UNIVERSITÉ PARIS 13, SORBONNE PARIS CITÉ
U.F.R. LETTRES, SCIENCES DE L'HOMME ET DES SOCIÉTÉS
LABORATOIRE LEXIQUES, DICTIONNAIRES, INFORMATIQUE

THÈSE DE DOCTORAT

En vue de l'obtention du grade de

DOCTEUR DE L'UNIVERISTÉ PARIS 13

EN SCIENCES DU LANGAGE

TITRE

ÉTUDE DE LA FONCTION ARGUMENTALE DANS LA PERSPECTIVE DE L'ACQUISITION AUTOMATIQUE DU VOCABULAIRE

Présentée et soutenue publiquement

le 10 juin 2016

Par
Xiaoqin HU

DIRECTEUR DE THÈSE

Pierre-André BUVET

JURY

M. Xavier BLANCO, Université Autonome de Barcelone

M. Juan-Manuel TORRES-MORENO, Université d'Avignon

M. Pierre-Patrick HAILLET, Université de Cergy-Pontoise

M. Pierre-André BUVET, Université Paris 13

À mes parents

« J'ai décidé d'être heureux parce que

c'est bon pour la santé. »

Voltaire

Remerciements

Je tiens tout d'abord à remercier mon directeur de thèse, M. Pierre-André Buvet à l'Université Paris 13, pour avoir accepté de diriger mes travaux de recherche. Je le remercie de m'avoir guidé et fourni des ressources linguistiques importantes pour réaliser cette thèse.

J'adresse également mes remerciements à M. André Dugas, professeur en linguistique et mathématiques appliquées à l'Univeristé du Québec et M. Gaston Gross, professeur émérite en linguistique à l'Université Paris 13. Ils m'ont donné beaucoup de conseils très précieux en linguistique, en traitement automatique des langues et concernant la rédaction de la thèse. Leur aide et leur soutien m'ont beaucoup encouragée pour achever ma thèse.

Je tiens à remercier mes camarades de laboratoire avec qui j'ai partagé beaucoup de moments chaleureux et sympathiques. Les discussions enrichissantes que nous avons partagées m'ont beaucoup inspirée sur ma recherche en sciences du langage.

Je remercie également tous les memebres du laboratoire pour leur soutien et leur sympathie.

Finalement, je tiens à exprimer ma gratitude à ma famille, en particulier mes parents qui m'ont soutenue, encouragée et financée depuis le tout début de mes études en France.

Table des matières

Liste des figures

Liste des tableaux

Notations

N	Nom
Nc	Nom d'autres classes sémantiques
PRON	Pronom
Prép	Préposition
DET	Déterminant
ADJ	Adjectif
V	Verbe
ADV	Adverbe
GN	Groupe nominal
NAF	Nom d'artefact
NMP	Nom de métier
Préd	Prédéterminant
Dnom	Déterminant nominal
Arg	Argument
Préd	Prédicat
PrédN	Prédicat nominal
Vpp	Verbe au participe passé
Vinf	Verbe à l'infinitif
Vpr	Verbe au présent
GNAF	Groupe nominal de nom d'artefact
SN	syntagme nominal
SP	syntagme prépositionnel
NMOD	N+Modifieur
C.C.	Complément circonstanciel
C.O.D.	Complément d'objet direct
C.O.I.	Complément d'objet indirect
C.M.	Complément de moyen

Introduction

L'objectif de cette thèse est d'élaborer une méthode permettant d'acquérir automatiquement les termes de façon efficace et pertinente en fonction des caractéristiques linguistiques de la fonction argumentale définies dans le modèle de données « Trois fonctions primaires ». Nous avons effectué une étude sur la fonction argumentale et développé trois méthodes d'acquisition automatique des termes : la méthode distributionnelle, la méthode morphosémantique et la méthode combinatoire qui associe les deux premières approches. Deux vocabulaires spécifiques sont étudiés pour la recherche de cette thèse : les noms d'artefacts et les noms de métiers. Les noms de métiers se définissent d'une façon plus large comme des listes d'activités humaines ne faisant pas de distinction stricte entre métier, profession, artiste, fonctionnaire… Les noms d'artefacts sont également une liste hétéroclite d'outils, d'ustensiles, de mécanismes, de véhicules… La méthode distributionnelle est expérimentée à partir du vocabulaire des noms d'artefacts et la méthode morphosémantique l'est à partir du vocabulaire des noms de métiers. La méthode combinatoire s'applique aux deux vocabulaires. Finalement, nous avons comparé ces trois méthodes et effectué une série d'analyses sur les caractéristiques linguistiques de la fonction argumentale dans la perspective du traitement automatique des langues.

Dans cette thèse, on expose premièrement la problématique, les hypothèses et la méthodologie du travail. On explicite d'abord l'objectif et la problématique générale du travail. Ensuite, on pose les questionnements afférents : l'approche distributionnelle est-elle pertinente pour l'acquisition automatique du vocabulaire ? L'approche morphosémantique permet-elle d'étiqueter du vocabulaire ? Une approche mixte permet-elle d'améliorer les résultats par rapport aux deux autres approches ? Et puis, on présente les hypothèses correspondantes à chaque questionnement posé. Finalement, on expose le modèle de données, la technologie et la ressource utilisée pour réaliser les trois méthodes développées.

Deuxièmement, on focalise la présentation sur les trois méthodes d'acquisition automatique des termes. Pour la méthode distributionnelle, la présentation commence par une évaluation du profilage du corpus, y compris la problématique concernant la constitution du corpus, le corpus sélectionné pour la méthode distributionnelle et le moyen de constitution du corpus pour la méthode distributionnelle. On explique ensuite chaque étape dans la méthode distributionnelle (supervisée et semi-supervisée), allant du prétraitement (levée des ambiguïtés morphosyntaxiques et le traitement des séquences polylexicales), comprenant l'extraction automatique des structures prédicat-argument, à l'intersection et jusqu'au calcul statistique des patrons syntaxiques. Enfin, on présente en détail la méthode d'évaluation et les résultats obtenus. Pour la méthode morphosémantique, on présente d'abord le corpus choisi et le moyen de constitution du corpus dans le cadre de la méthode morphosémantique. Ensuite, on expose la méthode d'exploitation des structures internes des unités lexicales pour l'acquisition automatique des termes. Finalement, on présente la méthode d'évaluation et les résultats obtenus. Pour la méthode combinatoire, on présente également le corpus constitué d'abord. Et puis, on explique respectivement le moyen de combiner les deux premières approches pour l'acquisition automatique des deux vocabulaires : celui des noms d'artefacts et celui des noms de métiers. Finalement, on expose la méthode d'évaluation et les résultats obtenus.

Dans la dernière partie de la thèse, on présente une analyse des résultats obtenus par les trois méthodes et on compare ces méthodes pour mieux comprendre leurs mécanismes, leurs avantages et leurs inconvénients. Ensuite, nous présentons également une série d'analyses de la représentation des structures syntaxiques et de celle des structures lexicales, des classes sémantiques et des langues de spécialité dans la perspective du traitement automatique des langues. On essaie d'expliquer la raison d'adopter telles stratégies ou telles méthodes pour constituer le corpus et pour exploiter les caractéristiques linguistiques de la fonction argumentale. Finalement, on présente des perspectives de notre recherche.

Première partie : Préalables méthodologiques

Chapitre 1 Problématique et questionnements afférents

1. Problématique générale

L'objectif de cette thèse est d'étudier la fonction argumentale et d'essayer de développer une méthode permettant d'acquérir automatiquement les termes de façon efficace et pertinente. Dans le modèle de données « Trois fonctions primaires », les unités lexicales sont catégorisées selon les parties du discours et sont recensées selon leurs syntaxes et leurs particularités structurales (prédicats, arguments, actualisateurs). La description linguistique basée sur « Trois fonctions primaires »se focalise sur la description des emplois des unités lexicales d'une langue. Dans la morphologie, les unités lexicales sont catégorisées en fonction de leurs formations (simples, dérivées, composées). La description morphologique se base sur les analyses des structures internes des unités lexicales. On se demande : est-ce que le modèle de données « Trois fonctions primaires » et la méthodologie morphologique sont exploitables pour l'acquisition automatique du vocabulaire ? Si oui, comment doit-on les exploiter et par quelles méthodes informatiques ?

2. Questionnements

2.1. L'approche distributionnelle est-elle pertinente pour l'acquisition automatique du vocabulaire ?

Une approche distributionnelle a pour objet d'utiliser un critère distributionnel pour calculer des liens de proximité sémantique entre les mots et visualiser les relations sémantiques. Si l'approche distributionnelle est pertinente pour l'acquisition automatique du vocabulaire, comment doit-on exploiter les structures prédicat-argument, quel corpus

est adéquat pour expérimenter l'approche distributionnelle et comment évaluer la pertinence des résultats à la fin ? Dans ce qui suit, on fera une discussion détaillée sur ces trois questions.

2.1.1. Quel corpus pour la méthode distributionnelle ?

Un corpus n'est pas un ensemble de données langagières en vrac mais des données qu'on décide de regrouper pour une étude particulière (Habert et al., 1997). La représentativité du corpus est liée directement à la génération des résultats de l'analyse ou du projet dans le traitement automatique des langues. Un corpus adéquat doit comprendre non seulement les informations sur lesquelles on doit travailler mais aussi le moyen pour exploiter ces informations. La sélection du corpus dépend de la problématique scientifique du projet. Ainsi, quel corpus est adéquat pour la méthode distributionnelle devient une question pour savoir quel corpus contient les informations et le moyen d'exploitation de ces informations pour réaliser la méthode distributionnelle. Le corpus pour la méthode distributionnelle doit correspondre au modèle de données utilisé dans la méthode distributionnelle. Tous les corpus ne s'adaptent pas à la méthode distributionnelle. Le profilage du corpus détermine directement la performance du système et la pertinence des résultats générés. Le choix d'un corpus à l'épreuve d'une méthode distributionnelle est une problématique importante et incontournable.

2.1.2. Comment exploiter les structures prédicat-argument ?

Les structures prédicat-argument permettent de visualiser des relations sémantiques. L'approche distributionnelle utilisée pour acquérir automatiquement le vocabulaire s'appuie sur l'exploitation des structures prédicat-argument. La méthode d'exploitation des structures prédicat-argument détermine directement la performance de l'approche distributionnelle. Cependant, comment peut-on identifier automatiquement les structures prédicat-argument à partir du corpus ? Et quelle stratégie doit-on adopter pour repérer le vocabulaire à travers les structures prédicat-argument ? Peut-elle nous aider à cerner la prédictibilité sémantique des prédicats ? De plus, les emplois prédicatifs polysémiques partagent une même forme mais ils ne sont pas les mêmes prédicats en raison de leur natures syntactico-sémantiques différentes.

La polysémie des prédicats révèle potentiellement des classes sémantiques différentes de celles du vocabulaire qu'on veut récupérer. Est-ce que la représentation des structures prédicat-argument permet de considérer la polysémie ? Si oui, comment doit-on exploiter les structures prédicat-argument pour prendre en compte la polysémie ? Comment exploiter les structures prédicat-argument est une question cruciale au cœur du développement de l'approche distributionnelle dans cette thèse.

2.1.3. Comment évaluer la pertinence des résultats obtenus ?

Les différentes technologies et les différents consommateurs des résultats d'évaluation (fondateur, développeurs et utilisateurs) déterminent les méthodes d'évaluation. Le type d'évaluation dépend aussi des entrées et des sorties du système. Une méthode d'évaluation adéquate permet d'évaluer précisément la performance du système. Pour les résultats obtenus par la méthode distributionnelle, quelle approche d'évaluation doit-on adopter ? Quelles sont les caractéristiques des entrées et des sorties de la méthode distributionnelle ? Par quels aspects la pertinence des résultats doit-t-elle être évaluée ? En plus d'une évaluation de qualité, est-ce qu'une évaluation quantitative est également nécessaire pour les sorties du système ? Toutes ces questions se posent à propos du choix de la méthode d'évaluation.

2.2. L'approche morphosémantique permet-elle d'étiqueter le vocabulaire ?

L'approche morphosémantique s'appuie sur les analyses de la morphologie et de la sémantique des unités lexicales. Si le vocabulaire de l'étude possède une série de caractéristiques morphologiques, est-ce que l'approche morphosémantique permet de l'étiqueter ? Si oui, comment doit-on exploiter la méthodologie morphologique et à quelles méthodes informatiques peut-on recourir pour réaliser l'approche morphosémantique ? Est-ce que tous les corpus sont adéquats pour l'approche morphosémantique ? Comment peut-on évaluer la pertinence du résultat ? Dans ce qui suit, on présente en détail cette série de sous-questions.

2.2.1. Quel corpus est adéquat pour l'approche morphosémantique ?

L'analyse morphosémantique a pour objet d'analyser la structure interne et la relation sémantique interne des unités lexicales. L'analyse morphosémantique se base sur les analyses morphologiques. Une série de questions sont posées : quel corpus doit-on choisir pour réaliser l'approche morphosémantique ? Quel corpus comprend assez d'informations morphologiques ? Est-ce qu'un corpus pauvre en contextes peut être utilisé pour réaliser l'approche morphosémantique où la richesse en contexte est nécessaire pour optimiser la performance de la méthode morphosémantique ? Par rapport à la méthode distributionnelle, est-ce que les informations et le moyen pour exploiter les informations sont différents ?

2.2.2. Comment exploiter l'analyse de la structure interne des unités lexicales dans le cadre de l'approche morphosémantique ?

Les unités monolexicales et les unités polylexicales sont distinguées en linguistique informatique selon qu'elles comportent ou non des séparateurs (espace, trait d'union, apostrophe). L'analyse de la structure interne des unités monolexicales s'effectue traditionnellement en se basant sur les morphèmes et celle des unités polylexicales est effectuée en se basant sur le lexique. Dans le cadre de l'approche morphosémantique, comment doit-on exploiter l'analyse de la structure interne des unités lexicales pour étiqueter le vocabulaire ? Les structures internes des unités monolexicales d'un vocabulaire possèdent-elles certaines propriétés permettant de les identifier ? Est-ce qu'une représentation des structures internes des unités polylexicales par les patrons morphosyntaxiques nous aide à réaliser l'approche morphosémantique ? Si oui, comment peut-on construire cette représentation morphosémantique ? Doit-on établir une liste de patrons morphosyntaxiques d'une manière la plus exhaustive possible ou s'appuyer sur des calculs statistiques pour rentabiliser automatiquement les patrons morphosyntaxiques à partir du corpus ? Toutes ces questions sont liées à l'exploitation de l'analyse morphosémantique. Elles exigent de bien réfléchir pour valoriser la méthode morphosémantique.

2.2.3. Comment évaluer la pertinence du résultat ?

Malgré que les sorties de l'approche morphosémantique soient aussi une liste de termes comme celles de l'approche distributionnelle, les entrées et la technologie dans le cadre de la méthode morphosémantique diffèrent toutes de celles dans l'approche distributionnelle. Ainsi, est-ce qu'on a besoin de faire appel à une approche différente pour évaluer la pertinence du résultat obtenu par la méthode morphosémantique ? Si oui, quelle méthode est adéquate et par quels aspects doit-on évaluer les résultats afin de visualiser la performance de la méthode morphosémantique ? Est-ce qu'il faut faire une évaluation qualitative en plus de l'évaluation quantitative en fonction des caractéristiques des résultats ? Si oui, quelles approches que doit-on utiliser respectivement ?

2.3. Une approche mixte permet-elle d'améliorer les résultats obtenus avec les deux autres approches ?

La méthode distributionnelle et la méthode morphosémantique sont deux méthodes ayant les fonctionnements bien différents. Elles aboutissent aux différentes performances, comportent différents avantages mais aussi différents inconvénients. Ainsi, on se demande si une approche mixte qui combine ces deux méthodes permet une amélioration des résultats ? La combinaison de ces deux méthodes permet–elle de compléter l'une par l'autre ? Si la réponse est positive, comment doit-on combiner l'approche distributionnelle et l'approche morphosémantique ? Quel corpus est adéquat pour la méthode combinatoire et comment peut-on évaluer les résultats ? Dans ce qui suit, on présente en détail cet ensemble de questions.

2.3.1. Comment combiner l'approche distributionnelle et l'approche morphosémantique ?

Le moyen pour combiner l'approche distributionnelle et l'approche morphosémantique doit être décidé en fonction de la compatibilité des deux approches. L'objectif de la combinaison des deux méthodes est de mettre en valeur les avantages des

deux approches et compléter leurs désavantages dans le but d'améliorer au maximum le résultat. La question « Comment combiner l'approche distributionnelle et l'approche morphosémantique ? » peut se décliner par les questions suivantes : quel rôle joue chaque méthode dans l'approche mixte ? l'une est complémentaire ou filtrante pour l'autre ? à partir d'où doit-on intégrer une autre approche pour développer une approche mixte et de quelle manière ? est-ce qu'il nous faut ajouter d'autres processus pour valoriser la combinaison des méthodes ? auxquelles techniques informatiques ou linguistiques doit-on recourir afin de réaliser cette approche combinatoire ?

2.3.2. Quel corpus exploiter ?

L'approche distributionnelle est fondée sur l'exploitation des structures prédicat-argument, alors que l'approche morphosémantique s'appuie sur l'analyse morphologique des unités lexicales. Les corpus choisis pour ces deux méthodes correspondent aux caractères différents en raison des fonctionnements différents de ces deux méthodes. Cependant, quel est le corpus adéquat pour l'approche mixte qui combine ces deux approches de fonctionnement différent ? Est-ce qu'un corpus choisi pour l'approche distributionnelle ou un corpus adéquat pour l'approche morphosémantique peut aussi être exploité pour l'approche mixte ? Ou est-ce qu'il faut avoir un corpus de nouvelle nature en tenant compte des différents fonctionnements des deux approches ?

2.3.3. Comment évaluer la pertinence du résultat ?

Pour évaluer les résultats obtenus par l'approche mixte, est-ce qu'on a besoin d'une méthode d'évaluation différente de celles qu'on utilise pour l'approche distributionnelle ou pour l'approche morphosémantique ? Ou est-ce qu'il nous faut une nouvelle méthode d'évaluation pour prendre en compte à la fois l'influence de l'approche distributionnelle et celle de l'approche morphosémantique dans la méthode combinatoire ? Quels aspects doit-on considérer pour évaluer correctement la performance de la méthode mixte ? Est-t-il nécessaire d'évaluer les sorties intermédiaires pour mieux comprendre les résultats finaux ?

3. Hypothèses

Les hypothèses se concrétisent par les trois méthodes développées ci-dessous. On suppose que la méthode distributionnelle en exploitant les structures prédicat-argument est pertinente pour l'acquisition automatique du vocabulaire et que l'approche morphosémantique en s'appuyant sur les analyses de la morphologie et de de la relation sémantique interne des unités lexicales permet d'étiqueter du vocabulaire. La combinaison de l'approche distributionnelle et l'approche morphosémantique est effectuée à partir de l'hypothèse que cette méthode combinatoire permet d'améliorer les résultats obtenus par les deux autres approches.

3.1. Méthode distributionnelle

La méthode distributionnelle repose sur les relations sémantiques entre les mots. La possibilité d'exploitation des structures prédicat-argument permet de visualiser les relations sémantiques dans le corpus. L'exploitation des structures prédicat-argument est fondée sur la relation d'appropriation entre les prédicats appropriés et leurs arguments. Les prédicats appropriés ont un nombre de classes sémantiques d'arguments relativement contraint. Ce caractère des prédicats appropriés nous permet de prédire la classe sémantique dont leurs arguments font partie. Les classes d'arguments sont définies à partir d'un ensemble de prédicats appropriés. Ainsi, si l'on a un ensemble de prédicats appropriés, il est possible qu'on puisse récupérer une classe d'arguments. Dans ce qui suit, on décrit en détail les hypothèses sur le corpus adopté, le moyen d'exploitation des structures prédicat-argument et la méthode d'évaluation dans le cadre de l'approche distributionnelle pour l'acquisition automatique du vocabulaire.

3.1.1. Corpus spécialisé et corpus riche de structures prédicat-argument

On fait l'hypothèse que le corpus spécialisé est un corpus adéquat pour la méthode distributionnelle, puisque l'objectif de la méthode est d'acquérir automatiquement du vocabulaire spécifique. Un corpus spécialisé est un corpus en langue de spécialité. La langue

de spécialité est la production linguistique dans le cadre des communications de domaines spécifique. Il doit couvrir de nombreux termes et expressions spécialisées. Il est également possible que les prédicats appropriés associés aux arguments qui appartiennent à une langue de spécialité présente une grande fréquence.

La méthode distributionnelle est une méthode réalisée en exploitant les structures prédicat-argument. Ainsi, on suppose que plus un corpus est riche de structures prédicat-argument, plus il est aussi nécessaire. La richesse de structures prédicat-argument garantit la réalisation de la méthode et la qualité des résultats. Il est aussi possible qu'un corpus structuré dont la structure de langue est moins complexe permet de mieux exploiter les structures prédicat-argument.

3.1.2. Méthode supervisée et méthode semi-supervisée

En ce qui concerne le moyen d'exploitation des structures prédicat-argument, on fait deux hypothèses : l'une est fondée sur une méthode supervisée qui permet d'identifier les classes d'arguments à partir d'un ensemble de prédicats appropriés ; l'autre, sur une méthode semi-supervisée qui acquiert les prédicats appropriés à partir d'un ensemble d'arguments et récupère plus d'arguments avec les prédicats obtenus.

La méthode supervisée consiste à projeter une liste de prédicats appropriés sur un corpus et à identifier les structures prédicat-argument à l'aide des distributions syntactico-sémantiques des prédicats appropriés afin d'acquérir automatiquement les arguments, par exemple, la distribution syntactico-sémantique du prédicat approprié des noms d'artefacts (NAF) *fabriquer* est représentée par V+NAF et on peut ainsi repérer les arguments de *fabriquer* à l'aide du patron syntaxique V+NAF. En raison de la polysémie des unités lexicales, les prédicats polysémiques partagent la même forme mais ont des classes sémantiques d'arguments différentes. Néanmoins, un ensemble de prédicats appropriés permet de définir une classe sémantique d'arguments. Cet ensemble de prédicats appropriés sont les prédicats appropriés définitionnels de la classe d'arguments. L'idée est que nous établissons une liste de prédicats appropriés définitionnels d'une classe d'arguments pour

repérer les arguments de la classe à l'aide des patrons syntaxiques qui représentent les distributions syntactico-sémantiques des prédicats. Ensuite, nous appliquons l'opération d'intersection aux candidats arguments obtenus pour identifier les arguments de la classe sémantique. Cette opération d'intersection a pour objet de détecter les arguments communs de la liste de prédicats appropriés et ces arguments appartiennent à la classe sémantique des prédicats appropriés. Cependant, nous ne disposons pas d'une liste exhaustive de prédicats appropriés et pour cela, nous devons recourir à une méthode qui permet de récupérer automatiquement les prédicats appropriés à partir du corpus afin d'acquérir automatiquement les arguments.

