalué les performances des règles lexicales issues de la combinaison de nos modifications sur un texte de cinq pages (3024 mots) traitant de sujets ne se retrouvant pas dans notre corpus d'entraînement. On voit dans le Tableau 9 que la version sans modification des règles lexicales a produit 28,1 % d'erreurs sur ce texte, contre 27,8 % pour la version modifiée, soit une amélioration de 0,3 %. Une diminution de 1,1 % des erreurs.

Tableau 9 - **Évaluation sur un deuxième texte**

Sans modification	Fréquence minimale de 20, préfixe de 6 lettres et sans infixe
183 / 652 28,1 %	181 / 652 27,8 %

L'amélioration est moins grande dans le cas du deuxième texte à cause de la nature des modifications apportée et des caractéristiques du deuxième texte. Les modifications ne visaient pas à améliorer le traitement des mots n'ayant pas de morphologie particulière (ex. : noms propres, certains adverbes, mots outils, mots étrangers[29]). Or, dans le texte ayant servi à la deuxième

[29] Les mots étrangers ont une morphologie particulière, mais le logiciel ne dispose pas de suffisamment de cas pour créer des règles portant sur les morphèmes étrangers.

évaluation, il y a une plus grande proportion de ce type de mots (16 % dans le premier texte et 27 % dans le deuxième texte). Ainsi, la portion d'erreurs susceptible d'être touchée par les modifications est moins grande dans le deuxième texte. De plus, comme le deuxième texte traite de sujets qui n'étaient pas traités dans le corpus d'entraînement, le dictionnaire ne contient pas le vocabulaire de base des domaines traités. Les nouvelles règles qui utilisent l'ajout ou la suppression de suffixes de plus de quatre lettres n'ont donc pas pu être appliquées aux mots de ce texte puisque les mots reliquats obtenus n'étaient pas au dictionnaire.

3.3 Modifications vérifiées par simulations

3.3.1 Les préfixes

Les modèles de règles concernant les préfixes consistent à enlever ou à ajouter un préfixe (ex. : reconstruire > construire), à vérifier si le mot résultant est dans le dictionnaire (ex. : construire : verbe), et à trouver dans le corpus d'entraînement la catégorie la plus fréquente pour les mots contenant ce préfixe (ex. : *re-* + X = 65 % nom). D'autres modèles de règles consistent à trouver la catégorie la plus fréquente des mots comportant tel ou tel préfixe. Étant donné que les préfixes, contrairement aux suffixes lexicaux, ne modifient jamais la catégorie du mot auquel ils s'adjoignent (ex. : construire : verbe; reconstruire : verbe) et qu'un même préfixe peut s'adjoindre à des mots de différentes catégories (ex. : **re**construire / verbe, **re**construction / nom, **re**constructible / adjectif, **re**construit / participe), il serait plus efficace de donner la catégorie du mot reliquat trouvé dans le dictionnaire, par exemple la catégorie de *construire* à *reconstruire*. Étant donné que les préfixes ne changent jamais la catégorie d'un mot, ce type de règles serait très efficace.

Prenons le cas du préfixe *re-* qui, comme nous l'avons vu, peut s'adjoindre à plusieurs types de mots. Notre premier texte d'évaluation contient 121 mots commençant par cette chaîne de caractères. 42 d'entre eux sont des mots inconnus. 12 d'entre eux ont été mal catégorisés. Parmi ceux-ci, il y en a quatre (*recense*, *recommande*, *reprend*, *retrouvé*) auxquels on peut soustraire le préfixe *re-* et trouver le reliquat dans le dictionnaire. Dans tous les cas, la catégorie des mots reliquats trouvée dans le dictionnaire est la bonne[30]. Les huit autres mots (ex. : recule, relief, religion, etc.) ne se retrouvent pas au dictionnaire sans le préfixe *re-*. La modification éliminerait donc quatre erreurs sur 12 et n'en provoquerait pas de nouvelle. Dans le cas du préfixe *in-*, 3/23 erreurs seraient éliminées et aucune ne serait ajoutée. En modifiant ainsi le traitement des préfixes, nous éliminerions un minimum de 1,8 % (7/464) des erreurs du premier texte. Dans le cas du deuxième texte, 1,1 % (2/183) des erreurs seraient éliminées par la modification du traitement de *in-* et, 2,2 % (4/183) par la modification du traitement de *re-*. Il est donc clair que si le logiciel appliquait ce traitement à tous les préfixes (*co-*, *en-*, *dé-*, etc.) les résultats seraient considérablement améliorés. De plus, le fonctionnement des préfixes est le même dans de très nombreuses langues. La modification proposée pourrait donc être valable pour d'autres langues.

3.3.2 Les suffixes grammaticaux

Le traitement des suffixes grammaticaux de genre, de nombre, de temps, de mode et de personne serait aussi à revoir. Premièrement, ils ne modifient jamais la catégorie d'un mot. Leur présence ou leur absence n'a pas d'impact direct sur la catégorie des mots. C'est pourquoi on peut donner à *étudiant, étudiants, étudiante et étudiantes* la même catégorie. De plus, les catégories de mots possédant un genre et un nombre sont trop nombreuses pour que

[30] Il en est de même pour les 30 mots déjà bien catégorisés.

leur seule présence nous permette d'identifier la catégorie du mot. C'est pourtant ce que tente de faire le logiciel. Les règles qui traitent les suffixes spécifient qu'un mot se terminant par un suffixe donné appartient à une catégorie donnée. Cette catégorie étant celle qui est la plus fréquemment associée aux mots comportant ce suffixe dans le corpus de référence. Il faudrait donc les supprimer (ex. : *étudiantes* - *es*) ou les ajouter (ex. : *étudiant* + *es*) et voir dans le dictionnaire si le mot formé existe. Puis il faudrait donner au mot de départ la catégorie du mot obtenu (ex. : *étudiant* + *es* : nom). Si un seul de ces mots est présent au dictionnaire, il sera facile de bien catégoriser les autres. Les nouvelles règles créées maximiseraient donc l'utilité du dictionnaire et produiraient peu ou pas d'erreurs.