La méthode semi-supervisée permet d'identifier les prédicats en repérant les structures prédicat-argument à partir d'une classe d'arguments. Ensuite, avec les prédicats obtenus, nous revenons à la méthode supervisée pour récupérer les arguments. Les prédicats reconnus à partir d'un ensemble d'arguments sont également filtrés par l'opération d'intersection afin d'obtenir les prédicats appropriés de la classe d'arguments donnée. Il est possible que ces processus soient répétés pour récupérer plus d'arguments. Cependant, dans le cadre de la méthode semi-supervisée, on est incapable de prédire les distributions syntactico-sémantiques des prédicats associés avec une classe d'arguments donnée à l'avance. Ainsi, l'apprentissage automatique des distributions syntactico-sémantiques des prédicats appropriés devient nécessaire dans l'approche distributionnelle semi-supervisée. Nous essayons de prédire toutes les possibilités de distributions syntactico-sémantiques que les prédicats appropriés d'une classe sémantique peuvent avoir et faisons appel à un calcul probabiliste pour sélectionner le patron syntaxique le plus adéquat de chaque prédicat parmi les candidats patrons syntaxiques.

3.1.3. Évaluation quantitative et évaluation qualitative

En fonction des différents systèmes et de la nature différente des entrées et des sorties, on adopte les différentes technologies d'évaluation. Dans les applications d'analyses, l'évaluation peut être exécutée en créant un ensemble de sorties correctes (standard d'or) et engager une comparaison automatique entre les entrées du système et le standard d'or. Les sorties du système de production peuvent être évaluées par l'évaluation de sa qualité, de son

caractère informatif ainsi que de son influence sur l'efficacité et l'acceptabilité dans une tâche enchâssée. L'évaluation du système interactive dépend du jugement des experts, la réaction de l'utilisateur et les matrices de niveaux de tâche.

Les sorties de l'approche distributionnelle peuvent être évaluées au moyen de l'évaluation quantitative. Nous pouvons établir un ensemble de sorties correctes et comparer les sorties de l'approche distributionnelle avec le standard en faisant appel aux critères de mesure de performance : rappel et précision. Le standard peut être obtenu par l'annotation manuelle ou par des ressources linguistiques.

L'approche distributionnelle peut également être évaluée au plan de la qualité. En analysant les types d'erreurs ou les types de silences, le fonctionnement et les caractéristiques du système peuvent être plus clairement visualisés.

3.2. Méthode morphosémantique

On pose l'hypothèse que l'approche morphosémantique permet d'étiqueter le vocabulaire. Cela présuppose que les unités lexicales d'un vocabulaire spécifique partagent une série de propriétés morphologiques permettant de les identifier. Ainsi, l'étiquetage du vocabulaire s'effectue par l'identification des propriétés morphologiques des unités lexicales. Les propriétés morphologiques du vocabulaire sont dévoilées par une série d'analyses de la structure interne et de la relation sémantique interne des unités lexicales. Dans ce qui suit, on présente en détail les hypothèses sur le profilage du corpus pour l'approche morphosémantique, sur le moyen d'exploitation des analyses morphologiques et sur la méthode d'évaluation.

3.2.1. Corpus spécialisé et corpus pauvre en contextes

Le corpus choisi pour l'approche morphosémantique doit comprendre assez d'informations morphologiques. En fonction du principe de l'approche morphosémantique, on fait deux hypothèses sur le corpus adéquat : le corpus spécialisé est le corpus exploité par

l'approche morphosémantique ; un corpus pauvre en contextes s'adapte également à la méthode morphosémantique.

Puisque l'approche morphosémantique se base sur les analyses de la morphologie et la relation sémantique interne des unités lexicales, un corpus spécialisé qui fournit une grande quantité d'occurrences de termes doit être un bon choix. Un corpus riche de termes permet de fournir assez d'informations morphologiques. La richesse d'informations morphologiques est nécessaire dans le cadre de l'approche morphosémantique.

De plus, on suppose également qu'un corpus pauvre en contextes est quand même exploitable, car la méthode morphosémantique s'appuie sur les analyses des structures internes des mots et elle a besoin de peu d'informations contextuelles. Un corpus composé de listes de mots ne comprend pas de contextes mais est riche d'informations morphologiques. Ce type de corpus permet également l'exploitation des analyses de la structure interne des unités lexicales.

3.2.2. Analyse morphologique et analyse sémantique

Dans la méthode morphosémantique, on distingue les unités monolexicales et les unités polylexicales. Pour les unités monolexicales, nous étudions leur morphologie à partir des morphèmes et essayons de trouver les caractéristiques morphologiques permettant d'étiqueter ou d'enrichir le vocabulaire. La segmentation des unités monolexicales en morphèmes permet de nouvelles combinaisons des morphèmes et cela permet d'étiqueter de nouvelles unités lexicales (par ex., *psychiatre [psych(o)-, -iatre], géologue [géo-, -logue]* -> *psychologue [psych(o)-, -logue]*). La fonction des morphèmes affixaux est de faire entrer les unités dans une relation paradigmatique (par ex., les noms de métiers *technicien, physicien, gardien*, …se termient tous par le *-ien*) et cela facilite l'identification d'un vocabulaire spécifique. Pour les unités polylexicales, nous analysons leur construction à partir du lexique et représentons leurs structures internes par les patrons morphosyntaxiques, par exemple, la structure interne de l'unité polylexicale *batteur à œufs* peut être représentée par le patron

morphosyntaxique N1+à+N2. La projection de l'ensemble des patrons morphosyntaxiques sur le corpus permet de repérer les unités polylexicales.

La relation sémantique interne entre les constituants des unités lexicales est analysée pour visualiser le lien entre la sémantique et la formation des unités lexicales (par ex., dans *boîte à déjeuner*, *déjeuner* indique la fonction de *boîte*). La relation sémantique entre les unités lexicales et leurs constituants est aussi étudiée. Cela a pour objet de trouver les propriétés linguistiques permettant de distinguer les unités lexicales d'une classe sémantique de celles des autres classes sémantiques. Par exemple, la relation sémantique entre le nom de métier *directeur* (X) et son constituant *direct-* (*diriger*) (Y) peut être paraphrasée par « X est la personne qui Y » : « directeur est la personne qui dirige », alors que celle du nom d'artefact *accélérateur* ne le permet pas bien qu'elle ait le même suffixe *-eur*.

3.2.3. Évaluation quantitative et évaluation qualitative

Les sorties de l'approche morphosémantique peuvent également être évaluées en comparant avec le standard dans l'évaluation quantitative. Pareillement, nous établissons un ensemble de sorties correctes et comparons les sorties de l'approche morphosémantique avec le standard en calculant le taux de rappel et le taux de précision. Enfin, l'évaluation de la méthode morphosémantique du point de vue qualitative peut aussi être effectuée par l'analyse de la nature des erreurs et des silences.

3.3. Méthode combinatoire

On fait l'hypothèse que l'approche mixte qui combine la méthode distributionnelle et la méthode morphosémantique permet d'améliorer la pertinence des résultats obtenus avec les deux autres approches. On suppose qu'une approche est une partie complémentaire de l'autre ou qu'une approche est intégrée dans l'autre comme un module de filtrage. On fait également des hypothèses sur le corpus de l'approche mixte et les méthodes d'évaluation des résultats obtenus par l'approche mixte. Dans ce qui suit, on présente en détail cet ensemble d'hypothèses concernant la méthode combinatoire.

3.3.1. Mise en place d'un module complémentaire et mise en place d'un module de filtrage

L'approche distributionnelle a pour objet de récupérer le vocabulaire en localisant les relations sémantiques, alors que l'approche morphosémantique permet d'étiqueter le vocabulaire à partir de l'analyse de la structure interne et de la relation sémantique interne des unités lexicales. L'avantage de l'approche distributionnelle est qu'elle permet de repérer les distributions d'arguments afin d'identifier des classes sémantiques de termes. Pour l'approche morphosémantique, son avantage est qu'elle permet d'enrichir le vocabulaire (à la fois les unités monolexicales et les unités polylexicales) en exploitant les analyses morphosémantiques des unités lexicales. Pour combiner l'approche distributionnelle et l'approche morphosémantique, on fait également deux hypothèses. La première hypothèse est que l'approche morphosémantique peut être intégrée comme un module complémentaire dans l'approche distributionnelle pour étiqueter les unités polylexicales. La deuxième hypothèse est que l'approche distributionnelle est ajoutée comme un module de filtrage dans l'approche morphosémantique afin d'augmenter la pertinence d'étiquetage.

L'identification des structures prédicat-argument est réalisée à l'aide des patrons syntaxiques qui représentent les distributions syntactico-sémantiques des prédicats. On intègre les patrons morphosyntaxiques qui représentent les structures internes des unités polylexicales dans les patrons syntaxiques de la méthode distributionnelle à la position des distributions d'arguments afin de prendre en compte les noms composés. Certaines structures prédicat-argument peuvent également être exploitées pour former les candidats-termes composés (par ex., à partir des structures prédicat-argument *couper le jambon*, on peut obtenir *coupe-jambon*). De plus, à partir des termes obtenus, on peut continuer à enrichir la liste de candidats-termes composés à l'aide des opérations morphologiques, par exemple, à partir du nom d'artefact (NAF) *rouge à lèvre*, il est possible d'obtenir un autre nom composé *fixateur de rouge à lèvre* à l'aide du patron morphosyntaxique N+de+NAF (N <-de+NAF). L'approche morphosémantique est intégrée comme un module complémentaire dans la méthode distributionnelle.

Pour la mise en place du module de filtrage, la méthode distributionnelle consiste plutôt à délimiter une classe sémantique d'arguments pour l'approche morphosémantique. À partir de la classe sémantique d'arguments sélectionnée dans la méthode distributionnelle, on applique l'approche morphosémantique pour enrichir la liste de candidats-termes. Les structures prédicat-argument extraites peuvent également être exploitées pour fournir plus de termes de semence à partir desquels l'extension du vocabulaire est effectuée par les opérations morphologiques. Par exemple, dans la structure prédicat-argument *être spécialisé en* : *l'agence recrute un plieur numérique spécialisé en mécano-soudure*, on peut obtenir *mécano-soudeur* à partir de *mécano-soudure* en fonction de la relation sémantique interne entre le nom de métier (X) et sa base (Y) « X est la personne qui est spécialisée en Y (discipline/domaine) ». Dans ce cas-là, la méthode distributionnelle assure un rôle de filtrage sémantique pour la méthode morphosémantique.

3.3.2. Corpus spécialisé et corpus riche de structures prédicat-argument

La méthode combinatoire est une méthode qui combine l'approche distributionnelle et l'approche morphosémantique. Tant la mise en place du module complémentaire que la mise en place du module de filtrage, les deux approches gardent quand même leurs propres fonctionnements. Ainsi, le corpus utilisé pour la méthode combinatoire doit correspondre à la fois aux caractéristiques de la méthode distributionnelle et à celles de la méthode morphosémantique. On croit que l'approche mixte nécessite un corpus riche de structures prédicat-argument, puisque la réalisation de l'approche distributionnelle s'appuie sur l'exploitation de ce genre de structures. Ainsi, on fait l'hypothèse qu'un corpus spécialisé qui fournit une richesse de termes et de contextes dans un domaine spécifique peut être un corpus adéquat pour la méthode mixte. Cela permet de mettre en place à la fois l'approche distributionnelle et l'approche morphosémantique.

3.3.3. Évaluation quantitative et évaluation qualitative

Les sorties de la méthode combinatoire équivalent aux listes de termes. On peut aussi évaluer la pertinence du résultat du point de vue quantitatif et qualitatif. En ce qui concerne

l'évaluation quantitative, nous établissons un standard et comparons le résultat obtenu par la méthode mixte avec le standard en calculant le taux de précision et le taux de rappel. Le standard peut être établi par une annotation manuelle. De même, l'évaluation qualitative est aussi indispensable pour évaluer le fonctionnement de la méthode combinatoire. Une série d'analyses concernant les types d'erreurs et les types de silences peuvent être effectuée pour évaluer qualitativement la méthode.

Chapitre 2 État de l'art

L'acquisition automatique des termes est une problématique importante dans le traitement automatique des langues naturelles (TALN). Cependant, cela présente beaucoup de difficultés. Ces difficultés peuvent provenir des limites des techniques linguistiques, de celles des techniques informatiques ou de celles des limites de théories en TALN. La plupart des méthodes qu'on utilise maintenant pour l'acquisition automatique des termes sont fondées sur les descriptions morphosyntaxiques ou sur des calculs statistiques.

Dans les applications de TALN, les techniques d'apprentissage et celles de fouille de textes occupent une place très importante. L'apprentissage automatique consiste à établir des algorithmes permettant à un ordinateur d'évoluer grâce aux connaissances qu'il a acquises automatiquement. La fouille de textes (dite également exploration de données textuelles) a pour objet d'extraire des connaissances à partir de grandes quantités de données textuelles de façon automatique ou semi-automatique. La plupart des techniques de fouille de textes reposent sur les représentations basées sur les mots. À l'aide de ces techniques, on est capable d'entraîner le classifieur, de mesurer la similarité, de calculer l'information mutuelle, etc., pour réaliser finalement les tâches de traitement automatique des langues. Notamment, dans les méthodes statistiques ou hybrides qu'on utilise actuellement pour le traitement automatique des langues naturelles, les techniques d'apprentissage et les techniques de fouille de textes sont nécessaires et très importantes. Nous proposons une présentation de l'apprentissage automatique et de l'exploration de données textuelles avant celle qui portera sur l'acquisition automatique des termes.

Dans la première partie de l'état de l'art, on va introduire la notion d'apprentissage et présenter les principaux algorithmes en apprentissage automatique utilisés pour l'acquisition automatique des termes ou l'extraction de l'information. La présentation de fouille de textes sera orientée sur certaines techniques très populaires dans les applications de TALN et surtout

dans l'acquisition automatique des termes, telles que la similarité, l'information mutuelle, le Bayes naïf (Ibekwe-SanJuan, 2007 : 99), etc.

Par la suite, on présente l'état de l'art sur l'acquisition automatique des termes sous trois angles : les méthodes linguistiques pour l'acquisition automatique des termes ; les méthodes statistiques pour l'acquisition automatique des termes et les méthodes hybrides qui combinent à la fois les approches linguistiques et les approches statistiques pour l'acquisition automatique des termes. Pour chaque méthode, on présente certaines applications représentatives, telles que TERMINO (Lauriston, 1994), LEXTER (Bourigault, Gonzalez-Mullier et Gros, 1995), ANA (Enguehard, 1993), etc.

Enfin, on doit également constater que la reconnaissance des termes complexes est une problématique incontournable dans l'acquisition automatique des termes. L'efficacité de l'identification des noms composés contribue à augmenter la pertinence de l'acquisition automatique des termes. L'extraction automatique des termes composés joue un rôle crucial dans le traitement automatique des langues. La troisième partie est ainsi consacrée à la présentation de l'état de l'art sur l'extraction automatique des termes composés. On va présenter d'abord deux systèmes semi-automatiques permettant de faciliter la validation des recherches des lexicologues : ACABIT (Boulakandel, Daille et Aboutajdine, 2008) et le système de Biskri (2004). Ensuite, on présentera la méthode de Maurel (1993) qui se base sur les grammaires régulières, la méthode de Lebarbé (2002) qui se pose sur la cooccurrence, le modèle parser de Green (2012) et le multi-modèle de Matthieu Constant (2012).

1. Apprentissage automatique et fouille de textes

Les techniques d'apprentissage et les techniques de fouille de textes sont les techniques importantes en vue des différentes applications de traitement automatique. L'apprentissage automatique est une notion en intelligence artificielle dont l'objectif est de faire «apprendre» des connaissances à l'ordinateur automatiquement ou semi automatiquement. La fouille de textes a pour objet de découvrir des connaissances ou des informations utiles à partir de gros volumes de données textuelles de façon automatique ou semi-automatique. De nombreuses

méthodes de traitement automatique des langues se basent sur l'algorithme d'apprentissage à l'aide des techniques de fouille de données. Dans cette partie, on va faire une introduction sur ces deux notions et présenter certaines techniques d'apprentissage et de fouille de textes utilisées dans l'acquisition automatique ou dans l'extraction de l'information.

1.1. Techniques d'apprentissage automatique

En apprentissage automatique, les différents algorithmes d'apprentissage peuvent être divisés en trois types : l'apprentissage supervisé, l'apprentissage semi-supervisé et l'apprentissage non supervisé. L'apprentissage supervisé est une technique d'apprentissage automatique qui dépend complètement des données étiquetées. On produit les règles d'apprentissage entièrement à partir des données entraînées. L'apprentissage semi-supervisé est une technique d'apprentissage qui produit les règles automatiquement à partir d'un petit ensemble de données entraînées et on entraîne plus de données avec les règles produites. L'apprentissage semi-supervisé se définit par un processus itératif. En ce qui concerne l'apprentissage non supervisé, qui est complètement indépendant de données entraînées, on fouille les données en fonction des connaissances statistiques acquises à partir des textes.

L'apprentissage supervisé est une technique d'apprentissage automatique qui dépend complètement des bases de données entraînées pour produire les règles. Il s'agit d'un algorithme qui généralise les entrées inconnues en fonction des données annotées à l'avance par les experts. Par exemple, pour repérer tous les synonymes du mot anglais *nice* (*beau*) à partir d'un grand corpus inconnu, on peut faire appel à WordNet qui est une ressource lexicale structurée en anglais dans laquelle les verbes, les noms, les adverbes et les adjectifs sont groupés en séries de synonymes. Cet algorithme est supervisé, car il extrait les termes désirés à partir des données inconnues à l'aide d'une base de données entraînée (WordNet). L'inconvénient de l'algorithme d'apprentissage supervisé est qu'il dépend complètement des bases de données entraînées dont la construction est extrêmement coûteuse. Elle nécessite beaucoup de temps et une imposante intervention humaine.

L'apprentissage semi-supervisé est un algorithme qui exploite les données non annotées en utilisant des règles produites automatiquement à partir d'un petit ensemble de données annotées pour compléter l'apprentissage supervisé. Par exemple, si l'on a un petit ensemble d'instances de relation d'hyponymie comme *voiture->véhicule, pomme->fruit, réfrigérateur->électroménager,* etc. et qu'on cherche chaque paire d'entités dans un corpus non-annoté pour enregistrer leur environnement de cooccurrence, les patrons syntaxiques de la relation d'hyponymie peuvent être produits à partir des environnements de cooccurrence identifiés et ces patrons syntaxiques permettent de reconnaître plus de paires d'entités d'hyponymie. À partir des nouvelles paires d'entités reconnues, les nouveaux patrons syntaxiques de relation peuvent être découverts et plus de paires d'hyponymes peuvent être reconnus. Si ce processus est répété itérativement, plus de patrons syntaxiques et de paires d'hyponymes seront obtenus. Dans ce cas-là, il s'agit de l'algorithme Bootstrapping qui est en fait un processus itératif. Par rapport à l'apprentissage supervisé, l'apprentissage semi-supervisé est moins dépendant des données entraînées.

L'apprentissage non supervisé est une méthode d'apprentissage automatique qui ne nécessite ni des ressources ni du pré-entraînement. La plupart des méthodes non-supervisées pour le traitement automatique des langues naturelles sont réalisées à l'aide des techniques de fouille de textes, telles que la mesure d'association (similarité), le log-vraisemblance, l'information mutuelle, etc. Par exemple, si l'on veut classer les paires d'entités en groupes disjoints dans lesquels chaque groupe représente une seule relation sémantique, on peut avoir une corbeille de « clusters » qui capturent les paires d'entités non-corrélatives ou les relations peu importantes, et le contexte de chaque paire d'entités révèle le contexte de chaque paire d'entités ainsi que celui des deux entités qui coexistent. Bref, l'idée principale de cette méthode non-supervisée pour l'extraction automatique des entités est de classifier les entités ou les paries d'entités en fonction de leurs caractéristiques contextuelles lexico-syntaxiques. Ce genre de classifications non-supervisées pour l'extraction de l'information est souvent réalisée en s'appuyant sur le modèle vectoriel. LSA (Analyse sémantique latente) (Landauer et Dumais, 1997) est une variante de modèle vectoriel classique utilisée dans l'extraction de l'information. Dans le cadre de l'extraction de l'information, il utilise la matrice pour

représenter les termes (ou les relations) et leurs caractéristiques associées (ce sont souvent les caractéristiques contextuelles lexico-syntaxiques) afin de réaliser une classification non-supervisée (ou supervisée) des termes (ou des relations) en fonction de la similarité de leurs caractéristiques. La similarité est souvent calculée à l'aide des techniques de fouille de textes.

Le modèle de Markov est aussi un modèle très important en apprentissage automatique. Il est une méthode stochastique souvent utilisée dans la segmentation des langues idéographiques (le chinois, le japonais) ou dans l'extraction automatique de l'information (Jiang, 2012 : 18). Le modèle de Markov est un modèle génératif dans lequel on pose que l'état de l'unité précédente génère ou est généré par celui de l'unité qui suit. Jiang (2012 : 18-22) a présenté trois types de modèles de Markov : HMM (Hidden Markov Model) est un modèle génératif dans lequel la génération de l'étiquette ne dépend que de l'étiquette précédente ou de centaines d'étiquettes précédentes et l'étiquette du mot est traitée comme un état caché (hidden) ; MEMMs (Maximum Entropy Markov Models) est un modèle discriminatif dans lequel la génération de l'étiquette d'un mot courant est déterminée par à la fois le mot courant et l'étiquette ou certaines étiquettes précédentes, et ce modèle a un taux d'erreurs plus bas que le modèle HMM ; CRFs (Conditional random fields) est un autre modèle discriminatif dans lequel la génération d'une étiquette peut non seulement dépendre de l'étiquette (ou des étiquettes) précédentes mais aussi de l'étiquette suivante. La méthode basée sur le modèle de Markov peut être semi-supervisée ou non supervisée selon qu'il y l'intervention des données entraînées ou pas.

1.2. Techniques de fouille de textes

Ibekwe-SanJuan (2007 : 32-36) a comparé trois définitions de fouille de textes : d'après Feldman, Aumann et al. (1998), la fouille de textes est la science qui extrait des motifs cachés à partir de grandes collections de textes. Kodratoff (1999) a aussi donné une définition sur la fouille de textes : la fouille de textes est la science qui découvre des connaissances dans les données textuelles. Les connaissances ici ont une conception particulière. Elles se distinguent des connaissances connues ou des règles triviales. Elles doivent nécessairement appartenir à un domaine en rapport avec l'activité humaine. De plus,

la conception des connaissances recouvre certaines notions subjectives. Par exemple, pour les experts en fouille de textes (FT), les connaissances sont les motifs découverts à partir des données ; pour les philosophes, les connaissances le sont plutôt au sens philosophique. Hearst (1999), spécialiste plus proche de l'ingénierie linguistique, nous en a donné une définition plus restrictive : la fouille de textes correspond à la découverte par l'ordinateur de nouvelles informations extraites de plusieurs sources, qui seront reliées pour former de nouvelles réalités ou de nouvelles hypothèses à explorer par d'autres moyens. En tous cas, ces trois définitions ont toutes mis l'accent sur le caractère nouveau des motifs ou des informations découvertes et sur la grande quantité de corpus sur lesquels la FT doit porter.

Les disciplines qui contribuent au développement des techniques de la FT sont : la linguistique et l'ingénierie linguistique, la statistique, l'apprentissage automatique, l'informatique, et la visualisation de l'information. Notamment, pour le développement de la fouille de textes, on ne peut pas négliger le rôle prépondérant des techniques linguistiques et de l'ingénierie linguistique. À travers ces relations entre disciplines, on peut bien voir la nature multidisciplinaire de la fouille de textes.

L'application de la fouille de textes concerne plusieurs domaines : les services ou les analyses dans l'économie, la gestion de ressources humaines, la veille concurrentielle, l'extraction de l'information, la catégorisation de textes, le résumé des textes, etc. Parmi ces applications, l'extraction de l'information, la catégorisation de textes et le résumé des textes appartiennent plutôt aux tâches de fouille et aux applications intermédiaires et les autres sont plutôt des applications terminales. Dans la partie suivante, on va présenter certaines techniques de fouille de textes souvent utilisées dans les applications dédiées au traitement de l'information, telles que la mesure de similarité, la pondération, le calcul de l'information mutuelle, la vraisemblance, etc.

Une mesure de similarité entre unités textuelles permet de mesurer la proximité ou l'éloignement entre deux unités textuelles ou entre les documents. Une mesure de similarité est une mesure d'association. La mesure de similarité est souvent utilisée pour regrouper ou classifier les unités lexicales ou les documents. Il existe plusieurs types de mesures de

similarité. Ici, on présente deux mesures de similarité souvent utilisées pour illustrer le principe et l'application de cette technique de FT. Le premier type de mesure de similarité est la mesure de similarité pour l'agrégation des termes en analyse des données qui est souvent utilisée pour les méthodes de classification automatique. Il existe plusieurs façons de calculer cette similarité, mais elles sont toutes fondées sur le principe que la force d'association entre deux mots est un rapport entre le nombre de cooccurrences de ces deux mots (f_{ij}) et de leurs occurrences séparées (f_i, f_j). Ci-dessous, on présente une liste de trois équations (trois façons différentes) permettant de calculer la similarité pour l'agrégation des termes en analyses des données :

$$E_{ij} = \frac{f_{ij}{}^2}{f_i \times f_j} \quad (1)$$

$$Dice = 2\,\frac{f_i \cap f_j}{f_i + f_j} \quad (2)$$

$$S_{ij} = \frac{f_{ij}}{\min(f_i, f_j)} \quad (3)$$

L'équation E_{ij} détermine s'il existe une égalité entre l'ensemble des documents indexés par les mots *i* et *j*. E_{ij} égale à 1 si les mots *i* et *j* apparaissent dans les mêmes documents. S_{ij} met en avant la relation d'imbrication entre deux mots *i* et *j*. S_{ij} égale à 1 si l'ensemble des documents contenant le mot *j* est inclus dans l'ensemble des documents qui contient le mot *i*. *Dice* divise le nombre d'occurrences par le nombre de cooccurrences des deux mots et le résultat est multiplié par 2. Le deuxième type de mesure de similarité est la mesure de similarité de distance du cosinus en recherche d'information. Cette mesure de similarité consiste à mesurer la proximité ou l'éloignement entre deux unités lexicales ou des documents. La mesure de similarité de la distance du cosinus est appliquée au modèle vectoriel dans lequel chaque colonne représente un document et chaque ligne représente un terme inclus dans le document correspondant. Le vecteur de documents est noté comme $U=(u_i, u_j, ...u_k)$ et le vecteur de termes est noté comme $V=(v_i, v_j, \ldots\ldots v_k)$. La similarité de distance du cosinus est calculée selon la formule suivante (Aggarwal et Zhai, 2012 : 89):

$$cosine\ (U, V) = \frac{\sum_{i=1}^{k} f(u_i) \cdot f(v_i)}{\sqrt{\sum_{i=1}^{k} f(u_i)^2} \cdot \sqrt{\sum_{i=1}^{k} f(v_i)^2}} \quad (4)$$

La pondération est une méthode qui a pour objet de choisir à partir d'un texte les mots dont les poids sont les plus informatifs. La pondération des termes est exécutée selon un modèle de pondération qui se base sur deux niveaux d'information : l'information de niveau local qui dépend de la fréquence d'un terme dans un document donné ; l'information de niveau global qui dépend de la distribution d'un terme dans l'ensemble du corpus.

L'information mutuelle consiste à mesurer le degré de dépendance au sens probabiliste entre deux variables. On considère que deux variables sont indépendantes si la réalisation d'un mot n'a aucun rapport avec l'autre. Dans ce cas-là, l'information mutuelle est nulle. En revanche, si la relation entre deux variables est linéaire, ces deux variables sont dans une relation de dépendance de corrélation. La formule pour calculer l'information mutuelle est la suivante :

$$I(X, Y) = \sum_{x,y} P(x, y) \log \frac{P(x,y)}{P(x)P(y)} \quad (5)$$

dans laquelle (*X*, *Y*) est un groupe de variables, *P*(*x*) indique la probabilité de variable *X*, *P*(*y*) représente la probabilité de *Y*, et *P*(*x*, *y*) signifie la probabilité de cooccurrence de *X* et *Y*.

Le mot *vraisemblance* est souvent considéré comme un synonyme de *probabilité*. La différence entre *vraisemblance* et *probabilité* est issue des différents rôles de sorties et de paramètres dans le calcul. La probabilité est une fonction qui nous fournit la possibilité d'un résultat en prenant un ensemble de paramètres donnés à l'avance en entrée. Par exemple, si l'on jette une pièce de monnaie en l'air 100 fois et que cette pièce de monnaie a deux côtés face identiques, quelle est la probabilité que cette pièce de monnaie tombe côté face chaque fois? La vraisemblance est néanmoins une fonction qui nous fournit la possibilité d'un paramètre comme sortie en prenant un résultat indiqué comme entrée. Par exemple, si l'on jette une pièce de monnaie en l'air 100 fois et que cette pièce tombe côté face 100 fois, quelle est la vraisemblance que cette pièce de monnaie a deux côtés face identiques ? le log-

vraisemblance est une fonction de vraisemblance en se basant sur la fonction logarithmique. Le logarithme d'une fonction atteint la valeur maximum à la même position de la fonction elle-même. Ainsi, le log-vraisemblance est souvent utilisé à la place de la vraisemblance pour l'estimation du maximum de vraisemblance ou les autres techniques liées.

Le classifieur bayésien naïf est en fait une règle, un modèle, ou une fonction statistique construit par la méthode d'apprentissage automatique permettant de faire la classification automatique (Ibekwe-SanJuan, 2007 : 99). Le classifieur peut être binaire ou multiple. Par exemple, un classifieur binaire peut définir une fonction qui prend deux valeurs vrai et faux selon qu'un document a la catégorie indiquée ou pas et il peut aussi définir un seuil d'appartenance pour cette catégorie. Il existe plusieurs types de classifieurs : le classifieur bayésien naïf, le classifieur k-plus, le classifieur SVM (machine à vecteurs supports), etc. Dans cette partie, on va présenter en détail le classifieur bayésien naïf souvent utilisé dans les applications de traitement automatique. « […], le classifieur bayésien naïf cherche à prédire la valeur d'un nouvel objet (document, texte, événement, état) à partir d'une estimation des probabilités prenant en compte des connaissances ou observations existantes (*propabilities a priori ou prior probabilities*) et des croyances (*beliefs*)» (Ibekwe-SanJuan, 2007 : 99). L'estimation de la probabilité est exécutée sur la base des connaissances existantes et elle peut être représentée par la formule mathématique suivante :

$$P(R = r|e) = \frac{P(e|R=r)\cdot P(R=r)}{P(e)} \qquad (6)$$

dans laquelle la variable *R* représente le document ou l'unité lexicale dont la valeur le bayésien naïf doit prédire ; $P(R=r)$ indique la probabilité de la variable *R* ; $P(e|R=r)$ est la probabilité de la variable R étant donné *e* ; $P(R=r|e)$ représente la probabilité de la variable R étant données les observations *e* et $P(e)$ signifie la probabilité générale des observations. $P(e)$ est une probabilité déterminée en fonction des connaissances d'arrière-plan. On présente un exemple donné par Kevin Murphy : l'objectif du problème de cet exemple est de calculer la probabilité qu'un patient ayant été testé positif à une maladie l'ait réellement. On note cette maladie par *M*, le test positif par $T=+$ et le test négatif par $T=-$. Ainsi, la probabilité d'avoir cette maladie si le test est positif est $P(T=+|\ M=vrai)$; la probabilité d'avoir cette maladie si

le test est négatif est $P(T= -|M=vrai)$; la probabilité de ne pas avoir cette maladie si le test est positif est $P(T=+|M=faux)$ et la probabilité de ne pas avoir cette maladie si le test est négatif est $P(T-|M=faux)$. De plus, on suppose que cette maladie est rare et selon cette connaissance d'arrière-plan, on peut attribuer à $P(M=vrai)$ un pourcentage. Ainsi, la probabilité qu'un patient ayant été testé positif à cette maladie l'ait véritablement est calculé selon la formule suivante :

$$P(M=vrai|T=+) = \frac{P(T=+|M=vrai)\cdot P(M=vrai)}{P(T=+|M=vrai)\cdot P(M=vrai)+P(T=+|M=faux)\cdot P(M=faux)} \quad (7)$$

L'entropie est une fonction mathématique qui a pour objet de mesurer l'incertitude de l'information. Du point de vue d'un récepteur, plus les informations émises d'une source sont différentes, plus l'entropie est grande. Autrement dit, si une source est censée envoyer toujours les mêmes informations, par ex., on envoie toujours un même mot *bonjour*, l'entropie de ces informations est nulle. En revanche, si la source est réputée émettre les différentes informations, par ex., on envoie *salut* la moitié du temps et *bonjour* l'autre moitié, le récepteur est incertain du prochain mot à recevoir et l'entropie de ces informations est relativement grande. Le calcul de l'entropie est représenté par la formule mathématique suivante :

$$H_b(X) = -\,\mathrm{E}[log_b\, P(X = x_i)] = \sum_{i=1}^{n} P_i\, log_b(\frac{1}{P_i}) = -\sum_{i=1}^{n} P_i\, log_b\, P_i \quad (8)$$

X représente une variable discrète contenant n symboles (x_i, x_j .... x_k). P_i représente la probabilité d'un symbole x_i et H est l'entropie de la source X. On peut également calculer l'entropie conjointe de deux variables X et Y :

$$H(Y, X) = -\sum_{i,j} P(X = x_i, Y = y_i)\log_2 P(X = x_i, Y = y_j) \quad (9)$$

et l'entropie conditionnelle de Y relativement à X :

$$H(Y|X) = - \sum_{i,j} P(X = x_i, Y = y_j)\log_2 P(Y = y_j| X = x_i) \quad (10)$$

En TALN, l'entropie peut être utilisée pour aider à faire le filtrage, la classification ou la recherche d'information. Cependant, comme les autres techniques statistiques, l'entropie dont la performance dépend largement de la quantité d'information semble aussi faible en face de certaines tâches de traitement automatique des langues. Par exemple, supposons qu'on a un texte composé de 27 caractères équiprobables et l'entropie associée à chaque caractère selon la formule présentée ci-dessus est : $log_2 27=4,75…$ Cependant, dans un corpus, il est très rare que chaque caractère soit équiprobable. Le plus souvent, dans un corpus, certains caractères sont très fréquents et d'autres le sont moins. Dans ce cas-là, l'entropie de chaque caractère n'est pas si élevée. Malgré tout, l'importance de l'entropie dans la fouille de textes et dans les applications de traitement automatique des langues n'est pas négligeable et l'entropie est quand même une approche statistique importante pour aider à résoudre les problèmes dans les tâches de traitement automatique des langues.

2. État de l'art sur l'acquisition automatique des termes

Au fur et à mesure de l'accroissement des données textuelles électroniques, on est de plus en plus exigeant sur l'efficacité d'analyse automatique des textes. L'acquisition automatique des termes est une application du traitement automatique des langues naturelles qui consiste à extraire automatiquement une liste de termes à partir d'un corpus spécialisé. La construction automatique des ressources lexicales d'une ou plusieurs langues à partir d'un gros corpus, la facilité du travail de terminologie, l'indexation automatique et la création d'index thématiques se basent toutes sur l'acquisition automatique des termes. De plus, le résultat de l'acquisition automatique des termes peut aussi servir aux autres applications de traitement automatique des langues naturelles, telles que la compréhension automatique du contenu des données textuelles, l'extraction automatique des relations, la classification automatique, etc.

Un terme, dans la notion traditionnelle, est un label linguistique d'un concept qui est envisagé comme un élément entrant dans une structure de connaissances et est appréhendé par une suite d'opérations de classement (Jacquemin et Bourigault, 2003 : 600). Cependant, cette

notion traditionnelle ne peut pas être appliquée aux applications de TALN, car les types de ressources nécessaires varient selon les différentes applications. D'après Jacquemin et Bourigault (2003 : 601), un terme peut être un mot ou une séquence extraite du corpus spécialisé dans la terminologie informatique. Un terme est une unité lexicale ayant un sens spécifique dans un domaine de spécialité donné et est considéré comme la sortie d'une procédure d'analyse terminologique. Dans ce cas-là, la construction automatique des ressources lexicales est considérée comme la construction d'une instance de structures terminologiques d'un corpus. Dubois (1994 : 440) a indiqué : un terme est une unité signifiante constituée d'un mot (terme simple) ou de plusieurs mots (terme complexe), qui désignent une notion de façon univoque à l'intérieur d'un domaine. Ainsi, le terme s'oppose aux mots par sa référence inhérente à un domaine. Dans le domaine du traitement automatique des langues, l'usage des termes concerne le processus automatique, tels que la traduction automatique, l'extraction automatique de l'information, l'acquisition automatique des connaissances langagières, etc. Les termes dans le traitement automatique des langues sont considérés comme un type particulier de données lexicales. Les bases de termes comprennent les unités de multi-mots et les unités monolexicales. Elles peuvent aussi être améliorées ou enrichies sans cesse.

L'acquisition automatique des termes a pour objet de reconnaître automatiquement les termes d'un corpus spécialisé. Le résultat d'une extraction est une liste de candidats-termes. Jacquemin et Bourigault (2003 : 604) croit que l'acquisition automatique des termes est un axe de recherche associée à l'indexation automatique des textes et la génération de textes où la connaissance de structures idiomatiques est essentielle. Elle peut être divisée en deux sous-domaines : l'acquisition des termes initiaux et l'enrichissement des termes. Les trois applications principales basées sur les termes sont les suivantes : l'acquisition automatique des termes, l'acquisition automatique des termes de deux ou d'une langue et l'indexation automatique.

Les méthodes principales utilisées actuellement pour l'acquisition automatique des termes peuvent être divisées en trois types : les méthodes linguistiques basées sur les règles,

les méthodes purement statistiques et les méthodes hybrides en combinant les deux méthodes précédentes. Les méthodes linguistiques se basent sur des descriptions syntaxiques, grammaticales ou sémantiques. Par exemple, TERMINO (Lauriston, 1994), qui repose sur l'hypothèse qu'il y a des traces lexicales et syntaxiques permettant de détecter les termes dans le texte, identifie les termes complexes à l'aide des règles grammaticales génératives. Les méthodes statistiques en acquisition automatique des termes peuvent reposer sur la probabilité de cooccurrence, sur la fréquence d'apparition, sur le modèle de Markov, etc. L'extraction des termes basée sur la probabilité a pour objet de classifier les entités ou les paries d'entités en fonction de leurs caractéristiques contextuelles lexico-syntaxiques. Le calcul de ces caractéristiques contextuelles lexico-syntaxiques est du type probabiliste. Il existe de nombreuses fonctions statistiques pour calculer ces caractéristiques contextuelles lexico-syntaxiques, telles que l'information mutuelle, le C-value, le log-vraisemblance, le TF-IDF, le NC-value, etc. La plupart des méthodes adoptées actuellement pour l'acquisition automatique des termes sont les méthodes hybrides qui combinent les approches linguistiques et les approches statistiques. La combinaison des approches linguistiques et des approches statistiques permet d'augmenter largement l'efficacité de l'extraction automatique des noms composés, comme le permettent ACABIT (Boulaknadel, 2008), la méthode de Biskri (2004), etc.

Dans la partie suivante, on va présenter en détail les méthodes principales utilisées pour l'acquisition automatique des termes sous trois classifications : les méthodes linguistiques basées sur les descriptions syntaxiques des termes, les méthodes statistiques pour l'acquisition automatique des termes et les méthodes hybrides en combinant les approches linguistiques et statistiques.

En premier lieu, on va présenter quatre méthodes d'acquisition automatique des termes fondées sur les descriptions linguistiques : TERMINO，LEXTER (Bourigault, 1996), la méthode d'Hearst (1992) et la méthode de Snow (2004), etc. Ces méthodes identifient les termes complexes à l'aide des patrons morphosyntaxiques. TERMINO et LEXTER sont deux

systèmes semi-automatiques très populaires pour l'extraction automatique des noms composés.

En deuxième lieu, on va présenter trois méthodes statistiques classiques pour l'acquisition automatique des termes. Selon le principe des méthodes statistiques, ces trois méthodes sont de trois types : la méthode qui se base sur la cooccurrence, la méthode qui se base sur la théorie harrissienne et la vraisemblance, et la méthode qui se base sur la fréquence d'apparition. ANA (Enguehard, 1993) est une approche semi-supervisée qui apprend automatiquement les règles à partir d'un petit ensemble de corpus annoté pour reconnaître plus de termes. Les méthodes proposées dans la thèse de Morlane-Hondère (2014) sont une série de méthodes distributionnelles qui extrait les termes automatiquement en se basant sur le modèle vectoriel à l'aide des techniques de fouille de textes : le calcul de l'information mutuelle, la mesure de similarité, etc. La méthode de Boulbry (2004) est une méthode statistique non-supervisée qui extrait les termes spécialisés en fonction de la fréquence d'apparition.

Finalement, on va présenter des méthodes hybrides utilisées le plus souvent actuellement pour l'acquisition automatique des termes : la méthode d'Enguehard (2005), la méthode de Meilland et Bellot (2003), une stratégie de multi-extraction de Plaisantin Alecu, Thomas et Renahy (2012), etc.

2.1. Méthodes linguistiques basées sur les descriptions morphosyntaxiques

A. Automatic recognition of complex terms: Problems and the TERMINO solution (Andy Lauriston, 1994)

TERMINO est une approche basée sur la grammaire. Elle repose sur l'hypothèse à l'effet que le lexique et la syntaxe ont les traces avec lesquelles on peut repérer les termes dans le corpus. TERMINO est un système semi-automatique qui comprend trois sous-systèmes : un

pré-éditeur, un analyseur morphosyntaxique et une facilité de rédaction d'enregistrement (une interface permettant aux terminologues de produire les enregistrements de termes, des informations offertes du sous-système morphosyntaxique). En général, la méthode est la suivante : on fait un traitement dans le but premier de filtrer le texte et d'enlever les caractères de format ; ensuite, on parcourt le texte en produisant une analyse morphologique pour trouver des groupes nominaux ; finalement on génère les termes. Dans la partie suivante, on présente en détail chaque module de TERMINO.

Dans le module pré-éditeur, on fait un prétraitement qui filtre le texte et on enlève les caractères de format. Dans cette étape, on segmente d'abord les textes en phrases et on segmente ensuite chaque phrase en tokens. En même temps, on identifie les noms propres à l'aide des dictionnaires intégrés.

Le deuxième module, analyseur morphologique, comporte trois sous-modules : un analyseur morphologique, un analyseur syntaxique et un détecteur de synapsie. La synapsie est définie par David (1990 : 145) comme une séquence qui comprend un substantif-tête suivi ou précédé d'un modifieur (un adjectif, un syntagme prépositionnel ou une proposition relative). L'analyseur morphologique a deux fonctions : racinisation et étiquetage. La racinisation est une technique de transformation des dérivés ou des mots fléchis en leurs radicaux. L'étiquetage a pour objet d'associer les informations morphologiques et grammaticales (telles que les préfixes, les suffixes, les parties du discours, etc.) à chaque mot. Ensuite, l'analyseur syntaxique traite l'ambiguïté lexicale et apparie la structure syntaxique à la structure superficielle des tokens étiquetés avec les groupes structuraux dans les parenthèses. Le détecteur de synapsie comporte deux sous-modules : le constructeur de synapsie et le comparateur de synapsie. Le constructeur de synapsie sélectionne le substantif-tête et applique récursivement le module de représentation d'expansion. Le module de représentation d'expansion est un sous-module du constructeur de synapsie. Les autres sous-modules dans le constructeur de synapsie sont les suivants : le classifieur de substantif-tête, l'identifieur d'expansion, le catégoriseur d'expansion et le générateur de synapsie. Le

comparateur de synapsie a pour objet de reconnaître les similarités à la fois structurale et lexicale des candidats-structures.