1814 mots de notre premier texte d'évaluation se terminent par *-s*. Sur ce nombre, 415 sont inconnus et 135 de ceux-ci sont mal catégorisés. Pour 113 de ces 135 mots, le *-s* final représente bel et bien le pluriel et peut être enlevé sans que la catégorie du mot impliqué ne change. Pour les 22 autres mots, le *-s* n'est pas la marque du pluriel (ex. : *autobus*). Le logiciel ne trouvera qu'un de ces 22 mots sans *-s* au dictionnaire (*moi-s* : pronom), ce qui occasionnera une erreur. Parmi les 113 autres mots, le logiciel trouvera au dictionnaire 80 mots reliquats et tous auront la bonne catégorie. Cette modification permettrait donc d'éliminer 80 erreurs / 464 (17,2 %) au minimum puisqu'en appliquant le même traitement au suffixe *-e* et *-es* les résultats seraient similaires. Dans le cas du deuxième texte d'évaluation, la modification éliminerait 24 % (44/183) des erreurs. De plus, le fonctionnement des suffixes grammaticaux est le même dans de très nombreuses langues. La modification proposée pourrait donc être valable pour d'autres langues.

3.3.3 Les suffixes lexicaux

Il faudrait aussi revoir le traitement des suffixes lexicaux (ex. : *-ique*). Les mots formés à partir de certains d'entre eux sont ambigus (ex. : *-ique* peut donner lieu à un nom ou à un adjectif) et d'autres non (ex. : *-emment* : adverbe). De plus, ils modifient la catégorie des mots auxquels ils s'adjoignent (ex. : *mange*/verbe; *mangeable*/adjectif), et un même suffixe lexical donnera toujours lieu à des mots de mêmes catégories. Pour le moment, les règles obtenues ne permettent que d'attribuer la catégorie qui est le plus fréquemment associées aux mots comportant un suffixe donné. Car, une règle qui permettrait d'assigner une catégorie moins fréquente causerait plus d'erreurs qu'elle n'en résoudrait. Il faudrait que le logiciel vérifie lesquels des suffixes sont ambigus et pour résoudre leur ambiguïté, il faudrait ajouter une vérification du contexte. Les nouvelles règles seraient du type : un mot se terminant par *-ique* est adjectif si un nom le précède. En vérifiant deux critères plutôt qu'un les règles donneraient de meilleurs résultats.

Notre premier texte d'évaluation contient 123 mots se terminant par *-ique*. 46 sont inconnus et 10 sont mal catégorisés. La catégorie donnée à un mot ayant ce suffixe est adjectif, ce qui est valable dans 36 cas, mais dans huit cas elle aurait dû être substantif. Les deux autres cas sont les mots *quoique* et *revendique*, subordonnant et verbe conjugué qui sont mal catégorisés et le resteront après la modification. Dans ces huit cas le contexte aurait permis de bien catégoriser le mot. En effet, dans tous les cas le mot était précédé d'un déterminant et/ou suivi d'un adjectif, et dans aucun cas il n'y avait de nom à proximité. 8/464 (1,7 %) erreurs auraient donc pu être corrigées par cette modification. Dans le cas du deuxième texte d'évaluation 0,5 % (1/183) auraient été éliminées par cette modification. Le fonctionnement des suffixes

lexicaux est lui aussi le même dans de très nombreuses langues, la modification proposée pourrait donc être valable pour d'autres langues.

3.4 Récapitulation des résultats obtenus

Dans le Tableau 10, on voit que la modification du traitement des préfixes *re-* et *in-* permet d'éliminer 1,5 % des erreurs sur le premier texte et 3,3 % sur le deuxième texte; la modification du traitement du suffixe grammatical *-s* 17,2 % des erreurs sur le premier texte et 24 % sur le deuxième texte; la modification du traitement du suffixe lexical *-ique* 1,7 % des erreurs sur le premier texte et 0,5 % sur le deuxième texte. Ces modifications permettent donc d'éliminer 20,4 % des erreurs sur le premier texte et 27,9 % sur le deuxième texte. 5,8 % des erreurs ont été éliminées par les modifications effectuées par implémentation sur le premier texte, contre 1,1 % sur le deuxième texte. Au total, les modifications ont permis d'éliminer 26,3 % des erreurs sur le premier texte et 29 % sur le deuxième texte.

Tableau 10 – **Tableau récapitulatif des améliorations effectuées**

	Premier texte		Deuxième texte	
Préfixes				
re-	4/464	0,9 %	4/183	2,2 %
in-	3/464	0,6%	2/183	1,1 %
Suffixes grammaticaux				
-s	80/464	17,2 %	44/183	24 %
Suffixes lexicaux				
-ique	8/464	1,7 %	1/183	0,5 %
Implémentation	27/464	5,8 %	2/183	1,1 %
Total	122/464	26,3 %	53/183	29 %

Il ne reste donc que 343 erreurs portant sur les 1558 mots inconnus du premier texte pour une efficacité de 78% et 130 erreurs sur les 652 mots inconnus du deuxième texte pour une efficacité de 81,1%. La performance globale, pour les deux textes, qui était au départ de 70,7 % est maintenant

de 78,6 % (voir le Tableau 11). Les résultats se rapprochent donc de ceux obtenus par Brill 1994 qui étaient de 82 %.

Tableau 11 – **Performance après modifications**

	Évaluation contrôle		Évaluation après modification	
1er texte	464/1558	29,8 %	343/1558	22 %
2e texte	183/652	28,1 %	130/652	19,9 %
Total	647/2210	29,3 %	473/2210	21,4 %

Conclusion

Après avoir vu brièvement le type d'ambiguïté que l'approche de Brill résout, soit l'ambiguïté grammaticale, nous avons vu les différentes approches utilisées pour faire de la catégorisation grammaticale automatique. Il existe deux grands types d'approches : les approches supervisées et les approches non supervisées. Les approches supervisées utilisent des corpus précatégorisés pour apprendre à catégoriser de nouveaux textes. Les logiciels fondés sur une approche supervisée utilisent des algorithmes qui leur permettent d'apprendre à reproduire la catégorisation du texte exemple. Les approches non supervisées ont un fonctionnement très différent. L'apprentissage se fait à partir de corpus non catégorisés, par une analyse distributionnelle. Les logiciels utilisant cette approche doivent donc d'abord déterminer quelles sont les classes de mots pour ensuite apprendre à bien catégoriser du texte. Ces deux grandes approches se subdivisent encore en trois : catégoriseurs à base de règles, catégoriseurs statistiques et catégoriseurs neuronaux. La première grande famille regroupe les catégoriseurs qui utilisent et apprennent des règles. Ces règles portent normalement sur le contexte et la morphologie. Les catégoriseurs statistiques n'utilisent pas de règles, mais compilent plutôt des statistiques sur les mots et les suites d'étiquettes contenues dans un texte. Une fois entraînés, ils peuvent déterminer la catégorie d'un mot en fonction de la probabilité que ce mot ait telle ou telle catégorie et que la séquence d'étiquettes obtenue apparaisse. Les catégoriseurs neuronaux sont eux implémentés dans des réseaux neuronaux qui sont conçus pour apprendre des fonctions « intelligentes ». Ils prennent du texte en entrée et vont moduler le poids à donner à chaque élément pour que le résultat soit le bon. Les résultats sont sensiblement les mêmes pour les trois approches (96-97 %).