Le troisième module de TERMINO consiste à fournir une interface permettant de faciliter le travail de terminologues pour choisir et enregistrer les termes à l'aide des informations morpho-syntaxiques offertes par le système. Ce semi-système réussit à minimiser le travail de terminologues dans l'identification des termes.

La précision de la version 1.0 de TERMINO atteint 70%. La complexité des contextes linguistiques provoque beaucoup de difficultés dans TERMINO pour reconnaître les structures syntaxiques. La coordination, les acronymes et les noms communs en lettres majuscules sont les trois principaux facteurs qui font causer des erreurs d'identification dans TERMINO. Les bruits obtenus par TERMINO sont de deux types : les bruits syntaxiques et les bruits terminologiques. Les bruits syntaxiques sont provoqués dans le parcours et les bruits terminologiques se produisent plutôt parce que les mots reconnus comme candidats font en fait partie de la langue générale. En ce qui concerne la perspective, le participe passé est une source de problèmes futurs. Les limites de TERMINO proviennent encore des limites des techniques linguistiques. Il y a encore un énorme développement à faire dans le domaine des techniques sémantico-syntaxiques. On doit accorder plus d'attention aux caractéristiques graphiques des textes techniques que sont l'utilisation des acronymes, la capitalisation du mot au début, l'utilisation de la ponctuation et la présence des conjonctions.

B. LEXTER, a Natural Language Processing Tool for Terminology Extraction (Didier Bourigault, Isabelle Gonzalez-Mullier et Cécile Gros, 1996)

LEXTER est une méthode de combinaison de modèles et une méthode de restrictions syntaxiques apprises et sélectionnées. Elle est accompagnée d'une interface pour valider et organiser les candidats-termes récupérés d'un corpus. Dans cette méthode, on étiquette le corpus à l'aide des patrons lexico-syntaxiques qui se basent sur des éléments lexicaux et des

catégories syntaxiques. Les opérations principales sont les suivantes : premièrement, on identifie les groupes nominaux à l'aide de transducteurs ; ensuite, les groupes nominaux sont segmentés en substantif-tête et expansion, et les sorties sont considérées comme candidates ; après, les candidats-termes sont automatiquement organisés dans le réseau en fonction des éléments lexicaux partagés dans des positions syntaxiques similaires ; finalement, les termes sont entrés dans les bases de données et doivent être validés par les experts. Dans la partie suivante, on va présenter en détail le principe et les processus de cette approche.

Le principe basique de LEXTER est de repérer les frontières des groupes nominaux. Dans LEXTER, on étiquette chaque mot avec l'information morphologique, grammaticale (partie du discours), et la forme lemme ; puis, on segmente le texte en localisant les frontières potentielles. Les informations syntaxiques et grammaticales sont fournies par un analyseur morphologique développé par la société Gsi-Erli. Pour repérer les frontières potentielles, on fait reconnaître les séquences lexico-syntaxiques qui ne peuvent pas faire partie des groupes nominaux terminologiques. Ce genre de séquences peut indiquer les frontières de groupes nominaux. Par exemple, les frontières peuvent être simplement un verbe, un pronom ou une unité polylexicale : Prép+article possessif, par exemple, *à une, aux, par une, pour la*, etc. Ainsi, on obtient un ensemble de segments. La plupart de ces segments sont les groupes nominaux. Les groupes nominaux reconnus de cette façon peuvent être les candidats-termes.

Néanmoins, on a constaté qu'on doit avoir une information syntaxique de sous-catégorisation pour traiter précisément certains cas de segmentation précisément. Par exemple, pour la séquence *une armoire de contrôle sensible à une élévation de température*, on fait une segmentation à la position *à une* selon le patron de frontière. Cependant, la segmentation doit être exécutée à la position de l'adjectif *sensible*, à savoir le système segmente à partir de *sensible à une*, en raison de la contrainte de cohérence syntaxique locale. Le système étiquette *à une* avec l'information syntaxique complément suivi de l'adjectif *sensible*. Pour résoudre ce problème, la solution la plus classique est la méthode CBEL (Corpus-Based Endogenous Learning) (Frérot et al., 2003) dans laquelle on fournit une liste d'adjectifs qui ont la possibilité d'apparaître à la position prédicative suivie de la préposition *à*. Avec cette liste, le

système segmente la séquence *logiciel disponible à la direction informatique* en (*logiciel*)/ (*disponible à la direction informatique*) et la séquence *contrôle sensible à une évaluation de température* en (*contrôle sensible*) / (*à une évaluation de température*).

Le module de Parser de LEXTER opère la décomposition de candidat-terme (groupe nominal) en deux constituants : un substantif-tête et une expansion. De cette façon, on génère les sous-groupes. Le module d'analyseur syntaxique de LEXTER est composé de règles du style qui peuvent indiquer quels sous-groupes à extraire à partir des groupes nominaux.

La désambiguïsation s'appuie sur la méthode basée sur le corpus (Bourigault 1993). Par exemple, on a une séquence nom1+Prép+nom2+Adj, (*pomme de terre pourrie, fard à paupière bleu, chemisier à pois rouges, pantalon avec ceinture vert*) et on a deux possibilités de la parcourir. Les deux schémas suivants expriment bien ces deux façons d'analyses :

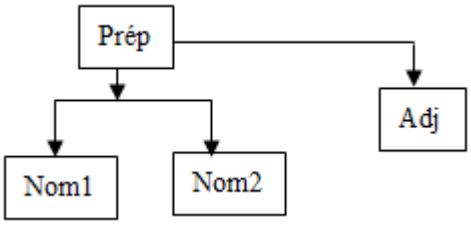

Figure 1 Parser 1

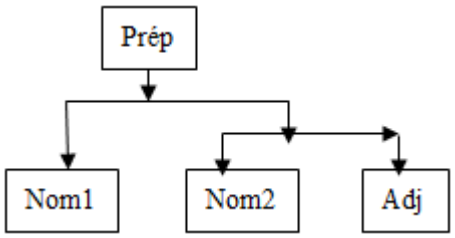

Figure 2 Parser 2

Ainsi, on obtient une liste de sous-séquences : nom2+Adj *(*comme *terre pourrie, paupière bleu, poids rouges, ceinture vert…)*, nom1+Adj (comme *pomme pourrie, fard bleu, chemisier rouges, pantalon vert...*) et nom1+Prép+nom2 (comme *pomme de terre, fard à paupière, chemisier à poids, pantalon avec ceinture...*). On cherche les occurrences de ces sous-séquences non ambigües dans le corpus. Si la sous-séquence nom2+Adj est trouvée, on choisit le parser1. Si l'on trouve la sous-séquence nom1+adj ou nom1+Prép+nom2, on choisit le parser2. Si ce qu'on trouve ne correspond pas à aucun cas ci-dessus, on choisit le parser1.

Le module de structuration consiste à structurer tous les candidats-termes. Il les organise en un réseau appelé le réseau terminologique. Dans ce réseau, chaque candidat-terme composé est lié avec les candidats-termes qui constituent sa tête et son expansion.

Le système LEXTER a été réalisé dans le but de contribuer à la construction et à la mise à jour des bases de données lexicales Thesaurus. Aujourd'hui, il est plutôt utilisé pour des applications de gestion des documents électroniques. Il peut être appliqué sur les différents corpus pour construire les différents types de produits terminologiques. Dans l'acquisition de l'information et des termes, le système LEXTER occupe une place signifiante et importante.

C. Automatic Acquisition of Hyponyms from large Text Corpora (Marti A., Hearst, 1992)

La méthode proposée par Hearst pour l'acquisition automatique des hyponymes est aussi une méthode très classique basée sur la description linguistique. Le principe de la méthode de Hearst (1992) est d'identifier les patrons lexico-syntaxiques marquant la relation d'hyponymie pour repérer les hyponymes. Les patrons lexico-syntaxiques correspondent aux critères suivants : ils sont fréquents et présents dans plusieurs genres de textes ; ils indiquent toujours une relation d'intérêt ; ils peuvent être reconnus avec peu de connaissance encodée à l'avance. Par exemple, le patron lexico-syntaxique *such NP as {NP,}*{(or|and)}NP* peut reconnaître les séquences comme *…works by such authors as Herrick, Goldsmith, and Shakespeare* dans laquelle on acquiert les paires d'hyponymes: hyponyme ("author", "Herrick"), hyponyme ("author", "Goldsmith"), hyponyme ("author", "Shakespeare"). La formule suivante est celle des modèles que Hearst (1992 : 3) propose pour acquérir les hyponymes:

1) NP{,NP}*{,} or other NP

Par ex., *Bruises, wounds, broken bones or other injuries… (Contusions, plaies, os cassés et autres blessures)*

hyponyme: ("bruise" (contusion), "injury" (blessure))

hyponyme: ("wound" (plaie), "injury" (blessure))

hyponyme: ("broken bone" (os cassé), "injury" (blessure))

2) NP{,NP}*{,} or other NP

Par ex.,*….temples, treasuries, and other important civic buildings (…temples, trésorerie, et autres bâtiments civils importants)*

hyponyme: (“temple” (temples), “civic building” (bâtiment civil))

hyponyme: (“treasury” (trésorerie), “civic building” (bâtiment civil))

3) NP{,}especially{NP,}*{or|and}NP

Par ex., *…most Europe countries, especially France, England, and Spain (la plupart des pays européens, nommément la France, l’Angleterre et l’Espagne)*

hyponyme: (“France” (la France),”European country” (pays européen))

hyponyme: (“England” (l’Angleterre),”European country” (pays européen))

hyponyme: (“Spain”(L’Espagne),”European country” (pays européen))

Selon ces patrons lexico-syntaxiques，on peut acquérir automatiquement les hyponymes du texte. Néanmoins, pour identifier tous les hyponymes d’un grand corpus, on a besoin d'un nombre suffisant de patrons lexico-syntaxiques. Ainsi, l’efficacité de l’acquisition automatique des relations lexicales est directement déterminée par le nombre de patrons lexico-syntaxiques découverts automatiquement. À partir des patrons lexico-syntaxiques préétablis, pour en trouver de nouveaux automatiquement, on adopte les procédures suivantes : premièrement, il faut décider une relation lexicale, par exemple, *équipe/membre* (une relation d’hyponymie) ; deuxièmement, il faut recueillir une liste de termes dans lesquels cette relation est exploitée, ex, *pays/Angleterre, verre/ustensile, téléviseur/électro-ménager, etc.*; troisièmement, il faut repérer chaque paire de termes dans le corpus et enregistrer l’environnement syntaxique et structural de chaque groupe de termes; finalement, une fois qu’un patron lexico-syntaxique est identifié, on l’applique pour trouver d’autres instances de relation.

Pour évaluer cette méthode, on compare les résultats obtenus avec les informations trouvées dans WordNet. On enregistre WordNet dans une structure d’arbre. On compare chaque résultat de groupes d’hyponymes (N0, N1) avec le contenu dans la hiérarchie

nominale de WordNet. Si N0 et N1 se trouvent tous les deux dans WordNet et leur relation d'hyponymie existe aussi dans la hiérarchie, la paire de termes (N0, N1) est validée comme groupe de termes ayant la relation d'hyponymie. Si N0 et N1 se trouvent tous les deux dans WordNet, mais leur relation d'hyponymie n'existe pas dans la hiérarchie nominale de WordNet, la relation d'hyponymie de cette paire de termes est mise en doute. Si l'un ou les deux membres de N0 et N1 sont absents dans WordNet, on les considère comme entrées. La précision de l'identification automatique des relations d'hyponymie est assez forte. Cependant, la plupart des termes dans la hiérarchie nominale de WordNet sont les noms non modifiés ou les noms avec un seul modifieur. Ainsi, on n'extrait que les relations consistant en noms non modifiés ou avec un modifieur unique dans les rôles hyponymes ou hyperonymes. Comme toutes les autres méthodes, la méthode de Hearst a aussi ses limites. Par exemple, cette méthode paraît légèrement inefficace pour l'identification de la méronymie ; la sous-spécification des hyponymes entraîne la difficulté de reconnaître les hyponymes pertinents (par ex., {"anguille", "espèces"}, mais quelles "espèces" ?); la dépendance de relations entre certains termes (par ex., hyponyme {"vaisseau", "navire"} est dépendant du contexte ; certains hyponymes sont des termes de niveau très élevé (par ex., hyponyme {"substance", "forme"}), etc. Ces limites peuvent occasionner des problèmes importants dans la recherche et le développement futurs dans ce domaine.

D. Learning syntactic patterns for automatic hypernym discovery (Rion Snow, Daniel Jurafsky et Andrew Y.Ng, 2004)

Snow, Jurafsky et Y.Ng (2004) ont proposé une autre méthode pour l'acquisition automatique des hyponymes. Cette méthode est aussi appuyée sur des patrons lexico-syntaxiques. On identifie les paires d'hyponymes à l'aide d'un petit nombre d'expressions régulières pré-élaborées manuellement et on exécute cette étape récursivement pour reconnaître plus de paires d'hyponymes. Les opérations principales sont les suivantes : on collecte les paires de noms à partir du corpus en identifiant les exemples de paires d'hyponymes avec WordNet ; pour chaque paire, on collecte les phrases dans lesquelles tous les deux noms sont présents ; on parcourt les phrases et on extrait automatiquement les

patrons d'arbre syntaxique, puis on exploite un classifieur d'hyperonymie basé sur ces patrons; pour le tester, on met une paire de noms dans la séquence de tests et on extrait l'environnement syntaxique entre ce groupe de noms, puis on utilise le classifieur pour déterminer si la paire de noms est dans la relation d'hyponymie.

Bref, on découvre les patrons lexico-syntaxiques indiquant la relation d'hyponymie d'une façon automatique et on entraîne un classifieur statistique à partir des groupes de noms dans la relation d'hyponymie identifiée. On applique le classifieur utilisé pour reconnaître plus de groupes d'hyponymes plutôt que de se servir de patrons syntaxiques. Les expérimentations prouvent que l'extraction des groupes d'hyponymes avec le classifieur arrive à une pertinence plus élevée. De plus, le classifieur peut être amélioré en ajoutant une ressource de connaissances.

En ce qui concerne les trois premières opérations, l'objectif est de trouver les patrons lexico-syntaxiques qui peuvent indiquer la relation d'hyponymie. On adopte la méthode "chemins dépendants" pour extraire les patrons lexico-syntaxiques. Dans cette méthode, un analyseur syntaxique de dépendance produit un arbre de dépendance qui représente les relations syntaxiques entre les mots. Dans l'arbre de dépendance, chaque mot est sous la forme lemmatisée et correspond à un nœud spécifique, et chaque relation est la relation syntaxique directe entre deux mots successifs l'un à côté de l'autre. On définit notre espace de patrons lexico-syntaxiques comme espace comprenant tous les chemins les plus courts dans l'arbre de dépendance. Pour recueillir les patrons lexico-syntaxiques, on parcourt premièrement chaque phrase dans le corpus et extrait les paires de noms de chaque phrase en utilisant WordNet. Les paires de noms sont étiquetées *Hyperonyme connu* ou *non Hyperonyme connu* en fonction de relation entre les mots. Ensuite, on met les paires de noms dans l'analyseur de dépendance pour trouver les chemins les plus courts entre chaque paire de noms. Les chemins les plus courts sont les patrons lexico-syntaxiques indiquant la relation sémantique entre chaque groupe de noms. Ainsi, on extrait les patrons syntaxiques de chaque phrase contenant des paires de noms.

La quatrième opération a pour but d'entraîner un classifieur pour l'acquisition automatique d'un plus grand nombre d'hyponymes. Le classifieur est entraîné à partir des patrons lexico-syntaxiques qu'on extrait à l'aide des techniques de fouille de textes. Le plus souvent, on entraîne le classifieur en utilisant le classifieur bayésien naïf multinomial, le classifieur bayésien naïf complémentaire ou la régression logistique (Rennie, Shih, Teevan et Karger, 2003).

Néanmoins, il existe quand même beaucoup de place à l'amélioration en ce qui concerne le classifieur. Pour l'améliorer, on peut établir une série de classifieurs de termes de coordination et combiner les patrons de termes de coordination et les patrons d'hyperonymes en tenant compte du fait que les "termes de coordination" dans le glossaire de WordNet sont considérés comme des noms ou des verbes qui ont le même hyperonyme. On définit deux probabilités de relations entre les noms : la probabilité que n_i soit l'hyperonyme de n_j et la probabilité que n_i et n_j sont les termes de coordination. Sur la base de ces deux probabilités, on compte une nouvelle probabilité d'hyperonyme que n_k est l'hyperonyme de n_i : $P_{new}(n_i<n_k) \propto \lambda_1 P_{old}(n_i<n_k)+ \lambda_2 \sum_j P(ni \sim nj) Pold(nj < nk)$.

Le classifieur de patron d'hyperonyme entraîné sur le corpus de Wikipédia est le classifieur le plus performant qu'on entraîne. Ce classifieur montre 54% d'amélioration par rapport au classifieur entraîné en se basant sur WordNet. Le classifieur entraîné par la régression logistique sur le corpus Newswrite a 16% d'amélioration en F-score par rapport au meilleur classifieur entraîné en se basant sur WordNet et le classifieur entraîné par le modèle en combinant les hyperonymes et les coordinateurs a relativement 40% d'amélioration de F-score.

En résumé, en ce qui concerne la méthode basée sur les règles, on établit une série de règles manuellement et on provoque l'apprentissage automatique des règles pour étiqueter le corpus à fouiller d'après un ensemble de documents entraînés (étiquetés) ou des centaines de ressources correspondantes (telles que WordNet，Thésaurus, etc.). De cette façon, plus de règles sont obtenues de façon automatique. On divise les méthodes d'apprentissage automatique des règles en deux types : la méthode « top-down » dans laquelle les règles sont

définies pour couvrir la plupart des instances entraînées (on a un ensemble de documents étiquetés dans lesquels chaque terme est étiqueté) ; la méthode « bottom-up» dans laquelle les règles sont définies pour reconnaître aussi les instances entraînées qui ne sont pas encore couvertes par les règles existantes.

2.2. Méthodes statistiques pour l'acquisition automatique des termes

En fonction de principes théoriques sur lesquels s'appuient les méthodes statistiques pour l'acquisition automatique des termes, on divise celles-ci en trois classes : celles qui sont basées sur la cooccurrence, celles qui le sont sur les caractéristiques contextuelles lexico-syntaxiques et celles qui le sont sur la fréquence d'apparition. Dans cette partie, nous détaillons trois méthodes statistiques pour l'acquisition automatique qui couvrent les trois types de méthodes décrites ci-dessus. Elles sont respectivement : ANA (1993), la méthode distributionnelle présentée dans la thèse de Morlane-Hondère (2014) et la méthode de Boulbry (2004) pour extraire les termes concernant la nostalgie. Du point de vue de l'apprentissage de machine, la méthode ANA est une méthode statistique semi-supervisée basée sur la cooccurrence ; la méthode présentée par Morlane-Hondère est statistique non supervisée qui s'appuie sur la théorie de distribution sémantique de Harris (1954) et la méthode basée sur la fréquence d'apparition est aussi une méthode statistique non supervisée.

A. Acquisition de terminologie à partir de gros corpus (Chantal Enguehard, 1993)

Bootstrapping (Efron et Tibshirani, 1993) est une méthode permettant d'apprendre de façon itérative plus d'instances de modèles d'une manière automatique à partir d'un petit ensemble d'instances de modèles. Ces modèles ne sont pas forcément des modèles linguistiques. On peut aussi dire que dans Bootstrapping, on étiquette un petit ensemble de corpus à partir desquels on entraîne un classifieur pour acquérir plus d'autres modèles.

ANA a adopté l'algorithme Bootstrapping et se base sur la cooccurrence pour l'acquisition automatique des termes. L'idée est qu'on enrichit notre liste d'entités en cherchant d'autres paires d'entités qui coexistent à partir d'un petit ensemble d'instances de relations ou de paires d'entités coexistant (une liste de semences). Dans ce cas-là, le contexte où les paires d'entités coexistent peut être un modèle de la relation à chercher pour identifier d'autres entités de la même relation. On ajoute cette paire dans la liste de semences et on répète le processus itérativement. Dans la partie suivante, on va présenter avec détails cette méthode semi-supervisée en s'appuyant sur la cooccurrence pour l'acquisition automatique des termes.

Le principe du système ANA modélise la capacité humaine à reconnaître des informations dont la morphologie varie alors que la sémantique reste la même. ANA représente les processus d'induction et de généralisation. Le développement du système ANA a pour but de répondre aux problèmes que posent des systèmes documentaires qui enregistrent des documents en réponse à la requête de l'utilisateur. Les descripteurs dans un thésaurus prédéfini sont nécessaires pour la qualité d'indexation. Néanmoins, il n'existe pas beaucoup de thésaurus préétablis, surtout dans les domaines de sciences et de techniques récentes. De plus, si l'on constitue un thésaurus d'un domaine manuellement, cela prend beaucoup de temps et de mains d'œuvre, ce qui coûte très cher. Ainsi, l'établissement d'un thésaurus d'un domaine de la manière automatique devient important.

Le système ANA est un système d'acquisition de termes par induction inspiré de l'apprentissage humain de la langue maternelle. Par exemple, pour apprendre les noms de dénominations, tels que *banane, eau, voiture,* etc., les enfants les apprennent en généralisant la concomitance fréquente de l'objet (Enguehard, 1993 : 5). Autrement dit, c'est la cooccurrence fréquente des entités qui facilite l'apprentissage des noms de dénominations pour les enfants. De plus, le développement du système ANA s'inspire aussi de la capacité humaine de reconnaître la même notion sous différentes formes pour reconnaître les termes sous différentes formes en utilisant les variations morphologiques et syntaxiques. Par exemple, *la pompe fuit, fuite de la pompe* et *fuite à la pompe* sont tous d'une pompe qui fuit. Le système ANA est composé de deux modules : Familiarisation et Découverte. Le module

Familiarisation consiste à extraire automatiquement les entités sous la forme de quatre listes : la liste de mots fonctionnels, la liste de mots fortement liés, la liste de mots de schémas et la liste Bootstrap (Efron et Tibshirani, 1993). Le module Découverte utilise les quatre listes extraites ci-dessus pour sélectionner les termes du domaine indiqué. L'architecture générale du système ANA est la suivante :

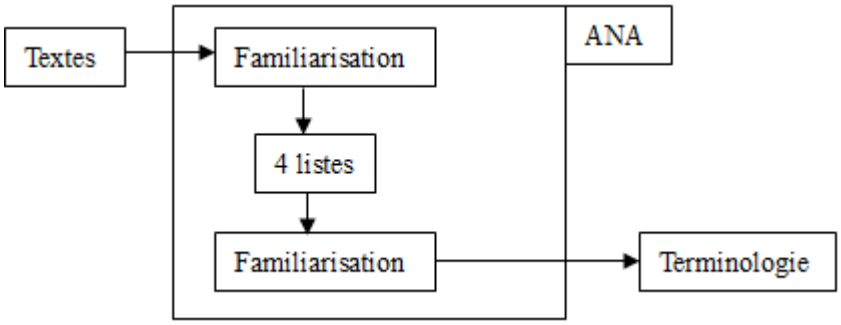

Figure 3 Architecture du système ANA

Dans le module Familiarisation, la liste de mots fonctionnels comprend des articles, des pronoms et quelques adverbes ; la deuxième liste comprend les mots fortement liés qui sont les mots contenant des espaces. Ce sont les mots provenant de la variation morphologique, par exemple, *de l'*, *de la*, *est en*, *et la*, *est le*, *on a*, etc. ; la troisième liste comprend les mots de schémas qui sont des mots fonctionnels structurant des syntagmes et la quatrième liste est une liste de Bootstrap qui contient un ensemble de termes de semences du domaine qui nous permet de faire le bootstrapping.

Le cœur d'ANA est le module Découverte qui consiste à extraire plus de termes à partir de la liste de Bootstrap par la cooccurrence des évènements. Un événement dans ANA est défini comme une occurrence d'un mot quelconque. « Deux événements sont cooccurrents s'ils sont séparés par W mots (non fonctionnels) ou moins. » (Enguehard, 1993 : 8). Il y a trois types de cooccurrences dans ANA : le type d'expressions qui contient les cooccurrences de deux termes ; le type de candidat qui est composé des cooccurrences séparées d'un terme, d'un mot ou d'un mot de schéma ; le type d'expansion qui groupe les cooccurrences d'un terme et d'un mot. On parcourt les textes et on enregistre l'ensemble des trois types de cooccurrences. Ensuite, le système trie les cooccurrences en fonction de leur fréquence. Pour les occurrences comprenant deux termes, si l'une d'elles est assez fréquente, le système

l'enregistre comme nouveau terme. Pour les occurrences de type de candidat qui comprend un terme, un mot de schéma ou un mot, si l'une d'elles est assez fréquente, le terme, le mot de schéma ou le mot compris est considéré comme un nouveau terme. Pour les occurrences ne comprenant qu'un terme et aucun mot de schéma, si l'une cooccurrence d'entre elles est assez fréquente, on prend la chaîne de caractères qui inclut le terme et un autre mot non fonctionnel comme nouveau terme. Par exemple, quand l'occurrence *structure* n'a ni un terme ni un mot de schéma et si la chaîne de caractère *structure interne* est assez fréquente, cette enchaîne est considérée comme le nouveau terme.

On a appliqué la méthode au corpus de 120000 mots issus du retour d'expérience de Super-Phénix[1] et on a recensé 3000 nouveaux mots. On a également appliqué la méthode aux autres corpus parmi lesquels il y a un corpus de 30000 mots traitant la commercialisation du miel à partir duquel on a extrait 350 nouveaux termes. Sur un corpus scientifique de 22000 mots, on a identifié 200 nouveaux termes. Le développement du système répond aux contraintes suivantes : la mauvaise qualité du corpus, l'absence de règles de grammaire, de dictionnaires et de la non-intervention de spécialistes. Pour améliorer le taux de satisfaction des spécialistes par rapport aux résultats fournis, on extrait automatiquement à partir des termes corrects une grammaire qui permet de faire le tri entre les éléments acceptables ou non.

B. Une approche linguistique de l'évaluation des ressources extraites par analyse distributionnelle automatique (François Morlane-Hondère, 2014)

Dans la thèse de Morlane-Hondère (2014), on présente une série de méthodes permettant d'extraire des termes par le procédé d'une analyse distributionnelle. Ce sont des méthodes qui se basent sur la distribution sémantique représenté par les caractéristiques contextuelles lexico-syntaxiques. L'analyse distributionnelle provient de la théorie de sémantique distributionnelle de Harris (1979) dans laquelle on pose qu'une unité linguistique se définit en fonction de ses contextes d'apparition dans les textes. Selon Harris (1979), le sens d'un mot

[1]http://fr.wikipedia.org/wiki/Superph%C3%A9nix#Retour_d.27exp.C3.A9rience_sur_la_fili.C3.A8re_.C3.A0_neutrons_rapides

est défini par l'ensemble des contextes qui le contiennent. L'analyse distributionnelle consiste à décrire la façon de combiner des unités qui sont les phonèmes et les morphèmes. Deux mots qui se trouvent dans les mêmes contextes ont la possibilité d'avoir une relation sémantique entre eux. La méthode utilisée pour une analyse distributionnelle afin d'extraire des termes dans la thèse de Morlane-Hondère est du type statistique non supervisé. Son application est réalisée à l'aide des techniques de fouille de données, telles que l'information mutuelle, les mesures d'association (la similarité), le log-vraisemblance, etc. Dans la partie suivante, on va présenter avec plus de détails les opérations principales de ces méthodes distributionnelles appliquées dans l'extraction automatique des termes.

Premièrement, on fait une série de prétraitements : la « tokenization » qui permet la segmentation des textes en tokens, la lemmatisation pour enlever les marques de genre, de nombre et de conjugaison de chaque token, et l'étiquetage morphosyntaxique qui fournit les informations morphosyntaxiques (les parties du discours).

Ensuite, on essaie de définir les conditions pour contrôler si deux mots appartiennent à un même contexte. On a deux façons de considérer le contexte : par la conception syntaxique du contexte et par la cooccurrence simple. Dans l'approche syntaxique, on pose qu'un mot et son contexte (qui est un ensemble de mots) entretiennent une relation de dépendance syntaxique.

Troisièmement, on extrait les contextes d'un mot à partir d'un grand corpus et on procède à des mesures de pondération. Cette procédure a pour objet de choisir les contextes les plus caractéristiques du mot, car on considère que certains contextes d'un mot sont meilleurs pour décrire ce mot par rapport aux autres. Par exemple, *devenir_suj* ou *se trouver_suj* ne nous donnent que peu d'informations sur la nature sémantique de *couvent*, car les contextes *devenir_suj* et *se trouver_suj* sont présents très fréquemment. Ils ont une classe d'arguments très générale. En revanche, les contextes *inhumer_à* et *moine_de* sont moins fréquents. Ils ont un spectre très réduit des termes qui coexistent avec eux. Leur cooccurrence avec *couvent* est beaucoup plus significative : ils nous indiquent plus de caractéristiques du mot *couvent*. On peut calculer ce rapport d'exclusivité entre un mot et ses contextes à l'aide des techniques de fouille de données : les mesures d'association comme TF-IDF (Sparck Jones, 1972), le T-

score (Church et Hanks, 1990), la log-vraisemblance (Dunning, 1993) ou l'information mutuelle (Manning et al., 2008). Le calcul de l'information mutuelle est une étape nécessaire pour la génération des ressources distributionnelles dans de nombreux modèles. Dans la thèse de Morlane-Hondère, on a éprouvé la procédure de pondération par le calcul de l'information mutuelle :

$$IM\ (x, y) = log\ (freqTot \cdot \frac{freq(xy)}{freq(x) \cdot freq(y)}) \quad (11)$$

Dans l'étape suivante, on compare les contextes pondérés et on attribue à chaque mot du corpus un score de proximité distributionnelle. Il existe deux modèles pour mesurer la proximité distributionnelle : le modèle géométrique et le modèle probabiliste. Dans le modèle géométrique, la distribution d'un mot est considérée comme un vecteur à *n* dimensions dans un espace. Le nombre de dimensions est égal au nombre de contextes qui permettent de caractériser le mot. L'information mutuelle sert à mesurer le degré de propriété des contextes du mot. Si le score de l'information mutuelle est plus élevé, la propriété du contexte est plus grande. Dans le modèle probabiliste, on calcule la probabilité que deux mots aient des distributions similaires. « Les distributions ne sont plus envisagées comme des vecteurs dans des espaces sémantiques mais comme des ensembles d'attributs dont il s'agit de mesurer le recouvrement (Hindle, 1990; Ruge, 1992; Grefenstette, 1994b; Lin, 1998b). » (Morlane-Hondère, 2014 : 50)

La dernière étape consiste à classifier les mots les plus similaires du point de vue de leurs contextes. On adopte la méthode de la similarité cosinus pour calculer cette dernière. La formule est la suivante :

$$cos\ (x, y) = \frac{\sum_{i=1}^{n} x_i \cdot y_i}{\sqrt{\sum_{i=1}^{n} x_i^2} \cdot \sqrt{\sum_{i=1}^{n} y_i^2}} \quad (12)$$

x et *y* sont les deux vecteurs de *n* dimensions. L'idée est de calculer le cosinus de deux vecteurs afin de mesurer leur degré de similarité. Pour la classification, on a les méthodes hiérarchiques, k-moyennes, centres mobiles, etc. De cette façon, on obtient les classes de mots

partageant des propriétés distributionnelles. Ces classes de mots ont les mêmes contextes d'apparition.

« Les travaux qui ont été menés par Harris et son équipe sur les corpus de textes médicaux reposent sur l'hypothèse d'une corrélation entre les classes distributionnelles et les classes d'objets : […] Habert et Nazarenko (1996 : 10) ont pu faire le même constat dans leur étude pionnière sur des textes en français : De la même manière, la syntaxe ne permet pas directement de délimiter des classes de mots reflétant une notion. » (Morlane-Hondère, 2014 : 58)

C. De l'analyse lexicale à la construction d'Echelles psychométriques : Application à la mesure du tempérament nostalgique (Gaëlle Boulbry, 2004)

Pour l'acquisition automatique des termes spécialisés, on utilise aussi la méthode statistique basée sur la fréquence d'apparition. Les techniques statistiques basées sur la fréquence d'apparition sont généralement utilisées pour établir des comparaisons de corpus, c'est-à-dire pour comparer un corpus d'analyse à un corpus de référence afin d'identifier des termes propres au corpus d'analyse. C'est plutôt une tâche de la lexicométrie qui consiste à repérer les mots caractéristiques d'un corpus, d'un thème ou d'une population étudiée. Les techniques de cette analyse de la spécificité peuvent aussi servir à mesurer le degré de figement syntaxique. Cette méthode est fondée sur l'hypothèse à l'effet que les unités qui sont fréquentes dans un domaine, qui sont présentes seulement dans un domaine ou qui sont plus fréquentes dans un domaine que dans l'usage général, ont une probabilité plus forte d'être des termes. Par exemple, les différents individus de métiers emploient souvent un vocabulaire qui varie pour parler d'un même objet. Il existe un écart de fréquence d'apparition entre les termes caractéristiques d'un domaine particulier et les termes généraux. On utilise la technique de probabilité pour détecter le lexique caractéristique de chacun.

L'analyse statistique du discours concerne différents aspects : les analyses thématiques qui concernent le thème du discours, les analyses syntaxiques qui s'occupent de sa structure, et les analyses lexicales qui traitent ses caractéristiques lexicales (richesse du vocabulaire, répétition des mots, etc.). L'étude de la spécificité du vocabulaire a pour objet de reconnaître les termes spécialisés du corpus. Les termes spécialisés peuvent être considérés comme les mots caractéristiques du corpus. Boulbry et Kercadio (2004 : 157) a souligné la problématique de l'étude de la spécificité du vocabulaire et a présenté une approche statistique basée sur la fréquence d'apparition pour extraire ceux qui sont caractéristiques du corpus. Dans la partie suivante, on va présenter cette méthode avec plus de détails.

Pour identifier les termes spécialisés et analyser la spécificité du vocabulaire par la fréquence, on segmente le corpus en tokens et on trie d'abord les tokens selon la fréquence d'occurrence. On distingue ensuite les termes spécifiques des termes généraux en comparant les distributions de ces termes dans des corpus monothématiques. L'idée est que plus un terme peut présenter un thème, plus il est spécifique au corpus de ce thème et plus il est rare dans le corpus des autres thèmes. Gaëlle Boulbry adopte l'approche qui construit l'indicateur de spécificité en tenant compte de l'ambiguïté des mots.

L'approche adoptée par Sphinx Lexica (Moscarola, 1995) permet d'identifier les termes spécialisés par la construction d'un indicateur de spécificité n'utilisant qu'une estimation empirique de la probabilité des mots. Plus précisément, supposons qu'on a un estimateur de probabilité $P(m|T)$ [4] du mot *m* dans les documents du thème *T* et d'un estimateur de probabilité $P(m)$ du même mot dans n'importe quel document, l'indicateur d'utilité thématique du mot *m* est calculé selon la formule suivante : $\frac{p(m/T)}{p(m)}$. En fait, c'est la probabilité de trouver *m* dans un corpus T sur la probabilité de trouver *m* dans n'importe quel corpus.

Une autre approche pour calculer l'indicateur de spécificité est plus simple à mettre en œuvre. Dans cette approche, on compare la probabilité de la présence du mot *m* dans le thème *T* avec sa probabilité de présence dans le corpus des autres termes *T'*. La formule est la suivante : utilité=$\frac{P(m/T)}{P(m/T')}$ (P (m/T') ≠0).

L'ambiguïté des mots baisse la pertinence de l'entraînement de l'indicateur. Par exemple, le terme *avions* dans le domaine des aéronefs est un nom qui signifie un appareil de navigation aérienne, mais il peut aussi correspondre au verbe conjugué avoir dans les autres contextes. Ainsi, Boulbry et Kercadio (2004 : 167-170) a introduit la notion d'ambiguïté de sens dans l'indicateur. Pour désambiguïser le sens des termes, on fait appel aux rapports de vraisemblance utilisant les expressions suivantes : Utilité=$\frac{P(m/T)}{p(m/T')}$; Ambiguïté=$\frac{P(m'/T)}{m'/T'}$. Cette technique pour caractériser les termes a été prouvée depuis longtemps. Le principe est que plus un mot est ambigu dans le discours *T*, plus il est absent dans le corpus *T* que dans le corpus des autres thèmes *T'*. Cette méthode est adoptée par la plupart des systèmes visant les évaluations internationales organisées par l'organisme américain NIST (National Institute of Standards and Technology- en français l'institut national de standard et technologie) sur la recherche d'information. La spécificité d'un terme peut être mesurée selon le poids de l'offre : OW=$P\ (m/T)\log\frac{\frac{P(m/T)}{P(m/T')}}{p(m'/T')}$. Si le poids de l'offre d'un terme est positif, ce terme est plus spécifique au corpus de *T* que celui de *T'*.

En linguistique, il est possible que *T* soit un petit corpus et que *T'* soit volumineux. Afin que l'estimateur de poids de l'offre permette un calcul dont les résultats ne changent pas en fonction de la taille de corpus, on note le terme du corpus *T* comme ε_t et le terme du corpus *T'* comme $\varepsilon_{t'}$. Ainsi, on obtient la formule : $\varepsilon_t = \frac{1}{1+\frac{total\,T'}{total\,T}}$.

2.3. Méthodes hybrides combinant les approches linguistiques et statistiques

Actuellement, pour l'extraction automatique des termes, la plupart des méthodes utilisées sont des méthodes hybrides qui combinent les approches linguistiques et les approches statistiques. Les expérimentations prouvent que la combinaison des approches linguistiques et des approches statistiques augmente souvent largement la pertinence de l'identification des termes. Dans cette partie, on va présenter une méthode semi-automatique proposée par Enguehard (2005) et deux autres méthodes hybrides qui sont entièrement automatiques.

A. Un banc de test pour la reconnaissance de termes en corpus (Chantal Enguehard, 2005)

Enguehard (2005, p.273-286) a présenté une méthodologie de construction d'un banc de test pour la reconnaissance des termes d'une manière semi-automatique. Cette méthodologie a pour objet de faciliter le travail des spécialistes et de minimiser le temps de leur travail. Dans cette méthodologie, on a un système de reconnaissance de termes (SRT) qui calcule les occurrences de termes identifiés et les propose aux terminologues pour valider les candidats-termes. Cette approche est exécutée itérativement pour construire la ressource lexicale. Les SRT identifient les termes d'un corpus et produisent un corpus indexé. Le schéma suivant montre bien le fonctionnement des SRT :

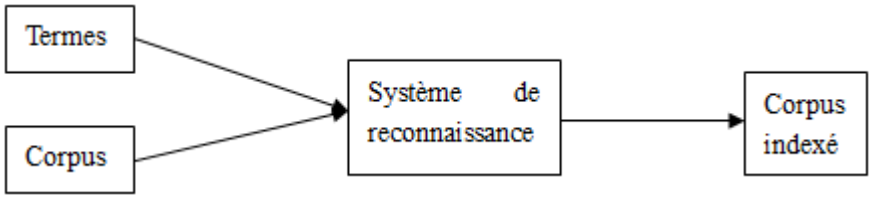

Figure 4 Architecture du système SRT

Les opérations sont les suivantes : les SRT indexent le corpus en signalant les emplacements de chacune des variantes du terme reconnu ; l'outil d'alignement permet de comparer le corpus indexé avec le corpus de référence dans les SRT et produit les formulaires de validation où sont regroupées les variantes (qui ne sont pas listées dans le corpus de référence) identifiées par les SRT ; finalement, les spécialistes valident les termes et l'outil d'alignement ajoute les termes validés dans le corpus de référence ce qui produit un nouveau corpus de référence. Le schéma suivant nous montre bien chaque opération :

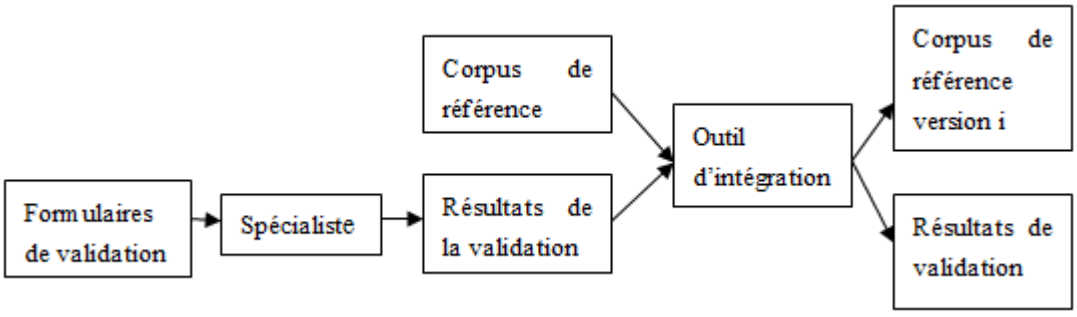

Figure 5 Operations dans le système SRT

On applique cette série d'opérations successivement et le nombre de termes dans le corpus de référence augmente petit à petit.

Cette méthodologie est appliquée au corpus « métal » qui est fourni par l'Institut de l'Information Scientifique et Technique. Ce corpus comprend 1280 résumés d'articles scientifiques dans le domaine de la chimie des métaux et une liste de 6582 termes du domaine (5239 termes complexes et 1343 termes simples) assortis d'une traduction en français et d'un étiquetage morphosyntaxique sans définition. Les textes sont rédigés en anglais. On a utilisé deux systèmes de reconnaissance FASTER et SYRETE. FASTER qui se fondent sur les règles morphologiques des mots composés. Par exemple, dans la métarègle *Coor (X1->X2X3)=X1->X2C4X5X6X3* qui décrit la coordination, *X1* est la balise de métarègle et cette règle qui traite une succession de deux mots *X2* et *X3* produit une autre règle dans laquelle *X2* et *X3* sont coordonnés par *C4* et les mots *X5* et *X6*. C'est plutôt une méthode qui se base sur la description morphosyntaxique. Néanmoins, SYRETE est un modèle statistique. Il s'appuie sur la mesure de distance entre deux chaînes de caractères qui est calculée selon la formule suivante :

$$WD(x, y)=dist(x, y)// (|x|+|y|) \quad (13)$$

dans laquelle $|x|$ indique le nombre de symboles apparaissant dans *x*. *WD* est égal à 0 quand les chaînes sont strictement égales, et il égale à 1 quand les chaînes n'ont aucun symbole en commun. La référence et les corpus indexés produits par SRT sont tous au format XML, car les SRT sont un système de communication. Le format XML facilite l'utilisation des ressources produites par d'autres modules dans le système.

FASTER permet d'identifier 4409 occurrences dont 2934 sont des variantes de termes. SYRETE permet d'identifier 2818 occurrences de variantes de termes dont 1048 ne figurent pas dans la version 2 du corpus. Le taux de rappel calculé pour les variantes de termes identifiées par FASTR est seulement de 63% et le taux de précision est de 89%. Pour SYRETE, on a un taux de rappel de 68.21% et un taux de précision de 98.84%. L'échec d'identification des occurrences est la cause des variations typographiques des termes, telles

que la permutation de caractères de la casse, l'insertion d'un tiret, etc. Malgré ce dysfonctionnement, le fonctionnement de SRT est quand même considéré bon.

B. Extraction automatique de terminologie à partir de libellés textuels courts (Jean-Claude Meilland et Patrice Bellot, 2003)

Meilland et Bellot (2003) ont présenté une approche permettant d'extraire automatiquement la terminologie d'un corpus. Supposons un corpus composé de textes de libellés courts (dépêches journalistes, petites annonces, descriptions de produits commerciaux, publicités, forums de discussion) qui ont une sémantique particulièrement riche et dans lesquels il y a relativement peu d'ambiguïtés. Ce corpus utilisé dans l'application de la méthode de Meilland et Bellot est issu des données du projet industriel SmartyCart[2] obtenues à partir des différents sites de magasins en ligne (www.ooshop.fr, www.telemarket.fr, www.houra.fr, www.auchan-direct.fr). L'objectif de cette méthodologie est de reconnaître les nouveaux termes au fur et à mesure du développement et du changement des libellés courts des produits commerciaux provenant de la grande distribution. Dans ce cas-là, l'extraction du terme complexe *lait écrémé* (NADJ) permet à la fois l'identification d'un nouveau type de produit particulier (puisque de nombreux libellés débutent par ces deux mots) et la désambiguïsation du mot *lait* (*lait à boire* ou *lait de beauté*).