Nous avons ensuite vu le défi que représente la catégorisation des mots inconnus quelle que soit l'approche utilisée. Ils ne sont bien catégorisés que dans 85-88 % des cas. La difficulté est plus grande dans le cas de mots inconnus, car ils peuvent appartenir à toutes les catégories et non pas seulement à deux ou trois d'entre elles. Leur morphologie, qui est souvent ambiguë, est utilisée pour les désambiguïser ainsi que le contexte dans certains cas.

Puis nous avons vu le détail du fonctionnement du catégoriseur de Brill qui est un catégoriseur supervisé, à base de règles, utilisant la fréquence. Le logiciel de Brill prend en entrée du texte catégorisé à partir duquel il crée un dictionnaire, des règles contextuelles et des règles lexicales. Il range dans le dictionnaire tous les mots du texte avec leurs catégories possibles, la plus fréquente étant placée en tête de liste. Puis, pour créer les règles, il enlève les étiquettes du texte. Ensuite, il assigne aux mots du texte leur catégorie de défaut. Puis il crée des règles pour corriger le résultat en se basant sur des modèles de règles. Une fois l'apprentissage terminé, le logiciel dispose de tout ce dont il a besoin pour catégoriser de nouveaux textes.

Ensuite, nous avons vu la méthodologie utilisée dans cette recherche. Nous avons entraîné le logiciel sur 150 pages de texte tiré de 15 *Que Sais-je ?* et nous avons évalué les performances sur un premier texte de 15 pages tiré des mêmes *Que Sais-je ?* et sur un deuxième texte de cinq pages tiré de cinq autres *Que Sais-je ?*. Pour évaluer les performances du logiciel, nous avons simplement compté le nombre d'erreurs produites. Nous avons considéré qu'il y avait une erreur quand l'étiquette donnée à un mot ne correspondait pas à la définition donnée. À partir de cette évaluation, nous avons établi une typologie des erreurs. Les erreurs peuvent être regroupées en deux grandes classes. Il y a d'abord les erreurs qui portent sur les mots connus et

représentent 24,8 % des erreurs. Puis il y a les erreurs qui portent sur des mots inconnus qui représentent 75,2 % des erreurs.

Après avoir analysé les erreurs, nous avons effectué des modifications soit par implémentation, soit par simulation pour augmenter la performance des règles lexicales qui était de seulement 70,7 %. Les modifications effectuées par implémentation ont été de considérer des suffixes d'au maximum six lettres plutôt que de quatre lettres puisque les suffixes français sont plus longs que les suffixes anglais ; éliminer la recherche d'infixes puisqu'il n'y en a pas en français ; n'utiliser que les mots de basse fréquence, les mots de haute fréquence ayant une morphologie très différente des mots inconnus qui sont par définition des mots rares. La combinaison de ces modifications a permis d'éliminer 5,8 % des erreurs dans le premier texte d'évaluation. Puis nous avons évalué les performances des nouvelles règles sur le deuxième texte d'évaluation. Dans ce texte, 1,1 % des erreurs ont été éliminées. Nous croyons que l'écart est moins grand dans le cas du deuxième texte parce que celui-ci contenait plus de mots inconnus ne pouvant être touchés par les modifications (noms propres, mots étrangers, sigles, etc.) et que le logiciel ne disposait pas du vocabulaire de base des domaines traités. Les modifications effectuées par simulations ont été de modifier le traitement des préfixes, des suffixes lexicaux et des suffixes grammaticaux en fonction de leurs caractéristiques propres. Ces modifications permettent d'éliminer 20,4 % des erreurs contenues dans le premier texte et 27,9 % des erreurs contenues dans le deuxième texte. La performance des règles lexicales passe donc à 78 % sur le premier texte et à 81,1 % sur le deuxième texte. Les résultats obtenus se rapprochent de ceux obtenus par Brill 1994 (82 %).

Notre objectif était d'adapter l'approche de Brill au français et d'améliorer les résultats des règles lexicales. Cet objectif est en partie atteint bien qu'il reste encore beaucoup de travail à faire avant que le catégoriseur soit aussi

efficace que pour l'anglais et avant que les performances des règles lexicales soient satisfaisantes. Nous espérons que notre contribution pourra être reprise et servir dans un contexte où tous les éléments seraient là pour produire un vrai bon catégoriseur pour le français.

D'autres modifications pourraient être apportées qui amélioreraient fort probablement l'approche. L'ajout d'un module qui générerait des contraintes capables de bloquer l'application de certaines règles dans un contexte donné serait probablement profitable. Le jeu d'étiquettes pourrait aussi être travaillé. La catégorisation des adjectifs postnominaux et prénominaux serait probablement facilitée si les différents types d'adjectifs (ex : adjectif de couleur de forme, etc.) portaient des étiquettes distinctes. Finalement, la considération de sous-catégories pourrait aussi être profitable. Pour le moment le logiciel considère les substantifs singuliers comme des entités totalement différentes des substantifs pluriels. Il serait plus profitable de considérer le nombre comme une sous-catégorie.

Bibliographie

Atwell, E. (1983), “Constituent-Likelihood Grammar” in *ICAME Journal*, vol. 7.

Baayen, H. et R. Sproat (1996), « Estimating Lexical Priors for Low-Frequency Morphologically Ambiguous Forms » in *Computational Linguistics*, vol. 22, no 2, p. 155-166.

Brill, E. (1995a), « Unsupervised Learning of Disambiguation Rules for Part of speech Tagging » in *Natural Language Processing Using Very Large Corpora*, p. 1-13.

Brill, E. (1995b), « Transformation-Based Error-Driven Learning and Natural Language Processing : A Case study in Part of Speech Tagging » in *Computational linguistics*, vol. 21, no 4, p. 543-565.

Brill, E. (1994), « A Report of Recent Progress in Transformation-Based Error-Driven Learning » in *Proceedings of the Twelfth National Conference on Artificial Intelligence (AAAI-94)*, Seattle, WA.

Brill, E. (1993), *A Corpus-Based Approach to Language Learning*, University of Pennsylvania, Department of Computer and Information Science, Ph.D. Thesis, IRCS Report 93-44, 166 p.

Brill, E. (1991), « Discovering the Lexical Features of a Language » in Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics, Berkeley, CA.