Pour cela, on catégorise le corpus de libellés et on définit la terminologie de chaque catégorie. Par comparer la présence et la cooccurrence des termes d'un nouveau produit dans les différentes catégories de libellés, on classe les nouveaux termes. Dans la partie suivante, on va présenter en détail cette méthodologie. Cette méthode se base sur les collocations similaires de Manning et Schütze (1999) et des critères d'association de Daille (1994). L'idée est que l'on associe à chaque terme les statistiques relatives dans une catégorie, les statistiques relatives dans les autres et les statistiques relatives dans le corpus afin d'identifier le nouveau terme d'un produit et le classifier. Dans la partie suivante, on va présenter cette méthodologie avec plus de détails.

[2] http://smartycart.com/presentation.htm

SmartyCart constitue le corpus catégorisé par les moteurs de recherche des sites web de la grande distribution. Ensuite, on catégorise et filtre les produits manuellement pour extraire les termes représentatifs de chaque catégorie et les étiqueter. De cette façon, on construit un corpus d'apprentissage pour des algorithmes de classification et de structuration. Le tableau suivant montre un exemple :

Catégorie Lait écrémé	
Noms	Lait, Silhouette
Marques	Candia, Lactel, Danone, Carrefour
Quantités Emballages Unités	 Brique, Bouteille Litre
Adjectifs (Qualifiants)	Ecrémé, Vitaminé, Bébé

Tableau 1 Catégorie Lait écrémé

Pour extraire les termes les plus représentatifs de chaque catégorie d'un corpus de départ catégorisé, on calcule la variance et certains critères d'associations des termes aussi bien dans chaque catégorie que dans l'ensemble du corpus. Notre méthode se base sur les mesures de distances et les critères d'associations. Nous évaluons chaque critère dans une catégorie et dans le corpus en entier. Pour cela, nous utilisons une évaluation graphique sur des valeurs normalisées et triées. Ainsi un critère est retenu si sa courbe met en évidence des seuils. Pour les études de collocations, on calcule la moyenne des distances séparant les deux mots d'un couple de termes dans sa catégorie ($\sigma^2 = \frac{\sum_{i=1}^{n}(di-u)2}{n-1}$). *n* est le nombre de fois où le couple apparaît, d_i est la distance en nombre de mots séparant les deux mots du couple dans le *i*-ème libellé où il apparaît et *u* est la distance moyenne séparant les deux mots du couple. En ce qui concerne les études de critères d'associations, « D'un point de vue statistique, les deux lemmes qui forment un couple sont considérés comme deux variables qualitatives dont il s'agit de tester la liaison. » (Daille 1994 : 116). La matrice de contingence suivante est à la base des critères d'association testés :

	L_j	L_j' avec j'≠j
L_i L_i' avec i'≠i	a c	b d

Tableau 2 Matrice de contingence

a est le nombre d'occurrences du couple de mots (L_i, L_j). *b* est le nombre d'occurrences des couples où L_i est le premier élément d'un couple et L_j n'est pas le second. *c* est le nombre d'occurrences des couples où L_j est le second élément du couple et L_i n'est pas le premier. *d* est le nombre d'occurrences de couples où ni L_i ni L_j n'apparaissent. Le coefficient de proximité simple SMC (simple matching coefficient) est : SMC = $\frac{a+b}{a+b+c+d}$. Le coefficient d'Ochiai donne le résultat plus fin que SMC : OCH = $\frac{a}{\sqrt{(a+b)(a+c)}}$.

C. La "multi-extraction" comme stratégie d'acquisition optimisée de ressources (non) terminologiques (Blandine Plaisantin Alecu, Izabella Thomas, Julie Renahy, 2012)

Plaisantin Alecu, Thomas et Renahy (2012) ont proposé une autre méthode hybride pour l'acquisition automatique des termes. Les travaux faits par l'équipe de Plaisantin Alecu prouvent que la multi-extraction (la coopération de plusieurs extracteurs de termes (EdT)), donne des résultats significativement meilleurs que l'extraction via un seul outil. Dans la partie suivante, on va présenter premièrement les notions de langue contrôlée et l'unité lexicale de la langue contrôlée introduites dans la méthode de Plaisantin Alecu. Ensuite, on va présenter les opérations précises de la méthode.

Une langue contrôlée (LC) est une langue circonscrite à un domaine et à un environnement de rédaction précis, c'est-à-dire pour un public restreint et pour un type de textes particulier. On considère toutes les unités du lexique de la langue contrôlée (LLC) comme des Unités Lexicales. Un texte écrit en LC inclut les différents types d'UL, les termes du domaine, ceux d'un autre domaine, les unités lexicales (UL) du lexique général portant un sens spécifique dans le domaine traité ou celles du lexique général dans la composition de termes.

En se basant sur les notions introduites sur LC et LLC, on commence par la constitution d'un corpus. On a constitué une base de 14 modes opératoires d'immunobiologie de l'Etablissement Français du Sang (EFS) en Bourgogne France-Comté et aussi un LLC de

référence à partir de cette base. Le lexique de référence contient 1512 UL pour 1729 formes fléchies, 7 catégories syntaxiques fonctionnelles (ADV, N, V, etc.), 92 matrices morphologiques distinctes et 2 statuts lexico-terminologiques distincts.

Nous utilisons les EdT (extrateurs de termes) suivants : Acabit qui procèdent par l'identification des groupes nominaux complexes sur des matrices syntagmatiques pour l'extraction de bi-termes, par le regroupement de variantes et par le filtrage statistique à la fin ; YateA qui enchaîne l'identification des groupes nominaux à partir des frontières morphosyntaxiques, le calcul de leurs structures en substantif-tête et modifieur, et l'exploitation de ces structures pour l'analyse des groupes nominaux restants ; TermoStat qui fonctionne par la détection de CT sur patrons morphosyntaxiques, la pondération et le filtrage selon la spécificité de chaque CT. Finalement, on a fait une comparaison sur les résultats d'un Edt, cumulés et communs (intersection) afin d'estimer la capacité des EdT à recenser les UL.

En ce qui concerne l'évaluation, l'objectif est d'évaluer le recensement des UL (l'ensemble des UL d'un LLC) et le recensement des termes (des UL de statut terminologique d'un LLC). Pour chaque tâche, nous devons évaluer les résultats de chaque Edt pris séparément, évaluer les résultats cumulés de tous les Edt et évaluer des résultats consolidés ou communs. Sur l'évaluation des résultats de chaque Edt, le recensement des UL du LLC est de 52% en utilisant un seul Edt. Sur l'évaluation des résultats cumulés, le cumul des résultats des 3 Edt permet de couvrir presque les trois-quarts du lexique de référence et pour l'évaluation des résultats communs, la meilleure approximation est de 74%. Dans la deuxième tâche, la meilleure précision obtenue du recensement des résultats d'un Edt est de 28%. La meilleure précision des résultats cumulés des 3 Edt atteint 23% et 37% pour les résultats communs. En résumé, cumuler les résultats de tous les EdT permet de couvrir 79% des termes et la meilleure façon de déterminer le statut d'une UL est d'utiliser les résultats communs aux 2 Edt TermoStat et YaTeA.

3. État de l'art sur l'extraction automatique des noms composés

La reconnaissance automatique des mots composés est toujours une problématique incontournable pour l'acquisition automatique des termes, car les termes dans le traitement automatique des langues contiennent non seulement les termes simples mais aussi les termes complexes. L'efficacité de la reconnaissance automatique des noms composés influence directement la pertinence de l'acquisition automatique des termes. La reconnaissance des termes composés dans le texte courant est une tâche cruciale dans la phase d'extraction des termes.

Cependant, jusqu'à présent, il existe quand même de nombreuses difficultés pour l'identification automatique des noms composés. Ces difficultés proviennent de l'abstraction et la complexité de la langue naturelle. Les difficultés de la reconnaissance automatique des termes composés peuvent être divisées en deux types : les difficultés découlant de la reconnaissance et les difficultés de la désambiguïsation. Ces difficultés indiquent les différents contenus aux différents niveaux de la langue. Au niveau lexical, la première difficulté est d'identifier les frontières des termes composés et la deuxième difficulté est de faire la désambiguïsation morphosyntaxique, par exemple, *terminal* dans *le terminal d'ordinateur* est un nom, alors qu'il est un adjectif dans *équipement terminal*. Au niveau syntaxique, la première difficulté est de distinguer les constituants des termes composés des constituants de la phrase, par exemple, *The service has closed user groups (Le service a fermé les groupes d'utilisateurs/Le service a les groupes d'utilisateurs fermés) (closed (fermé)* peut être un verbe ou un constituant du mot composé *closed user groups (les groupes d'utilisateurs fermés))* et la deuxième difficulté est de distinguer les mots composés des groupes nominaux libres. Au niveau sémantique, les frontières de termes composés peuvent être déterminées par certaines analyses sémantiques.

En ce qui concerne la définition des mots composés, on a deux moyens de définir : le moyen de critères linguistiques et le moyen de critères statistiques. « La plupart des critères

linguistiques pour déterminer si une combinaison de mots est une expression multi-mots (EMM) sont basés sur des tests syntaxiques et sémantiques » (Constant, Sigogne et Watrin, 2012 : 2). Par exemple, l'expression *caisse noire* est une EMM, car elle n'accepte pas de variations lexicales (**caisse sombre*) ni d'insertions (**caisse très noire*). Au moyen de mesures statistiques associatives, on identifie les associations habituelles de mots (fondées sur la notion de fréquence). Dans ces définitions linguistiques, les expressions multi-mots sont classées selon leurs parties du discours globales. Les termes complexes peuvent aussi être divisés selon le degré de figement : certaines EMM acceptent un certain degré de variation et certaines n'acceptent rien du tout. Les termes composés peuvent aussi être classés en catégories : idiomes nominaux composés de noms propres, noms communs étrangers et noms communs. Les termes composés acceptent l'inflexion mais pas l'insertion. Cependant, certaines expressions multi-mots comportent un nom ou un adjectif permettant l'insertion de modifieurs.

Actuellement, la plupart des méthodes utilisées pour l'extraction automatique des termes composés sont les méthodes qui combinent les approches linguistiques et les approches statistiques. Dans la partie de l'état de l'art de l'acquisition automatique des termes, on a présenté deux méthodes linguistiques basées sur les descriptions morpho-syntaxiques : TERMINO et LEXTER. Ces deux méthodes sont aussi tous les deux les méthodes classiques pour la reconnaissance automatique des termes composés. Dans la partie suivante, on va présenter cinq autres méthodes de l'extraction automatique des termes composés. Dans un premier temps, on va présenter un système semi-automatique de Biskri (2004) et un outil semi-automatique ACABIT de Boulaknadel (2008) pour l'extraction automatique des noms composés. Dans le deuxième temps, on va présenter deux méthodes statistiques : une se base sur les cooccurrences, qui est la méthode proposée par Lebarbé (2002) et l'autre se base sur des techniques de probabilité, qui est une méthode de Parser de Green (2012). Finalement, on va présenter la méthode hybride de Constant (2012) qui s'appuie sur à la fois la probabilité et l'analyse syntaxique.

A. L'extraction des termes complexes : une approche modulaire semi-automatique (Ismaïl Biskri, Jean-Guy Meunier et Sylvain Joyal, 2004)

Biskri et al. (2004) nous ont proposé un outil semi-automatique pour le repérage des termes complexes. Il a adopté une méthode hybride qui combine les approches linguistiques et les filtres statistiques qui reposent sur l'approche bayésienne (Ibekwe-SanJuan, 2007 : 99) et sur le N-gramme de mots (Cavnar et Trenkle, 1994). Pour extraire les termes, certains linguistes informaticiens préconisent d'utiliser le filtre linguistique avant d'utiliser le filtre statistique pour les raisons suivantes : les méthodes statistiques sont sensibles à la taille du corpus ; la fréquence des termes est parfois erronée sans une lemmatisation préalable ; on risque de perdre les termes modifiés par un adverbe ou un adjectif et d'introduire trop de bruits si l'on utilise le filtre statistique en premier lieu. Cependant, certains privilégient l'utilisation du filtre statistique avant l'utilisation du filtre linguistique, car ils croient que l'étiquetage des termes prend trop de temps, que les règles syntaxiques n'arrivent pas à couvrir tous les cas et que le sens des mots varie très souvent selon le contexte.

La méthode de Biskri et al. (2004) combine un calcul bayésien avec les filtres statistiques et linguistiques. Dans leur méthode, le n-gramme de mots est défini par une suite de mots, telle qu'une suite de deux mots (bi-gram), de trois mots (tri-gram), etc. La probabilité du dernier mot de la chaîne n-gramme est définie comme la probabilité que n-gramme soit admis comme terme. La formule correspondante est la suivante :

$$P(W_{1\ldots n}) = \Pi_{1\ldots k}\, P(W_k|W_{1\ldots k-1}) \quad \text{(équation bayésienne générale)} \tag{14}$$

$$P\,(W_{1\ldots n}) \approx \Pi_{1\text{-}k}\, P(W_k|W_{k-1}) \quad \text{(équation bayésienne pour les bi-grams)} \tag{15}$$

Plus la probabilité d'un n-gramme est haute, plus il est possible de le considérer comme un terme complexe.

Les opérations concrètes sont les suivantes : le système permet l'extraction du lexique du corpus et laisse l'usager sélectionner les mots qui l'intéressent ; à partir des mots sélectionnés, on repère des candidats-termes complexes qui seront entrés dans le filtre semi-automatique

statistique et linguistique ; le filtre statistique élimine les candidats-termes dont la probabilité est inférieure à un certain seuil donné par l'usager ; le filtre linguistique élimine ensuite les candidats-termes qui commencent ou se terminent par un mot fonctionnel, un verbe, un adverbe, un adjectif ou une préposition et les candidats-termes qui commencent ou se terminent par les mots spécifiques sélectionnés par l'usager. La première étape consiste à construire un ensemble de mots pôles, ce qui nous permet de déterminer si les termes complexes examinés appartiennent au domaine indiqué par l'usager ou non. À l'aide des mots de pôles, on repère plus de termes complexes du corpus dans la deuxième étape. Par la suite, on procède aux filtres statistiques et linguistiques qui sont indépendants. L'ordre de l'application de ces filtres qui aboutit aux différents résultats peut être choisi par l'usager. Ces filtres nous permettent d'éviter trop de bruits et d'argumenter le taux de précision. Cependant, l'augmentation du taux de précision baisse le taux de rappel du système. Pour surmonter ce problème, on adopte une règle d'apprentissage : « *un terme candidat ne peut être éliminé par un* » (Biskri et al., 2004 : 6).

Pour évaluer le système, on a pris un article scientifique de 20 pages en français dont le sujet est le traitement automatique des langues. On a constaté que le filtre statistique baisse le taux de rappel et que le filtre linguistique augmente le taux de précision. Au cas où l'on fixe le seuil de probabilité à 0,001, on a récupéré 92 candidats-termes parmi lesquels on n'a validé que 10 candidats. Le taux de précision est de 11%. Au cas où l'on fixe le seuil de probabilité à 0,0005, le taux de précision baisse à 10%. On a accepté seulement 37 candidats sur 391 candidats-termes. Ensuite, on a fait deux évaluations sur les résultats obtenus par la combinaison des deux filtres. L'application du filtre linguistique après l'application du filtre statistique a un taux de précision de 80% avec un seuil d'acceptabilité de 0,001 et un taux de précision de 80% également avec un seuil d'acceptabilité de 0,0005. On en déduit que le seuil d'acceptabilité utilisé dans le filtre statistique doit être le plus bas possible pour obtenir un meilleur taux de rappel et que le filtre linguistique peut être amélioré par la règle d'apprentissage (Biskri et al., 2004 : 7).

B. Acabit : un outil d'extraction des termes complexe (S. Boulaknadel, B. Daille et D. Aboutajdine, 2008)

ACABIT est un outil de construction de terminologie semi-automatique qui propose les candidats-termes à valider à l'intention des terminologues. ACABIT se base sur les règles linguistiques morpho-syntaxiques et les modèles statistiques. Premièrement, on extrait automatiquement les candidats-termes du corpus étiqueté et lemmatisé en s'appuyant sur les descriptions linguistiques. Ensuite, on trie les candidats en calculant le coefficient de vraisemblance de chaque candidat-terme. Finalement, on fournit la liste ordonnée des candidats-termes de base à valider pour les terminologues. Bref, ACABIT est composé de deux étapes : l'extraction des candidats-termes du corpus pré-étiqueté à l'aide des règles linguistiques ; le filtrage statistique des candidats-termes selon le coefficient de vraisemblance. Dans la partie suivante, on va présenter plus concrètement chaque étape dans ACABIT.

Pour réaliser la première étape (l'extraction des candidats-termes), on a fait une analyse linguistique permettant d'extraire des structures morphosyntaxiques et leurs variantes à l'aide des règles morphologiques. En raison que la performance de l'approche statistique est dépendante de la quantité et de la qualité des échantillons, la construction d'un ensemble de termes d'échantillons de bonne qualité et en quantité suffisante est indispensable. Or, le nombre des termes binaires où uniquement les unités lexicales non fonctionnelles (telles que les noms, les adjectifs, les adverbes, etc.) sont considérées apparaît limité dans l'écriture. Ainsi, le système ACABIT se focalise sur l'extraction des termes binaires comme candidats-termes. Ces termes sont aussi appelés termes de base et leurs structures morphosyntaxiques sont les suivantes :

N+A : par ex., instruction publique

N1+Prep+N2 : par ex., principe d'égalité

N1+Prep+N2 : par ex., apprentissage de la lecture

N1+N2 : par ex., apprenti lecteur

N1+à+Vinf : par ex., savoir à enseigner

Leurs variations morphosyntaxiques peuvent être divisées en quatre types : variations graphiques, variations flexionnelles, variations morphosyntaxiques (variation de la préposition, présence ou absence d'un déterminant) et variations syntaxiques (insertion de modifieurs, coordination). On représente ces structures morphosyntaxiques et leurs variantes correspondantes par les grammaires locales dans le but d'extraire les candidats-termes du corpus.

Dans le deuxième temps, on utilise les différentes mesures statistiques pour évaluer les candidats-termes et essayer de trouver la meilleure mesure. Le système permet de trier les candidats-termes selon le score statistique assigné à chaque candidat et fournit en sortie une liste de couples ordonnée. Ces mesures ont pour objet de calculer le statut terminologique de la séquence rencontrée. Chaque mesure statistique se base sur un classement conceptuel des couples qui met en avant des expressions figées que des termes du domaine. On évalue les résultats obtenus par chaque mesure en faisant appel à une liste de référence des termes du domaine donné. On n'évalue que les 100 premiers couples extraits du corpus. Si le couple est repéré dans la liste de référence, il est validé comme un terme complexe ; sinon, on cherche sa traduction compositionnelle en se référant à la banque terminologique Eurodicautom.

Le système ACABIT a été appliqué à un corpus composé de textes extraits du web dans le domaine de l'environnement restreint aux thématiques suivantes : la pollution, la purification de l'eau, la dégradation du sol, la préservation de la forêt et les catastrophes naturelles. Ensuite, on a évalué les résultats obtenus du système par la mesure IM3 (Daille, 1994), T-score (Dunning, 1994), LLR (Dunning, 1994) et FLR (Nakagawa & Mori, 2003). Le tableau suivant liste les types de mesure et les taux de précision obtenu par chaque mesure :

Type	P(%)
FLR	60%
T-score	57%
LLR	85%
IM3	26%

Tableau 3 Types de mesure et taux de précision

Le taux de précision des résultats obtenus par la mesure LLR est le plus élevé. LLR permet de bien cerner le potentiel terminologique de certains des candidats-termes extraits du corpus.

C. Validation des relations de dépendance par la cooccurrence sur internet : présentation et critique (Thomas Lebarbé, 2008)

Lebarbé (2008) a développé un agent qui consiste à valider la structure des syntagmes nominaux selon les cooccurrences. Cet agent nous aide à décider si une structure comprenant plusieurs mots est un terme composé ou non. Le principe de cet agent s'appuie sur la notion de blocs d'Abney (1991). La délimitation de ces blocs est décidée par les règles de déduction contextuelle, les mots-outils et les terminaisons morphologiques. L'objectif d'analyses de Lebarbé est de développer un système d'analyse syntaxique automatique permettant d'insérer les autres modules qui sont alliés pour calculer si une structure syntaxique est la plus proche de celle qui est attendue. Le système peut percevoir la présence des voisins des unités linguistiques (initialement les mots), identifier l'état de ces mots en fonction d'une base de règles. Ensuite, en prenant en compte et en appliquant la perception des informations ci-dessus, le système associe les voisins à chaque unité linguistique initiale pour former une unité linguistique du niveau hiérarchique supérieur. Cette procédure est faite itérativement dans le but de reconnaître les syntagmes nominaux. Dans la partie suivante, on va présenter les opérations de cette méthodologie plus concrètement.

Le principe de la méthode de Thomas Lebarbé se base sur les cooccurrences. Par exemple, dans une chaîne de type N+pN1+pN2, nous ne pouvons pas calculer avec certitude si pN2 dépend de pN1 ou de N, mais nous pouvons émettre l'hypothèse que le plus fréquent des motifs entre N+pN2 et pN1+pN2, sur un corpus suffisamment grand, est le plus probable des schémas relationnels (pN réfère au bloc prépositionnel).

L'outil qu'on utilise est le moteur de recherche Google qui permet des requêtes avancées (http://www/google.fr/advanced_serach?hl=fr). Les requêtes sont exécutées dans une langue donnée sur un ensemble de termes coprésents dans un document, dans un domaine d'un réseau ou de tout le réseau en fonction d'une date de mise en ligne du document. Par exemple,

les arguments <<hl=fr>> signifient une requête sur les documents en français ; l'expression exacte à rechercher dans <<as_epa=%22expression à rechercher%22>>est indiquée entre les deux « % 22 » correspondant aux guillemets.