Brill, E. et M. Marcus (1993), « Tagging an unfamiliar text with minimal human supervision » in *Proceedings of the Fall Symposium on Probabilistic Approaches to Natural Language*, AAAI.

Church, K. W. (1985), « A stochastic parts program and noun phrase parser for unrestricted text » in *Proceedings of the Second Conference on Applied Natural Language Processing*, p. 136-143.

Cutting, D., A. Kupiec, A. Pedersen et P. Sibun (1992), « A practical part-of-speech tagger » in *Proceedings of the Third Conference on Applied Natural Language Processing*, Trento, Italy (ACL), pages 133-140.

Daille, B. et E. Morin (2000), « Reconnaissance automatique des noms propres de la langue écrite : les récentes réalisations » in *Tal Traitement automatique des langues pour la recherche d'information*, vol. 41, no 2, p. 601-622.

Franz, A. (1996), « Automatic Ambiguity Resolution in Natural Language Processing : An Empirical Approach » in *Lecture Notes in Artificial Intelligence*, no 1171, p. 155.

Freeman, A., *Brill's POS tagger and a Morphology parser for Arabic*, En ligne, 7 p. <http://www.elsnet.org/arabic2001/freeman.pdf>, Consultée le 11 août 2003.

Greene, B. et G. M. Rubin (1971), « Automatic grammatical tagging of English » in *technical report*, Department of Linguistics, Brown University, Providence, Rhode Island.

Grevisse, M. (1969), *Le bon usage : grammaire française*, ed. Duculot Gembloux, Paris, 1228 p.

Helmut, S. (1994), « Part-of-Speech Tagging with Neural Networks » in *Proceedings of the International Conference on Computational Linguistics*, Kyoto, Japan, p. 172-176.

Jelinek, F. (1985), « Markov Source modeling of text generation » in *J.K. Skwirzinski Ed., Impact of Processing Techniques on Communication*, Nijhoff, Dordrecht.

Kempe, A. (1993), « A stochastic Tagger and an Analysis of Tagging Errors » in *Internal paper. Institute for Computational Linguistics*, University of Stuttgart.

Lecomte J. (1998, décembre), *LE CATÉGORISEUR BRILL14-JL5 / WINBRILL-0.3*, En ligne, 36 p. <http://www.inalf.fr/winbrill/BRILL14-JL5_WinBrill.doc>, Consulté le 1 mars 2004.

Lippmann, R. P. (1989), « Review of Neural Networks for Speech Recognition » in *Neural Computation*, vol. i, p. 1-38.

Marques, N. C. et G. P. Lopes (1996), « A Neural Network Approach to Part-of-Speech Tagging » in *Proceedings of the Second Workshop on Spoken and Written Portuguese*, Curitiba, Brazil, p. 1-9.

Megyesi, B. (1998), *Brill's Rule-Based Part of Speech Tagger for Hungarian*, Stockholm University, Master's thesis, 120 p.

Megyesi, B. (1999a), « Improving Brill's PoS Tagger for an Agglutinative Language » in Proceedings of the Joint Sigdat Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC '99), University of Maryland, USA, p. 275-284.

Megyesi, B. (1999b), « Brill's PoS Tagger with Extended Lexical Templates for Hungarian » in *Proceedings of the Workshop (W01) on Machine Learning in Human Language Technology*, ACAI'99, Chania, Crete, Greece, p. 22-28.

Mikheev, A. (1996), « Unsupervised learning of part-of-speech guessing rules » in *Natural Language Engineering*, Cambridge University Press, vol. 2, no 2, p. 111-136.

Nakagawa, T., T. Kudoh et Y. Matsumoto (2001), « Unknown Word Guessing and Part-of-Speech Tagging Using Support Vector Machines » in *Proceedings of the 6th Natural Language Processing Pacific Rim Symposium (NLPRS2001)*.

Nouveau Petit Robert électronique (1996), Dictionnaires Le Robert.

Prütz, K. (2002), « Part-of-speech tagging for Swedish », in *Parallel Corpora, Parallel Worlds*, University, Sweden, p. 201-206.

Rabiner, L. R. (1989), « A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition » in *Proceedings of the IEEE*, vol. 7, no 2, p. 257-286.

Schütze, H. (1993), « Part-of-speech induction from scratch » in Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics, p. 251-258.

Vasilakopoulos, A. (2003), « Improved Unknown Word Guessing by Decision Tree Induction for POS Tagging with TBL » in *Proceedings of CLUK 2003*, Edinburgh.

Viterbi, A. J. (1967), « Error bounds for convolutional codes and asymptotically optimal decoding algorithm » in *IEEE Transactions on Information Theory*, no 13, p. 260-269.

Annexe A
Étiquettes grammaticales utilisées

Étiquettes	Catégories
SBC :sg/pl	Substantif commun (singulier ou pluriel)
SBP :sg/pl	Substantif propre (singulier ou pluriel)
ADJ :sg/pl	Adjectif (singulier ou pluriel)
ACJ :sg/pl	Verbe avoir conjugué (singulier ou pluriel)
ANCFF	Verbe avoir à l'infinitif
ANCNT	Verbe avoir au participe présent
APAR :sg/pl	Verbe avoir au participe passé (singulier ou pluriel)
ECJ :sg/pl	Verbe être conjugué (singulier ou pluriel)
ENCFF	Verbe être à l'infinitif
ENCNT	Verbe être au participe présent
EPAR :sg	Verbe être au participe passé (singulier, pas de pluriel)
VCJ :sg/pl	Autre verbe conjugué (singulier ou pluriel)
VNCFF	Autre verbe à l'infinitif
VNCNT	Autre verbe au participe présent
PAR :sg/pl	Autre verbe au participe passé (singulier ou pluriel)
ADV	Adverbe
PREP	Préposition
DTN :sg/pl	Déterminant non contracté (singulier ou pluriel)
DTC :sg/pl	Déterminant contracté (singulier ou pluriel)
PRV :sg/pl/++	Pronom verbal (singulier, pluriel ou nombre indéterminé)
PRO :sg/pl/++	Pronom non verbal (singulier, pluriel ou nbr. indéterminé)
COO	Coordonnant
REL	Relatif
SUB	Subordonnant
CAR	Cardinal (en chiffres ou en lettres)
ABR	Abréviation
FGW	Mot étranger
INJ	Interjection, Onomatopée, etc.
PFX	Préfixe détaché
SYM	Symbole ou Signe mathématique
PUL	Particule non indépendante

Annexe B
Extrait du corpus catégorisé servant à l'entraînement

Introduction/SBC:sg Qu'/PRO:sg est/ECJ:sg -/- ce/PRV:sg Que/PRO:sg LA/DTN:sg BIOASTRONOMIE/SBC:sg ?/?

I/CAR ./. -/-

Définition/SBC:sg Le/DTN:sg but/SBC:sg de/PREP la/DTN:sg bioastronomie/SBC:sg ,/, dont/REL un/DTN:sg synonyme/SBC:sg est/ECJ:sg "/" l'/DTN:sg exobiologie/SBC:sg "/" ,/, est/ECJ:sg la/DTN:sg recherche/SBC:sg et/COO l'/DTN:sg étude/SBC:sg scientifiques/ADJ:pl de/PREP la/DTN:sg vie/SBC:sg dans/PREP l'/DTN:sg univers/SBC:sg sous/PREP toutes/DTN:pl ses/DTN:pl formes/SBC:pl ./.

Les/DTN:pl approches/SBC:pl de/PREP ce/DTN:sg vaste/ADJ:sg domaine/SBC:sg sont/ECJ:pl très/ADV nombreuses/ADJ:pl et/COO font/VCJ:pl appel/SBC:sg à/PREP des/DTC:pl champs/SBC:pl de/PREP recherche/SBC:sg extrêmement/ADV variés/ADJ:pl ,/, qui/REL recouvrent/VCJ:pl notamment/ADV des/DTC:pl recherches/SBC:pl d'/PREP ordre/SBC:sg plus/ADV "/" terrestre/ADJ:sg "/" ,/, depuis/PREP l'/DTN:sg étude/SBC:sg de/PREP l'/DTN:sg origine/SBC:sg de/PREP la/DTN:sg vie/SBC:sg et/COO son/DTN:sg évolution/SBC:sg sur/PREP Terre/SBC:sg ,/, jusqu'/PREP à/PREP celle/PRO:sg des/DTC:pl structures/SBC:pl et/COO des/DTC:pl processus/SBC:pl qui/REL lui/PRV:sg sont/ECJ:pl associés/PAR:pl ,/, y/PRV:++ compris/ADJ:sg au/DTC:sg voisinage/SBC:sg d'/PREP autres/ADJ:pl étoiles/SBC:pl ./.

Bien/ADV sûr/ADJ:sg ,/, l'/DTN:sg exploration/SBC:sg in/FGW situ/FGW est/ECJ:sg la/DTN:sg voie/SBC:sg la/DTN:sg plus/ADV directe/ADJ:sg ,/, mais/COO elle/PRV:sg fait/VCJ:sg appel/SBC:sg à/PREP des/DTC:pl

moyens/SBC:pl technologiques/ADJ:pl -/- principalement/ADV spatiaux/ADJ:pl -/- qui/REL ne/ADV sont/ECJ:pl disponibles/ADJ:pl que/SUB depuis/PREP peu/ADV ,/, et/COO qui/REL restent/VCJ:pl encore/ADV pour/PREP de/PREP nombreuses/ADJ:pl décennies/SBC:pl limités/ADJ:pl au/DTC:sg système/SBC:sg solaire/ADJ:sg ./.

En/PREP dehors/SBC:sg du/DTC:sg système/SBC:sg solaire/ADJ:sg ,/, seules/ADJ:pl les/DTN:pl approches/SBC:pl indirectes/ADJ:pl peuvent/VCJ:pl être/ENCFF utilisées/PAR:pl :/: recherche/SBC:sg d'/PREP une/DTN:sg vie/SBC:sg primitive/ADJ:sg ou/COO même/ADV d'/PREP "/" intelligences/SBC:pl "/" extraterrestres/ADJ:pl ./.

En/PREP bref/ADJ:sg ,/, les/DTN:pl régions/SBC:pl de/PREP l'/DTN:sg univers/SBC:sg où/REL la/DTN:sg vie/SBC:sg est/ECJ:sg recherchée/PAR:sg ,/, vont/VCJ:pl des/DTC:pl objets/SBC:pl du/DTC:sg système/SBC:sg solaire/ADJ:sg (/(les/DTN:pl comètes/SBC:pl ,/, les/DTN:pl planètes/SBC:pl et/COO leurs/DTN:pl satellites/SBC:pl)/) aux/DTC:pl galaxies/SBC:pl lointaines/ADJ:pl en/PREP passant/VNCNT par/PREP le/DTN:sg milieu/SBC:sg interstellaire/ADJ:sg et/COO les/DTN:pl planètes/SBC:pl en/PREP orbite/SBC:sg autour/ADV des/DTC:pl étoiles/SBC:pl de/PREP notre/DTN:sg galaxie/SBC:sg ./.

Parallèlement/ADV à/PREP l'/DTN:sg extension/SBC:sg spatiale/ADJ:sg du/DTC:sg champ/SBC:sg d'/PREP investigation/SBC:sg ,/, une/DTN:sg réflexion/SBC:sg sur/PREP la/DTN:sg nature/SBC:sg et/COO la/DTN:sg définition/SBC:sg de/PREP la/DTN:sg vie/SBC:sg ,/, sur/PREP d'/DTN:pl autres/ADJ:pl formes/SBC:pl de/PREP vie/SBC:sg possibles/ADJ:pl s'/PRV:sg impose/VCJ:sg ./.

En/PREP effet/SBC:sg ,/, il/PRV:sg est/ECJ:sg indispensable/ADJ:sg ,/, afin/PUL de/PREP ne/ADV pas/ADV passer/VNCFF à/PREP côté/SBC:sg

de/PREP types/SBC:pl d'/PREP organisation/SBC:sg biologique/ADJ:sg intéressants/ADJ:pl ,/, de/PREP ne/ADV pas/ADV se/PRV:++ limiter/VNCFF aux/DTC:pl apparences/SBC:pl du/DTC:sg seul/ADJ:sg exemple/SBC:sg que/SUB nous/PRV:pl connaissons/VCJ:pl ,/, la/DTN:sg vie/SBC:sg terrestre/ADJ:sg ./.

II/CAR ./. -/-

Historique/SBC:sg La/DTN:sg question/SBC:sg de/PREP l'/DTN:sg existence/SBC:sg de/PREP formes/SBC:pl de/PREP vie/SBC:sg ailleurs/ADV que/SUB sur/PREP la/DTN:sg Terre/SBC:sg est/ECJ:sg la/DTN:sg plus/ADV ancienne/ADJ:sg des/DTC:pl questions/SBC:pl scientifiques/ADJ:pl bien/ADV posées/PAR:pl ./.

Elle/PRV:sg reste/VCJ:sg à/PREP ce/DTN:sg jour/SBC:sg non/ADV résolue/PAR:sg ./.