Par le moteur de recherche sur Internet, nous avons pu observer que :

- {demande} et {attestation} sont cooccurrents dans 43900 documents
- {arrestation} et {juge} sont cooccurrents dans 29000 documents
- {demande} et {juge} sont cooccurrents dans 202000 documents
- la chaîne exacte {demande d'arrestation} apparaît dans246 documents
- la chaîne exacte {arrestation du juge} apparaît dans 45 documents
- la chaîne exacte {demande du juge} apparaît dans 928 documents

On peut constater que {demande} et {juge} sont plus liés que {arrestation} et {juge}. Les relations de dépendance respectent les deux règles suivantes pour le français : un syntagme prépositionnel dépend d'un syntagme qui le précède ; les liens entre les syntagmes ne se croisent pas, par exemple, parmi les 4 syntagmes, si le syntagme A est relié au syntagme C, le syntagme D ne peut pas être lié au syntagme B. Cependant, lorsque les syntagmes sont distants, l'opération pour relier les syntagmes alourdit la structure et la rend plus difficile à interpréter. Ainsi, on prend en compte des distances entre blocs (par ex., {demande} et {juge}) en pondérant les valeurs retournées par le moteur de recherche. La combinatoire dans le cas de 4 syntagmes est présentée dans le schéma suivant :

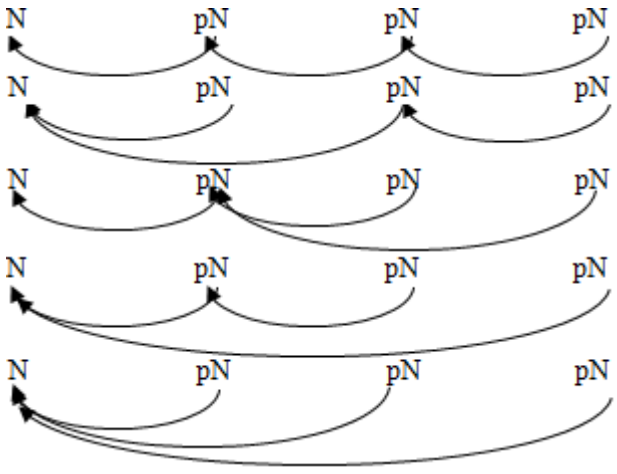

Figure 6 Combinatoire dans le cas de 4 syntagmes

Si l'utilisation de ces distances et coefficients de pondérations ne change rien au résultat, la dépendance entre {la demande} et du {juge} est validée.

Le taux de précision et de rappel de la méthode varie selon les corpus des différentes origines. De plus, il existe aussi quelques autres problématiques à étudier en ce qui concerne ce système. Premièrement, la fréquence de deux blocs est augmentée potentiellement selon la taille du document et la distance entre deux blocs. Deuxièmement, se baser sur l'existence d'une chaîne exacte pour déterminer la dépendance de deux blocs ne constitue pas encore une preuve empirique ou théorique. Le dernier point à critiquer du système est que le temps de récupération de la page correspondante à la requête peut beaucoup varier en fonction de l'encombrement du réseau. Le temps maximum de récupération dépasse la seconde et demie. Pour un système automatique, c'est un problème intolérable.

D. Parsing Models for identifying Multiword Expressions (Spence Green, 2012)

Les modèles Parser de Green (2012) sont d'autres méthodes hybrides pour identifier automatiquement les mots composés. Spence Green et son équipe ont développé deux modèles Parser : l'un se base sur les grammaires de contexte libre (CFG) et l'autre s'appuie sur l'arbre de grammaires de substitution (TSG) qui peut stocker les segments syntaxiques plus larges. Traditionnellement, on utilise la classification n-gramme pour reconnaître les mots composés. La classification n-gramme est une méthode basée sur la cooccurrence. Cependant, la méthode statistique basée sur la cooccurrence produit souvent des erreurs si le corpus ne couvre pas assez d'informations de cooccurrences. Par exemple, si la séquence multi-mots *partie du discours* est moins fréquente par rapport à la séquence *parties du discours*, on va probablement considérer à tort que seule la dernière est une des termes de multi-mots. Les deux modèles développés par Spence Green sont plus efficaces que la méthode n-gramme. Dans la partie suivante, on les présente en détail.

Le modèle Parser de contexte libre est un modèle génératif dans lequel les caractéristiques de grammaire sont composées de segments de balises dans les données

entraînées. La catégorie de segments de balises pour le français peut correspondre à un type de syntagmes adverbials, un type de noms composés, une partie du discours, une marque de préposition, des ponctuations, etc. L'annotation d'une partie du discours fournit une information sur le contexte externe. Les marques de prépositions contiennent les différentes caractéristiques qui marquent le contexte de prépositions, par exemple, la marque *P* sert à identifier les prépositions qui introduisent SPs modifiant un nom.

On a établi une liste de caractéristiques lexicales (telles que l'information morphologique, l'information de parties du discours, etc.) pour générer l'identification des types de mots rares ou non connus. On note la balise de caractéristiques lexicales par *t* et l'identification des mots par *s*. Le paramètre de signature *p(s|t)* est calculé par l'algorithme bayésien. L'idée est que l'on parcourt les termes par une représentation factorielle dans laquelle les facteurs peuvent être la forme d'un mot, le lemme d'un mot ou les caractéristiques grammaticales d'un mot, telles que le genre, le nombre, la personne, etc. Le Parser lexical factoriel sert à évaluer la probabilité de génération d'un mot par les balises du mot annotées dans la donnée entraînée. On définit le paramètre *p(t|s)* pour l'identification d'un mot non connu dans le corpus entraîné *c(w)* selon la formule suivante :

$$p(t|w) = \frac{c(t,w)}{c(w)} \tag{16}$$

$$p(t|w) = \frac{c(t,w)+\alpha(t|s)}{c(w)+\alpha} \tag{17}$$

Ensuite, on calcule le paramètre *p(w|t)* de la façon suivante :

$$p(w|t) = \begin{cases} \frac{p(t|w)p(w)}{p(t)} & if\ c(w) > \beta \quad (18) \\ \frac{p(t|w)p(w)}{p(t)} & if\ c(w) > 0 \quad (19) \\ \frac{p(t|s)p(s)}{p(t)} & else \quad (20) \end{cases}$$

Dans le Parser factoriel, chaque token a une analyse morphologique associée *m*. Le modèle génère le mot et l'analyse morphologique en fonction de la formule suivante :

$$p(w, m|t) = p(w|t)p(m|t) \quad (21)$$

Même si le type du mot *w* est inconnu, l'analyse morphologique associée *m* est toujours connue. Ainsi, la probabilité de toutes les balises nominales sont assignées zéro, car les noms ne portent jamais le temps. L'inconvénient du modèle CFG (grammaires de contexte libre) est qu'il ne détecte pas les expressions idiomatiques.

Le deuxième modèle est celui d'un arbre de grammaires de substitution qui peut servir à stocker les segments d'arbre lexicalisés comme règles. On peut stocker tout le segment lexicalisé dans la grammaire. Le DP-TSG peut être considéré comme un modèle de parser orienté par des données (DOP) avec l'évaluation de paramètre bayésienne. Un modèle P-TSG est un quintuple ($V, \Sigma, R, \blacklozenge, \theta$) où $c \in V$ sont les non-terminaux, $t \in \Sigma$ sont les non terminaux ; $e \in R$ sont les arbres élémentaires, $\blacklozenge \in V$ est un symbole de commencement unique et $c,e \in \theta$ sont les paramètres pour chaque segment d'arbre. L'extraction de la grammaire dans DP-TSG se réduit à une problématique de segmentation en utilisant une banque d'arbres *T* que l'on segmente en séries *R*. C'est un processus modélisé par le bayésien naïf :

$$p(R|T) \propto p(T|R)p(R) \quad (22)$$

$p(T \mid R)$ peut égaler à soit 0 soit 1. Le DP-TSG contient un DP prioritaire pour chaque $c \in V$ et génère un segment d'arbre *e* issu du non terminal *c* selon la formule suivante :

$$\begin{aligned} &\theta_c|c, \alpha_c, P_0(\cdot|c) \sim DP\ (\alpha_c, P_0) \\ &e|\theta_c \sim \theta_c \end{aligned} \quad (23)$$

Différentes des méthodes n-gramme classification, les méthodes de parser fournissent une sous-catégorisation syntaxique et n'exige pas un pré-filtrage heuristique des données entraînées. Les deux modèles de parser améliorent l'identification des mots composés par rapport aux méthodes n-gramme. L'amélioration par rapport aux méthodes n-gramme peut atteindre 50% pour le français. Les modèles de parser peuvent être appliqués à n'importe quelle langue pour laquelle les ressources linguistiques suivantes existent : une banque d'arbre syntaxique, un petit ensemble de termes complexes et un analyseur morphologique.

E. La reconnaissance des mots composés à l'épreuve de l'analyse syntaxique et vice-versa : évaluation de deux stratégies discriminantes (Matthieu Constant, 2012)

Constant (2012) a présenté et évalué deux stratégies discriminantes d'intégration du traitement des mots composés dans le processus d'une analyse syntaxique probabiliste : la première stratégie s'appuie sur la pré-segmentation lexicale au moyen d'un reconnaisseur de connaissances des mots composés à l'aide de l'application des champs markoviens aléatoires ; la deuxième stratégie est fondée sur une analyse à partir d'une grammaire incluant l'identification des mots composés, suivie d'une phase de réordonnancement des analyses à l'aide d'un modèle de maximum d'entropie intégrant des traits dédiés aux mots composés.

Dans la première stratégie, l'identification des mots composés est considérée comme une tâche d'annotation séquentielle selon le schéma IOB (Ramshaw et Marcus, 1995). Cependant, ce schéma a une limitation théorique : les mots composés doivent être continus. Green (2011) a proposé un schéma permettant d'intégrer les mots composés dans la grammaire et autorise des unités polylexicales discontinues. Dans la première stratégie, on propose d'associer les composants simples des unités polylexicales à une étiquette de la forme CAT+X où CAT est la catégorie grammaticale du mot composé et X détermine la position relative du token dans le mot composé (soit B pour le début -*Beginning*- et I pour les autres positions internes -*Inside*-). Les mots simples sont étiquetés par la lettre O, par exemple, *Jean/O adore/O les/O faits/N+B divers/N+I*. Ce processus d'annotation s'appuie sur le modèle des champs aléatoires markoviens linéaires (Tellier et Tommasi, 2011) [CRF] introduits par (Lafferty et al., 2001). Le modèle est défini par la formule suivante :

$$P\lambda(y|x) = \frac{1}{\mathrm{z(x)}} \cdot \sum_t^N \sum_k^K log\ \lambda_k \cdot f_k(t, y_t, y_{t-1}, x) \qquad (24)$$

dans laquelle $x=(x_1,x_2,\ldots,x_N)$ est une séquence de mots en entrée et $y=(y_1,y_2,\ldots y_N)$ est un ensemble d'étiquettes en sortie. $Z(x)$ est un vecteur de normalisation dépendant de x. Il est basé sur K traits qui sont définis par des fonctions binaires f_k. La fonction t marque la position

courante dans x et y_t indique la position courante dans y. La fonction f_k est déterminée par les paramètres : t, y_t, y_{t-1}, et x. Si une configuration particulière entre ces quatre paramètres est satisfaite (par ex., $f_k(t,y_t,y_{t-1},x)=1$), les traits sont activés. Chaque trait est associé à un poids λ_k.

Le deuxième modèle se base sur une grammaire incluant l'identification des mots composés, suivie d'une phase de réordonnancement des analyses. Le réordonnancement discriminant consiste à reclasser les n meilleures analyses et à sélectionner la meilleure. Les n meilleurs analyses sont produites par un analyseur syntaxique de base qui s'appuie sur un modèle de maximum d'entropie intégrant des traits dédiés aux mots composés. Plus précisément, pour une phrase, on a un ensemble de analyses candidates $P(s)= (p_1, p_2, p_3, \ldots p_i)$. L'analyseur de base fournit une probabilité pour chaque analyse candidate p à l'aide du modèle de maximum d'entropie. Pour chaque analyse candidate, on a défini un vecteur de traits f parmi lesquels le premier est la probabilité de p et un vecteur de paramètres θ qui est évalué lors des séances d'apprentissage du corpus de référence et lors des analyses générées par l'analyseur de base. Ainsi, pour chaque analyse candidate, on obtient une fonction de score V_θ :

$$V_\theta(p) = \theta \cdot f(p) = \sum_{j=1}^{m} \theta_j \cdot f_j(p) \quad (25)$$

Avec le score V_θ, on sélectionne la meilleur analyse p^* parmi les candidates selon la formule suivante :

$$p^* = argmax_{p \epsilon P(s)} V_\theta(p) \quad (26)$$

En ce qui concerne les traits dans les deux stratégies, on les génère à partir des patrons. Par exemple, on a défini un patron T dans la stratégie d'annotation séquentielle : $T=f(x,n)/y_n$ à chaque position n dans la séquence en entrée x et un patron T dans la stratégie de réordonnancement : $T=f(p,n)/label(m)/pos(p,n)$ à chaque feuille (à la position n) dominée par un nœud de type EMM m dans candidate analyse p. Dans cette définition, f est une fonction à définir ; y_n est l'étiquette du mot à l'indice n ; $label$ (m) est l'étiquette du nœud m et $pos(p,n)$ indique la position relative du mot par rapport à la position n dans l'unité polylexicale: B(position initiale), I(autres positions). Les traits peuvent être divisés en deux types : les traits

endogènes qui sont extraits directement des mots eux-mêmes et les traits exogènes qui proviennent totalement ou partiellement des données externes. Par exemple, on considère *coup de* comme un trait de noms composés, car il est souvent le préfixe de séquences de multi-mots, telles que *coup de pied, coup de foudre, coup de main*, etc. La séquence *coup de* est un trait endogène. Un autre exemple, si l'on étiquette chaque mot du corpus par sa partie du discours ou son pointage d'association (log-vraisemblance), les informations de ces étiquettes sont les traits exogènes.

Les deux stratégies décrites ci-dessus ont été appliquées et les expérimentations ont été évaluées. On a utilisé la F-mesure [F] pour évaluer l'efficacité de l'identification automatique des noms composés. Sur la première stratégie, le F-mesure de reconnaisseur qui intègre tous les traits s'élève à 75.9%. « Il est, en pratique, meilleur que celui proposé par (Green et al., 2011) qui a une F(EMM) de 71.1% sur les phrases inférieures à 40 mots[…] » (Constant, 2012 : 10). Le reconnaisseur de la première stratégie a atteint 74% pour les traits endogènes et 77.3% pour tous les traits.

Sur la deuxième stratégie qui intègre un outil de réordonnancement dans l'analyse syntaxique, on a utilisé plusieurs mesures d'évaluation : la F-mesure [F], la mesure UAS (Unlabeled Attachment Score) (Tsarfaty, Nivre et Andersson, 2011) et la mesure LA (Leaf Ancestors) (Sampson, 2000). Néanmoins, le résultat est moins satisfaisant par rapport à la première stratégie. La F-mesure de réordonnancement a une possibilité d'atteindre 76.9% au maximum et elle est plus élevée pour les traits généraux par rapport à tous les traits dédiés aux mots composés (81.98% vs. 81.64%). Pour la mesure LA, le résultat pour les traits généraux et les traits dédiés aux composés est de 93.41% contre 93.12%. Cependant, les résultats se dégradent pour la mesure UAS (84.40% vs. 84.98). Finalement, on n'a pas constaté une variation évidente en ce qui concerne les résultats pour les différents types de traits de n'importe quelle mesure d'évaluation.

Chapitre 3 Méthodologie

1. Modèle de données

« Trois fonctions primaires » est le modèle de données de références de la fonction prédicative, la fonction actualisatrice et la fonction argumentale. L'objectif de ce modèle est de décrire tous les emplois des unités lexicales d'une langue. Le postulat de ce modèle de données suppose que les phrases d'une langue sont constituées d'unités lexicales correspondant à des prédicats, des arguments et des actualisateurs. Le processus d'analyse des trois fonctions primaires consiste à construire la représentation des énoncés avec des éléments métalinguistiques qui sont reliés entre eux d'une façon structurée. Dans cette partie, nous présentons d'abord les notions de prédicat, d'argument et d'actualisateur. Ensuite, nous discutons d'une typologie de prédicats et nous présentons à la fin la relation d'appropriation.

1.1. Prédicats, arguments et actualisateurs

Le prédicat est une unité linguistique définie comme la forme langagière d'une relation entre deux entités. Les entités liées par cette relation sont les arguments (Gross G., 2012 :13). Par exemple, dans les phrases suivantes :

4) *Le président reçoit le premier ministre*.
5) *Paul a acheté une nouvelle voiture.*
6) *Les paysans ont jeté les pommes gâtées.*

recevoir, acheter et *jeter* sont tous des prédicats. *recevoir* indique la relation entre les deux entités *président* et *premier ministre, acheter* désigne la relation entre *Paul* et *voiture,* et *jeter* lie les deux entités *pommes* et *paysans*. *président* et *premier ministre* sont les deux arguments du prédicat *recevoir, Paul* et *voiture* sont les arguments d'*acheter,* et *pommes* ainsi que

paysans sont les deux arguments du prédicat *jeter*. La combinaison entre un prédicat et un argument est subordonnée à la sémantique du prédicat et la sémantique de l'argument. Par exemple, pour le prédicat *fabriquer*, son argument à la position du complément doit obligatoirement être de la classe sémantique de noms d'artefacts ; en revanche, les arguments qui sont de la classe sémantique de noms d'artefacts exigent de se combiner avec des prédicats qui décrivent des états, des actions ou des évènements associés aux caractéristiques sémantiques des noms d'artefacts, par exemple, pour le nom *ordinateur*, il ne peut être combiné qu'avec certains prédicats :

7) *fabriquer des ordinateurs*
8) *vendre des ordianteurs*
9) **manger des ordinateurs*
10) **élever des ordinateurs*

Les actualisateurs sont des éléments linguistiques qui permettent d'inscrire les prédicats et les arguments dans des énoncés afin d'obtenir une phrase grammaticalement correcte. Ils peuvent être les unités grammaticales, telles que les prépositions (par ex., *à, de,* …), les déterminants (par ex., *un, une, le, la*, …),..... ou les unités lexicales, telles que les verbes auxiliaires, les verbes supports, etc. Par exemple, dans la phrase

11) *Marie a fait couper ses cheveux.*

le verbe *faire* est le verbe support du verbe *couper*; dans la phrase

12) *Le directeur a démissionné.*

le verbe *a* est un verbe auxiliaire qui véhicule la valeur temporelle et aspectuelle du prédicat *démissionner*. *faire* et *avoir* dans les exemples cités sont tous les actualisateurs. L'actualisation des prédicats comprend à la fois des indications de nature temporelle (conjugaison) et des informations de nature aspectuelle. Les actualisateurs sont subordonnés soit directement aux prédicats, soit aux relations entre les prédicats et leurs arguments respectifs (Buvet, 2009a). Si les actualisateurs se rapportent aux prédicats, les occurrences des actualisateurs dépendent de celles des prédicats ; si les actualisateurs sont subordonnés aux

relations entre les prédicats et leurs arguments respectifs, les occurrences des actualisateurs dépendent des relations entre les prédicats et leurs arguments.

1.2. Typologie des prédicats

Du point de vue des emplois des prédicats, on peut avoir les prédicats verbaux, les prédicats nominaux, les prédicats adjectivaux, les prédicats prépositionnels et les prédicats adverbiaux. D'après Buvet (2009a : 89) : « les emplois prédicatifs sont définis du point de vue de leur particularité structurale. Les relations entre les prédicats et les arguments sont des relations de dépendance hiérarchisées : les prédicats prédominent les arguments et les sous catégorisent. […] Les emplois prédicatifs sont les différentes instances des prédicats dans les phrases. ». Les variations entre les emplois prédicatifs sont de nature morphosyntaxique ou interprétative. Un emploi prédicatif a comme racine son lemme. Les emplois prédicatifs qui se rapportent à un même prédicat malgré leurs différences morphosyntaxiques partagent la même racine. Un prédicat peut avoir un ou plusieurs emplois, par exemple, le verbe *se confier*, l'adjectif *confiant* et le nom *confiance* sont trois emplois du même prédicat dans :

13) *Marie se confie à sa mère.*
14) *Marie est confiante en sa mère.*
15) *Marie a confiance en sa mère.*

se confier est l'emploi verbal, *confiant* est l'emploi adjectival et *confiance* est l'emploi nominale. Les interprétations des emplois prédicatifs résultent de l'ensemble des propriétés des catégories : type d'état, type d'action, aspect processif et aspect statif. Un prédicat peut décrire un état, une action ou un évènement. Par exemple, dans la phrase

16) *Luc est confiant dans le succès.*

Le prédicat *confiant* décrit un état dans la phrase

17) *Le train est parti.*

partir est un prédicat d'action ; dans la phrase

18) Il y a eu une négociation hier.

négociation est un prédicat qui décrit un évènement.

Selon le type de relation prédicative, on classe les prédicats en prédicats monadiques, prédicats dyadiques et prédicats triadiques. Le prédicat du type monadique n'a qu'un argument et il indique une relation prédicative unaire, par exemple, dans

19) Le professeur est parti.

le prédicat *partir* indique une relation prédicative unaire et il n'a qu'un argument *professeur*. Le prédicat dyadique a deux arguments et il désigne une relation binaire, par exemple, dans la phrase

20) Paul a coupé l'arbre.

couper est un prédicat dyadique qui indique une relation binaire entre ses deux arguments *Paul* et *arbre*. Le prédicat triadique n'a également qu'un argument, mais il présente une relation internaire, par exemple, dans

21) Luc est courageux.

Luc est le seul argument du prédicat *courageux* qui décrit une propriété interne de *Luc* et il indique une relation internaire.

En fonction des classes sémantiques d'arguments des prédicats, les prédicats peuvent aussi être divisés en prédicats basiques et prédicats appropriés. Une classe sémantique est un ensemble de substantifs qui sont homogènes sémantiquement. Les verbes basiques ont un spectre lexical large. Ces verbes peuvent souvent être suivis des lexiques de presque toutes les classes sémantiques, par exemple, *donner, prendre, amener,* etc. Le prédicat approprié est spécifique à une certaine classe sémantique d'arguments, à savoir, ce genre de prédicats peut uniquement être suivi d'un groupe d'arguments appartenant à une classe sémantique particulière. Par exemple, pour le prédicat *boire,* son deuxième argument qui le suit ne peut être que les noms indiquant du liquide potable :

22) boire du vin/de l'eau/du jus d'orange

*23) *boire des biscuits/une table/une voiture*

pour le prédicat *enceinte*, son argument ne peut être qu'un être humain de sexe féminin :

24) Jeanne est enceinte.