Nous/PRV:pl sommes/ECJ:pl toutefois/ADV à/PREP une/DTN:sg époque/SBC:sg tout/ADV à/PREP fait/SBC:sg particulière/ADJ:sg car/COO ,/, grâce/SBC:sg notamment/ADV aux/DTC:pl missions/SBC:pl spatiales/ADJ:pl ,/, elle/PRV:sg sort/VCJ:sg d'/PREP une/DTN:sg phase/SBC:sg de/PREP latence/SBC:sg de/PREP quelques/DTN:pl millénaires/SBC:pl ./.

On/PRV:sg en/PRV:++ trouve/VCJ:sg ,/, semble/VCJ:sg -/- t/PUL -/- il/PRV:sg ,/, la/DTN:sg première/ADJ:sg mention/SBC:sg dans/PREP un/DTN:sg texte/SBC:sg d'/PREP Épicure/SBP:sg ,/, la/DTN:sg Lettre/SBC:sg à/PREP Hérodote/SBP:sg en/PREP 300/CAR av/ABR ./.

J/ABR ./. -/-

C/ABR ./.

Aujourd'hui/ADV la/DTN:sg question/SBC:sg subsiste/VCJ:sg :/: "/" Sommes/ECJ:pl -/- nous/PRV:pl seuls/ADJ:pl dans/PREP l'/DTN:sg univers/SBC:sg ?/? "/"

Depuis/PREP Épicure/SBP:sg ,/, des/DTC:pl dizaines/SBC:pl d'/PREP auteurs/SBC:pl ont/ACJ:pl réfléchi/PAR:sg à/PREP cette/DTN:sg question/SBC:sg ,/, de/PREP Giordano/SBP:sg Bruno/SBP:sg à/PREP Flammarion/SBP:sg en/PREP passant/VNCNT par/PREP Kepler/SBP:sg ,/, Fontenelle/SBP:sg ,/, Kant/SBP:sg ,/, Goethe/SBP:sg ./.

Huygens/SBP:sg ,/, au/DTC:sg lieu/SBC:sg de/PREP s'/PRV:sg en/PRV:++ tenir/VNCFF ,/, comme/SUB tous/DTN:pl ces/DTN:pl essayistes/SBC:pl ,/, à/PREP de/PREP fortes/ADJ:pl convictions/SBC:pl ,/, a/ACJ:sg été/EPAR:sg le/DTN:sg premier/SBC:sg à/PREP se/PRV:++ poser/VNCFF ,/, en/PREP astronome/SBC:sg observateur/SBC:sg ,/, la/DTN:sg question/SBC:sg :/: "/" Comment/ADV faire/VNCFF pour/PREP voir/VNCFF ces/DTN:pl autres/ADJ:pl mondes/SBC:pl ?/? "/"

Une/DTN:sg nouvelle/ADJ:sg étape/SBC:sg a/ACJ:sg été/EPAR:sg franchie/PAR:sg à/PREP la/DTN:sg fin/SBC:sg du/DTC:sg siècle/SBC:sg dernier/ADJ:sg par/PREP l'/DTN:sg astronome/SBC:sg franco/PFX -/- danois/ADJ:sg Jules/SBP:sg Janssen/SBP:sg ./.

De/PREP façon/SBC:sg très/ADV audacieuse/ADJ:sg ,/, il/PRV:sg a/ACJ:sg proposé/PAR:sg de/PREP rechercher/VNCFF ,/, par/PREP l'/DTN:sg analyse/SBC:sg spectroscopique/ADJ:sg ,/, des/DTC:pl composés/SBC:pl organiques/ADJ:pl dans/PREP les/DTN:pl autres/ADJ:pl planètes/SBC:pl du/DTC:sg système/SBC:sg solaire/ADJ:sg ./.

Dans/PREP les/DTN:pl années/SBC:pl 1960/CAR ,/, Van/SBP:sg de/PREP Kamp/SBP:sg a/ACJ:sg cru/PAR:sg avoir/ANCFF trouvé/PAR:sg de/PREP

forts/ADJ:pl indices/SBC:pl d'/PREP existence/SBC:sg d'/PREP une/DTN:sg planète/SBC:sg autour/ADV de/PREP l'/DTN:sg étoile/SBC:sg de/PREP Barnard/SBP:sg ,/, espoir/SBC:sg qui/REL a/ACJ:sg été/EPAR:sg démenti/PAR:sg par/PREP la/DTN:sg suite/SBC:sg ./.

À/PREP partir/VNCFF des/DTC:pl années/SBC:pl 1950/CAR ,/, a/ACJ:sg commencé/PAR:sg la/DTN:sg conjonction/SBC:sg de/PREP plusieurs/DTN:pl champs/SBC:pl :/: l'/DTN:sg astronomie/SBC:sg d'/PREP une/DTN:sg part/SBC:sg et/COO les/DTN:pl recherches/SBC:pl sur/PREP l'/DTN:sg origine/SBC:sg de/PREP la/DTN:sg vie/SBC:sg sur/PREP Terre/SBC:sg d'/PREP autre/ADJ:sg part/SBC:sg ont/ACJ:pl convergé/PAR:sg vers/PREP la/DTN:sg création/SBC:sg d'/PREP une/DTN:sg discipline/SBC:sg consacrée/PAR:sg à/PREP l'/DTN:sg étude/SBC:sg de/PREP la/DTN:sg vie/SBC:sg dans/PREP l'/DTN:sg univers/SBC:sg ./.

Dans/PREP les/DTN:pl années/SBC:pl 1960/CAR ,/, le/DTN:sg biologiste/SBC:sg américain/ADJ:sg Joshua/SBP:sg Lederberg/SBP:sg enrichissait/VCJ:sg le/DTN:sg vocabulaire/SBC:sg scientifique/ADJ:sg en/PREP inventant/VNCNT le/DTN:sg mot/SBC:sg "/" exobiologie/SBC:sg "/" ,/, terme/SBC:sg permettant/VNCNT l'/DTN:sg envolée/SBC:sg de/PREP la/DTN:sg biologie/SBC:sg vers/PREP des/DTC:pl lieux/SBC:pl extraterrestres/ADJ:pl ./.

Il/PRV:sg s'/PRV:sg intéressait/VCJ:sg depuis/PREP plusieurs/DTN:pl années/SBC:pl à/PREP la/DTN:sg question/SBC:sg de/PREP la/DTN:sg possibilité/SBC:sg de/PREP vie/SBC:sg ,/, ailleurs/ADV dans/PREP l'/DTN:sg univers/SBC:sg ./.

Il/PRV:sg prévoyait/VCJ:sg qu'/SUB avec/PREP le/DTN:sg développement/SBC:sg de/PREP l'/DTN:sg exploration/SBC:sg

spatiale/ADJ:sg -/- alors/ADV déjà/ADV en/PREP plein/ADJ:sg essor/SBC:sg -/- des/DTC:pl moyens/SBC:pl de/PREP recherches/SBC:pl puissants/ADJ:pl ,/, et/COO qui/REL n'/ADV avaient/ACJ:pl pas/ADV pu/PAR:sg être/ENCFF mis/PAR:sg en/PREP œuvre/SBC:sg auparavant/ADV ,/, allaient/VCJ:pl être/ENCFF bientôt/ADV disponibles/ADJ:pl pour/PREP étudier/VNCFF une/DTN:sg éventuelle/ADJ:sg vie/SBC:sg extraterrestre/ADJ:sg ,/, et/COO ce/PRO:sg ,/, de/PREP façon/SBC:sg rationnelle/ADJ:sg ./.

La/DTN:sg question/SBC:sg de/PREP la/DTN:sg vie/SBC:sg extraterrestre/ADJ:sg passait/VCJ:sg ainsi/ADV officiellement/ADV du/DTC:sg domaine/SBC:sg des/DTC:pl écrivains/SBC:pl et/COO de/PREP la/DTN:sg science/SBC:sg fiction/SBC:sg à/PREP celui/PRO:sg des/DTC:pl scientifiques/SBC:pl autres/PRO:pl que/SUB quelques/DTN:pl astronomes/SBC:pl isolés/ADJ:pl ,/, avec/PREP comme/SUB principal/ADJ:sg sponsor/SBC:sg ,/, les/DTN:pl agences/SBC:pl spatiales/ADJ:pl ./.

En/PREP effet/SBC:sg ,/, les/DTN:pl agences/SBC:pl nationales/ADJ:pl américaine/ADJ:sg (/(NASA/SBC:sg)/) ,/, française/ADJ:sg (/(CNES/SBC:sg)/) ,/, allemande/ADJ:sg (/(DASA/SBC:sg)/) ,/, européenne/ADJ:sg (/(ESA/SBC:sg)/) ,/, etc/ADV ./. ,/, ont/ACJ:pl à/PREP présent/SBC:sg inclus/PAR:pl dans/PREP leur/DTN:sg politique/SBC:sg scientifique/ADJ:sg le/DTN:sg soutien/SBC:sg à/PREP des/DTC:pl programmes/SBC:pl "/" exobiologiques/ADJ:pl "/" ./.

Jusqu'/PREP aux/DTC:pl années/SBC:pl 1990/CAR ,/, ceux/PRO:pl -/- ci/ADV ont/ACJ:pl été/EPAR:sg généralement/ADV proposés/PAR:pl par/PREP des/DTC:pl scientifiques/SBC:pl plus/ADV proches/ADJ:pl des/DTC:pl sciences/SBC:pl de/PREP la/DTN:sg vie/SBC:sg que/SUB de/PREP celles/PRO:pl de/PREP l'/DTN:sg univers/SBC:sg ./.

Or/COO il/PRV:sg existe/VCJ:sg à/PREP présent/SBC:sg une/DTN:sg communauté/SBC:sg d'/PREP astronomes/SBC:pl ,/, astrophysiciens/SBC:pl et/COO planétologues/SBC:pl qui/REL s'/PRV:pl intéressent/VCJ:pl de/PREP près/ADV au/DTC:sg problème/SBC:sg de/PREP la/DTN:sg vie/SBC:sg extraterrestre/ADJ:sg ./.

Certains/PRO:pl ont/ACJ:pl trouvé/PAR:sg que/SUB le/DTN:sg terme/SBC:sg "/" exobiologie/SBC:sg "/" est/ECJ:sg peu/ADV explicite/ADJ:sg et/COO sans/PREP doute/SBC:sg trop/ADV lié/PAR:sg aux/DTC:pl sciences/SBC:pl de/PREP la/DTN:sg vie/SBC:sg (/(terrestre/ADJ:sg)/) ,/, alors/ADV que/SUB le/DTN:sg domaine/SBC:sg de/PREP recherche/SBC:sg est/ECJ:sg extrêmement/ADV vaste/ADJ:sg et/COO très/ADV pluridisciplinaire/ADJ:sg ./.

Jugeant/VNCNT en/PREP particulier/SBC:sg qu'/SUB il/PRV:sg ne/ADV faisait/VCJ:sg pas/ADV apparaître/VNCFF l'/DTN:sg implication/SBC:sg des/DTC:pl sciences/SBC:pl de/PREP l'/DTN:sg univers/SBC:sg ,/, ils/PRV:pl ont/ACJ:pl décidé/PAR:sg d'/PREP introduire/VNCFF une/DTN:sg nouvelle/ADJ:sg appellation/SBC:sg ,/, mélange/SBC:sg de/PREP sciences/SBC:pl de/PREP la/DTN:sg vie/SBC:sg et/COO d'/PREP astronomie/SBC:sg :/: la/DTN:sg bioastronomie/SBC:sg ,/, qui/REL est/ECJ:sg en/PREP plein/ADJ:sg développement/SBC:sg actuellement/ADV ./.

Il/PRV:sg est/ECJ:sg nécessaire/ADJ:sg de/PREP rappeler/VNCFF une/DTN:sg approche/SBC:sg indispensable/ADJ:sg ,/, peut/ADV -/- être/ADV plus/ADV indirecte/ADJ:sg ,/, mais/COO soutenue/PAR:sg par/PREP une/DTN:sg incontestable/ADJ:sg logique/SBC:sg :/: l'/DTN:sg étude/SBC:sg des/DTC:pl origines/SBC:pl de/PREP la/DTN:sg vie/SBC:sg sur/PREP Terre/SBC:sg ./.

Car/COO rechercher/VNCFF ailleurs/ADV une/DTN:sg vie/SBC:sg signifie/VCJ:sg que/SUB nous/PRV:pl pensons/VCJ:pl que/SUB l'/DTN:sg apparition/SBC:sg de/PREP la/DTN:sg vie/SBC:sg n'/ADV est/ECJ:sg pas/ADV un/DTN:sg phénomène/SBC:sg unique/ADJ:sg ./.

Étudier/VNCFF l'/DTN:sg origine/SBC:sg de/PREP la/DTN:sg seule/ADJ:sg vie/SBC:sg que/SUB nous/PRV:pl connaissions/VCJ:pl actuellement/ADV ,/, comprendre/VNCFF les/DTN:pl processus/SBC:pl et/COO les/DTN:pl mécanismes/SBC:pl qui/REL ont/ACJ:pl permis/PAR:sg aux/DTC:pl premiers/ADJ:pl systèmes/SBC:pl vivants/ADJ:pl d'/PREP apparaître/VNCFF sur/PREP notre/DTN:sg planète/SBC:sg ,/, devrait/VCJ:sg nous/PRV:pl permettre/VNCFF de/PREP dresser/VNCFF une/DTN:sg liste/SBC:sg des/DTC:pl conditions/SBC:pl requises/ADJ:pl et/COO de/PREP vérifier/VNCFF ,/, éventuellement/ADV ,/, qu'/SUB il/PRV:sg n'/ADV est/ECJ:sg pas/ADV absurde/ADJ:sg de/PREP supposer/VNCFF que/SUB des/DTC:pl conditions/SBC:pl analogues/ADJ:pl ont/ACJ:pl pu/PAR:sg se/PRV:++ trouver/VNCFF ailleurs/ADV dans/PREP l'/DTN:sg univers/SBC:sg ./.

III/CAR ./. -/-

Qu'/PRO:sg est/ECJ:sg -/- ce/PRV:sg que/SUB la/DTN:sg vie/SBC:sg ?/?

Avant/PREP de/PREP partir/VNCFF à/PREP la/DTN:sg découverte/SBC:sg de/PREP la/DTN:sg vie/SBC:sg en/PREP dehors/SBC:sg de/PREP la/DTN:sg Terre/SBC:sg ,/, et/COO de/PREP considérer/VNCFF son/DTN:sg étude/SBC:sg éventuelle/ADJ:sg ,/, il/PRV:sg faut/VCJ:sg s'/PRV:++ entendre/VNCFF sur/PREP ce/PRO:sg qu'/SUB on/PRV:sg cherche/VCJ:sg ./.

On/PRV:sg ne/ADV peut/VCJ:sg par/PREP conséquent/SBC:sg éviter/VNCFF de/PREP se/PRV:++ demander/VNCFF :/: "/" Qu'/PRO:sg est/ECJ:sg -/- ce/PRV:sg que/SUB la/DTN:sg vie/SBC:sg ?/? "/"

Il/PRV:sg faut/VCJ:sg se/PRV:++ poser/VNCFF cette/DTN:sg question/SBC:sg pour/PREP deux/CAR raisons/SBC:pl au/DTC:sg moins/ADV ./.

1/CAR /// elle/PRV:sg a/ACJ:sg des/DTC:pl enjeux/SBC:pl philosophiques/ADJ:pl évidents/ADJ:pl sur/PREP lesquels/REL nous/PRV:pl reviendrons/VCJ:pl ;/; 2/CAR /// si/SUB l'/PUL on/PRV:sg n'/ADV y/PRV:++ prend/VCJ:sg pas/ADV garde/SBC:sg et/COO que/SUB l'/PUL on/PRV:sg adopte/VCJ:sg une/DTN:sg définition/SBC:sg trop/ADV restrictive/ADJ:sg de/PREP la/DTN:sg vie/SBC:sg ,/, on/PRV:sg risque/VCJ:sg de/PREP passer/VNCFF à/PREP côté/SBC:sg de/PREP formes/SBC:pl de/PREP vies/SBC:pl qui/REL sont/ECJ:pl pourtant/ADV à/PREP notre/DTN:sg portée/SBC:sg ./.

Cette/DTN:sg question/SBC:sg de/PREP la/DTN:sg définition/SBC:sg de/PREP la/DTN:sg vie/SBC:sg est/ECJ:sg en/PREP fait/SBC:sg très/ADV délicate/ADJ:sg et/COO demande/VCJ:sg une/DTN:sg réflexion/SBC:sg épistémologique/ADJ:sg ./.

Aussi/ADV ,/, c'/PRV:sg est/ECJ:sg par/PREP l'/DTN:sg étude/SBC:sg de/PREP la/DTN:sg vie/SBC:sg sur/PREP Terre/SBC:sg que/SUB nous/PRV:pl commencerons/VCJ:pl cet/DTN:sg ouvrage/SBC:sg ./.

L'/DTN:sg exemple/SBC:sg terrestre/ADJ:sg sera/ECJ:sg un/DTN:sg modèle/SBC:sg pour/PREP tenter/VNCFF de/PREP définir/VNCFF la/DTN:sg vie/SBC:sg ,/, mais/COO aussi/ADV et/COO surtout/ADV pour/PREP apprendre/VNCFF et/COO pour/PREP comprendre/VNCFF

la/DTN:sg vie/SBC:sg ,/, ses/DTN:pl processus/SBC:pl et/COO ses/DTN:pl structures/SBC:pl (/(chap/ABR ./.

I/CAR)/) ./.

Connaissant/VNCNT mieux/ADV le/DTN:sg vivant/SBC:sg terrestre/ADJ:sg ,/, nous/PRV:pl pourrons/VCJ:pl alors/ADV décrire/VNCFF l'/DTN:sg évolution/SBC:sg des/DTC:pl théories/SBC:pl sur/PREP l'/DTN:sg origine/SBC:sg de/PREP la/DTN:sg vie/SBC:sg sur/PREP Terre/SBC:sg ,/, le/DTN:sg scénario/SBC:sg actuellement/ADV retenu/PAR:sg (/(chap/ABR ./.

II/CAR)/) ,/, mais/COO aussi/ADV l'/DTN:sg évolution/SBC:sg de/PREP la/DTN:sg vie/SBC:sg et/COO ses/DTN:pl limites/SBC:pl actuelles/ADJ:pl (/(chap/ABR ./.

III/CAR)/) ./.

L'/DTN:sg exemple/SBC:sg terrestre/ADJ:sg servira/VCJ:sg ensuite/ADV de/PREP premier/ADJ:sg guide/SBC:sg pour/PREP chercher/VNCFF ailleurs/ADV ,/, en/PREP se/PRV:++ limitant/VNCNT dans/PREP un/DTN:sg premier/ADJ:sg temps/SBC:sg au/DTC:sg cas/SBC:sg d'/PREP une/DTN:sg vie/SBC:sg primitive/ADJ:sg :/: dans/PREP le/DTN:sg système/SBC:sg solaire/ADJ:sg (/(chap/ABR ./.

IV/CAR)/) et/COO hors/PREP du/DTC:sg système/SBC:sg solaire/ADJ:sg (/(chap/ABR ./.

V/CAR)/) ./.

Sommes/ECJ:pl -/- nous/PRV:pl trop/ADV peu/ADV inventifs/ADJ:pl dans/PREP nos/DTN:pl recherches/SBC:pl ?/?