*25) *Une vache est enceinte.*

1.3. Prédicats appropriés et relation d'appropriation

Les prédicats appropriés ont un nombre de classes sémantiques d'arguments relativement contraint. Ce caractère des prédicats appropriés nous permet de prédire la classe sémantique à laquelle leurs arguments font partie. Un prédicat approprié dans un sens spécifique peut définir une classe sémantique d'arguments. (Gross G., 2012 :73) La relation entre le prédicat approprié et ses arguments est la relation d'appropriation. La relation d'appropriation repose sur une observation de Zellig Harris. Selon Harris (2007), les combinaisons de certaines unités lexicales sont plus probables dans une phrase que d'autres.

Néanmoins, la polysémie de la plupart des prédicats nous exige de délimiter la classe d'arguments à l'aide de la conjonction de plusieurs prédicats appropriés à cette classe sémantique. Par exemple, pour le prédicat approprié *conduire*, on peut avoir des séquences suivantes en fonction de ses différents sens lexicaux :

26) conduire mon enfant à l'école (dans le sens de « transporter quelqu'un »)

27) conduire une voiture (dans le sens d' «assurer la direction d'un véhicule »)

28) conduire une entreprise (dans le sens de « diriger un groupe »)

La polysémie de *conduire* l'empêche d'isoler correctement la classe des moyens de transport routier. Néanmoins, avec un autre prédicat approprié de la classe sémantique de transport routier *garer*, on peut prédire que les arguments qui apparaissent à la fois après *conduire* et *garer* appartiennent à la classe sémantique des moyens de transport. Par exemple, dans

29) Le grand-père conduit sa voiture.

30) Le grand-père a garé sa voiture.

31) Je conduis mon enfant à l'école.

*32) *J'ai garé mon enfant.*

L'argument *voiture* peut apparaître à la fois après le prédicat *conduire* et *garer*, alors que l'argument *enfant* ne peut être utilisé qu'avec le prédicat *conduire* mais pas *garer*. Dans le traitement automatique, une classe sémantique d'arguments est délimitée à l'aide de la conjonction de plusieurs prédicats appropriés de cette classe. Cet ensemble de prédicats appropriés sont définis comme les prédicats appropriés définitionnels de cette classe d'arguments. Parmi ces prédicats appropriés définitionnels, on distingue les prédicats les plus représentatifs de la classe sémantique et les autres prédicats qui sont considérés comme périphériques. Cependant, les prédicats d'une même classe sémantique ont un domaine d'arguments commun. Les prédicats appropriés définitionnels caractérisent la sémantique de leur classe d'arguments.

2. Technologie : Unitex

L'exploitation des caractéristiques méthodologiques des trois fonctions primaires et l'exploitation des analyses de la structure interne des unités lexicales sont tous réalisées à l'aide des grammaires locales sur la plateforme Unitex. Unitex est un outil qui repose sur des technologies de type « automates à états finis» en intégrant de larges ressources linguistiques. Ces ressources linguistiques sont sous la forme de dictionnaires électroniques, de grammaires et de tables de lexique-grammaire. Elles sont issues des travaux initiés sur le français par Gross M. (1989) et qui ont été étendus à d'autres langues à travers du réseau de laboratoires RELEX. Unitex permet de construire, vérifier et appliquer les dictionnaires électroniques. Il permet de faire la correspondance avec les expressions régulières et de créer des transducteurs qui convertissent les entrées en représentant des grammaires locales sous la forme d'un automate à états finis.

Les dictionnaires électroniques dans Unitex recensent les unités monolexicales et les unités polylexicales en leur associant un lemme ainsi qu'une série de codes grammaticaux,

sémantiques et flexionnels. Deux sortes de dictionnaires électroniques sont distingués : le dictionnaire de formes fléchies, appelé DELAF (DELA de formes fléchies) ou DELACF (DELA de formes composées fléchies pour décrire les noms composés); le dictionnaire de formes non fléchies appelé DELAS (DELA de formes simples) ou DELAC (DELA de formes composées). Une entrée d'un DELAF respecte le schéma suivant : *Forme fléchie, Forme canonique.CodeGrammatical+CodeSémantique : CodeGenreCodeNombre*. Par exemple, dans l'entrée

33) mercantiles,mercantile.A+z1:mp:fp

mercantiles est la forme fléchie de l'entrée et *mercantile* est la forme canonique. La forme canonique est séparée de la forme fléchie par une virgule. Si la forme fléchie et la forme canonique sont pareilles, la forme canonique peut être omise alors que la forme fléchie est toujours obligatoire. Le code grammatical *A* indique la catégorie grammaticale de *mercantiles* qui est un adjectif et *z1* est un code sémantique qui signifie langage courant. *m* et *f* sont les codes sur le genre de l'unité lexicale (masculin ou féminin), et *s* ainsi que *p* désignent le nombre de *mercantiles* (simple ou pluriel). Certains mots composés comme *grand-mère* peuvent s'écrire avec des espaces, des tirets ou des caractères "=" qui seront remplacés par des espaces ou des tirets au cours de la compression du dictionnaire. Le format des DELAS est similaire à celui des DELAF. La différence est qu'on ne mentionne qu'une forme canonique suivie de codes grammaticaux et/ou sémantiques qui seront interprétés par le programme de flexion comme le nom de la grammaire à utiliser pour fléchir l'entrée, par exemple, dans l'entrée

34) cheval,N4+Anl

N4 signifie que l'unité lexicale *cheval* doit être fléchie avec une grammaire de flexion nommée N4. Les grammaires de flexion sont enregistrées sous forme de graphes (cf. Figure 7) dans Unitex et il est possible d'ajouter les nouveaux codes de flexion.

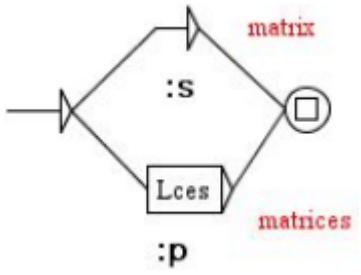

Figure 7 Grammaire de flexion

Les grammaires locales sont les grammaires régulières utilisées par l'analyseur syntaxique qui constituent la classe des grammaires les plus simples dans la hiérarchie de Chomsky (Maurel, 1992 : 150). Chomsky (1969) a présenté un modèle de grammaire construite à l'aide d'arbres syntaxiques et a proposé une théorie transformationnelle exploitant des arbres syntaxiques, à savoir un ensemble de règles permettant la transformation d'une structure d'arbre à une autre. Schématiquement, la forme d'un automate à états finis est un graphe dans lesquels les sommets sont appelés états et les flèches représentent les transitions. Par exemple, un ensemble de phrases en langue française *Paul aime des chocolats/Paul aime bien des chocolats/Paul aime des chocolats noirs* peuvent être représentée par la grammaire locale sous forme d'un automate à états finis comme ce qui est montré dans Figure 8.

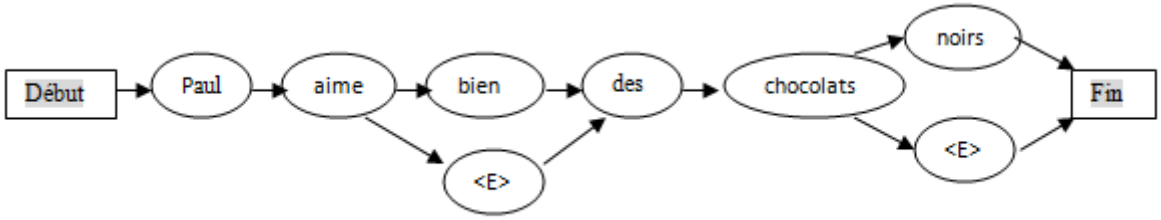

Figure 8 Automate à états finis

<E> signifie vide. Un transducteur fini est un automate à états finis avec sorties. Il opère sur les entrées textuelles. Au lieu de seulement accepter ou refuser les séquences textuelles entrées selon les grammaires locales, il transforme aussi les textes acceptés et génère une représentation en sortie. Les sorties peuvent être stockées dans les variables qui sont utilisées dans les graphes. Par exemple, dans la Figure 9, les noms de <titre> sont tous enregistrés dans la variable *title* et les sorties sont *[TITLE*=nom de <titre> reconnu*]*.

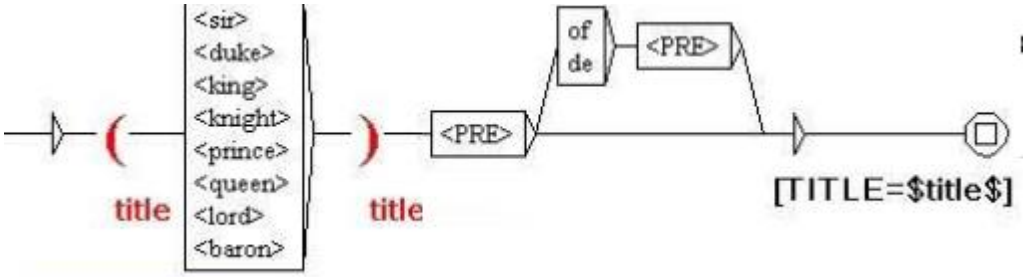

Figure 9 Transducteur à états finis

3. Ressources : Morfetik

Morfetik est un système modulaire incluant un moteur de flexion, des interfaces de consultation et d'interrogation et des outils d'exploitation. La ressource lexicale qui sert de base à Morfetik est un dictionnaire morphologique du français élaborée sous la direction de Michel Mathieu-Colas (2009). Il est un ensemble de données structurées sous forme de tables. Il recense des unités monolexicales du français en leur associant des informations morphologiques et flexionnelles. La ressource lexicale qu'on exploite pour le projet de la thèse est le dictionnaire morphologique Morfetik. Dans ce qui suit, on fait une présentation plus détaillée des données linguistiques de Morfetik et nous présentons également le moyen d'exploitation de Morfetik dans Unitex.

3.1. Présentation de Morfetik

Le recensement lexical de Morfetik a fait appel à diverses sources lexicographiques, tels que le DELAS (Dictionnaire électronique du LADL, cf. COURTOIS 1990), le Petit et le Grand Robert, le Petit Larousse illustré, le Lexis, le Grand Larousse encyclopédique et le Grand Dictionnaire encyclopédique Larousse (GDEL), le Trésor de la langue française, le Harrap's et le Robert & Collins, des dictionnaires d'argot, des tables de conjugaison (dont le Bescherelle (1842) et les Verbes logiques de A. DUGAS (1996)), le Bon Usage de GREVISSE et des dictionnaires de « difficultés » pour le traitement des cas problématiques. Morfetik comprend 106 884 unités monolexicales dont 69950 sont noms, 24405 sont adjectifs, 10232 sont verbes, 1894 sont adverbes, 200 sont interjections, 68 sont pronoms, 59 sont déterminants, 58 sont prépositions et 18 sont conjonctions.

Cinq groupes d'unités lexicales différents sont mis en place en fonction des catégories morphosyntaxiques : les noms, les adjectifs, les déterminants et pronoms, les verbes, et les mots invariables. Les mots invariables comprennent les prépositions, les adverbes, les conjonctions et les interjections. Les déterminants et pronoms, tout comme les mots invariables, nécessitent uniquement un listage. Cependant, pour les noms, les adjectifs et les verbes qui sont des catégories à flexion, il nous faut établir deux tables : une table de lemmes avec la catégorie grammaticale et les codes de flexion associés ; une table de flexions qui enregistre les informations flexionnelles à chaque lemme. Il y a au total 222 codes de flexion. La table de flexions comprend les champs suivants : un identifiant de code, un exemple de modèle, un radical (nombre de caractères à soustraire de la forme canonique), un radical-modèle, désinences de l'infinitif, désinences des formes conjuguées (45 champs), et désinences des participes (5 champs). La Figure 10 est la capture d'écran d'une partie de la table de flexions. De plus, on établit également une table de formes qui enregistre toutes les formes flexionnelles correspondant à chaque entrée. Dans cette table, chaque ligne enregistre une forme flexionnelle d'une unité lexicale en lui associant un identifiant, l'identifiant du lemme correspondant, le lemme, la catégorie grammaticale, le genre, le nombre, la personne et le temps. La Figure 11 est la capture d'écran d'une partie de la table de formes.

```
(`Code`, `Modele`, `Rad`, `R-modele`, `Inf::`, `Ind-pr:1:S`, `Ind-pr:2:S`, `Ind-pr:3:S`, `In
('001', 'aimer', 2, 'aim', 'er', 'e', 'es', 'e', 'ons', 'ez', 'ent', 'ais', 'ais', 'ait', 'i
('001.1', 'conster', 2, 'const', 'er', '-', '-', 'e', '-', '-', '-', '-', '-', '-', '-', '-'
('001.2', 'patafioler', 2, 'patafiol', 'er°', '-', '-', '-', '-', '-', '-', '-', '-', '-', '
('001.3', 'importer', 2, 'import', 'er', '-', '-', 'e', '-', '-', 'ent', '-', '-', 'ait', '-
('001.4', 'puer', 2, 'pu', 'er', 'e', 'es', 'e', 'ons', 'ez', 'ent', 'ais', 'ais', 'ait', 'i
('001.5', 'béer', 2, 'bé', 'er', 'e', 'es', 'e', 'ons', 'ez', 'ent', 'ais', 'ais', 'ait', 'i
('001.6', 'dinguer', 2, 'dingu', 'er', 'e°', 'es°', 'e°', 'ons°', 'ez°', 'ent°', 'ais°', 'ai
```

Figure 10 Table de flexions dans Morfetik

```
INSERT INTO `formes` (`forme_id`, `forme`, `lemme_id`, `lemme`, `catgram`, `gende
(1, 'abaisser', 'V1', 'abaisser', 'VRB', NULL, '', '', 'Inf', 0),
(2, 'abaisse', 'V1', 'abaisser', 'VRB', NULL, 'S', '1', 'Ind-pr', 0),
(3, 'abaisses', 'V1', 'abaisser', 'VRB', NULL, 'S', '2', 'Ind-pr', 0),
(4, 'abaisse', 'V1', 'abaisser', 'VRB', NULL, 'S', '3', 'Ind-pr', 0),
(5, 'abaissons', 'V1', 'abaisser', 'VRB', NULL, 'P', '1', 'Ind-pr', 0),
(6, 'abaissez', 'V1', 'abaisser', 'VRB', NULL, 'P', '2', 'Ind-pr', 0),
(7, 'abaissent', 'V1', 'abaisser', 'VRB', NULL, 'P', '3', 'Ind-pr', 0),
(8, 'abaissais', 'V1', 'abaisser', 'VRB', NULL, 'S', '1', 'Ind-imp', 0),
(9, 'abaissais', 'V1', 'abaisser', 'VRB', NULL, 'S', '2', 'Ind-imp', 0),
```

Figure 11 Table de formes dans Morfetik

3.2. De Morfetik à DELAF

Les dictionnaires électroniques utilisés dans Unitex correspondent au formalisme des DELA (Dictionnaires Electroniques du LADL). Ce formalisme permet de décrire les entrées lexicales simples et composées en leur associant des informations grammaticales, sémantiques et flexionnelles. La ressource lexicale Morfetik doit être transformée en format de DELA pour qu'elle soit exploitable par Unitex.

Dans Morfetik, la table de lemme enregistre les formes de toutes les unités lexicales recensées. Chaque forme d'unité lexicale est associée avec un identifiant, le lemme, les informations sur la catégorie grammaticale suivie du code de flexion, le nombre, le genre, la personne et le temps (si l'entrée est un verbe). La Figure 12 montre un exemple.

```
(`forme_id`, `forme`, `lemme_id`, `lemme`, `catgram`, `gender`, `number`, `pers
(1, 'abaisser', 'V1', 'abaisser', 'VRB', NULL, '', '', 'Inf', 0),
(2, 'abaisse', 'V1', 'abaisser', 'VRB', NULL, 'S', '1', 'Ind-pr', 0),
(3, 'abaisses', 'V1', 'abaisser', 'VRB', NULL, 'S', '2', 'Ind-pr', 0),
(4, 'abaisse', 'V1', 'abaisser', 'VRB', NULL, 'S', '3', 'Ind-pr', 0),
(5, 'abaissons', 'V1', 'abaisser', 'VRB', NULL, 'P', '1', 'Ind-pr', 0),
(6, 'abaissez', 'V1', 'abaisser', 'VRB', NULL, 'P', '2', 'Ind-pr', 0),
(7, 'abaissent', 'V1', 'abaisser', 'VRB', NULL, 'P', '3', 'Ind-pr', 0),
(8, 'abaissais', 'V1', 'abaisser', 'VRB', NULL, 'S', '1', 'Ind-imp', 0),
(9, 'abaissais', 'V1', 'abaisser', 'VRB', NULL, 'S', '2', 'Ind-imp', 0),
```

Figure 12 Table de lemmes dans Morfetik

Dans DELAF, les informations sur la forme flexionnelle, le lemme, la catégorie grammaticale, le nombre, le genre, la personne et le temps (si l'entrée est un verbe) respectent la forme comme ce qui suit : *abalourdiriez,abalourdir,.V+z1:F2p*. la forme flexionnelle (*abalourdiriez*) et le lemme (*abalourdir*) sont séparés par la virgule et le lemme est séparé des codes par le point. *V* est le code grammatical, *z1* est le code sémantique, *F* est le code sur le temps, *2* est le code sur la personne et *p* est le code sur le nombre. Dans *abaissée,abaisser,.V:Kfs*, *V* est le code grammatical, *K* est le code de temps, *f* est le code de genre et *s* est le code de nombre. Le code sémantique est optionnel.

Ainsi, pour transformer Morfetik à DELAF, c'est le format des données qui doit être changé. L'idée est que l'on enregistre respectivement les informations de différents champs

dans la table de formes de Morfetik dans une série de variables et qu'on réaffiche ces informations en respectant le format de DELAF. On remplace respectivement les codes de catégorie grammaticale, de genre, de nombre, de personne et de temps dans Morfetik par les codes définis pour indiquer respectivement les mêmes informations dans DELA. Par exemple, la catégorie grammaticale « verbe » qui est représentée par le code « VRB » dans Morfetik est remplacé par le code « V » qui est défini pour référer à la catégorie grammaticale « verbe » dans DELA ; le code « Ind-pr » qui désigne l'indicatif présent dans Morketik est remplacé par le code « P » défini pour désigner aussi l'indicatif présent dans DELA ; le code « S » (en majuscule) qui indique le nombre simple dans Morfetik est remplacé par le code « s » (en minuscule) défini pour indiquer le nombre simple dans DELA. Nous avons écrit un script perl en fonction de cet algorithme pour transformer Morfetik à DELAF. La Figure 13 est la capture d'écran d'une partie du résultat. En comparant avec la Figure 11, on peut voir que les entrées dans la table de formes de Morfetik sont réaffichées en respectant le format de DELAF.

```
abaisser,abaisser,.V:W
abaisse,abaisser,.V:P1s
abaisses,abaisser,.V:P2s
abaisse,abaisser,.V:P3s
abaissons,abaisser,.V:P1p
abaissez,abaisser,.V:P2p
abaissent,abaisser,.V:P3p
abaissais,abaisser,.V:I1s
abaissais,abaisser,.V:I2s
abaissait,abaisser,.V:I3s
abaissions,abaisser,.V:I1p
abaissiez,abaisser,.V:I2p
abaissaient,abaisser,.V:I3p
```

Figure 13 Résultat de transformation de Morfetik en DELAF

Deuxième partie : Analyse des données

Chapitre 1 Méthode distributionnelle

La méthode distributionnelle repose sur la relation d'appropriation entre les prédicats appropriés et leurs arguments. Un ensemble de prédicats appropriés définitionnels caractérisent une classe sémantique. Nous choisissons un vocabulaire spécifique (les noms d'artefacts) et appliquons la méthode distributionnelle (supervisée et semi-supervisée) à ce vocabulaire pour une expérimentation. Dans ce chapitre, on présente premièrement la constitution du corpus pour la méthode distributionnelle. Ensuite, on présente la méthode distributionnelle (supervisée et semi-supervisée). Enfin, on expose l'évaluation par rapport à la pertinence des résultats.

1. Profilage du corpus

Le corpus assure un rôle crucial à la fois dans les analyses linguistiques et dans les analyses statistiques. Le profilage du corpus détermine directement la qualité des résultats des analyses. Quel genre de corpus faut-il choisir ? Est-ce qu'il est représentatif ? Est-ce qu'il nous faut choisir un corpus de textes spécialisés ? Ce sont toutes les questions que l'on se pose à propos du profilage du corpus. Aujourd'hui, il devient de plus en plus facile d'accéder aux informations ou aux données électroniques avec l'évolution d'Internet. Cependant, on doit savoir quelle est la particularité du discours d'Internet, quels sont les traits langagiers contenus dans ces discours, comment doit-on les exploiter et par quel moyen le corpus numérique doit-il être constitué ; ce sont les questions que l'on doit se poser pour l'analyse et la constitution du corpus.

Dans ce qui suit, on présentera en premier lieu la problématique et l'état du point de vue scientifique du profilage du corpus. Ensuite, on exposera l'état de l'art sur l'aspirateur de site web et les problématiques sur la constitution du corpus issu du web. Et puis, on présentera le corpus utilisé dans le cadre de la méthode distributionnelle. Finalement, on expliquera le

principe de l'outil RENE (Récupération, Extraction, Nettoyage et Encodage) développé pour la construction du corpus dans le cadre du projet.

1.1. Problématique et état scientifique

Un corpus n'est pas un ensemble de données langagières qui sont en vrac mais des données qu'on décide de regrouper pour une étude particulière (Habert et al., 1997). La construction du corpus doit correspondre aux objets scientifiques et permet de répondre aux problématiques.

La représentativité, la clôture ainsi que le genre textuel et le type de texte sont trois problématiques classiques dans les études de corpus. Dans cette section, on exposera d'abord l'état scientifique sur ces problématiques. Et puis, on présentera la typologie des corpus. Ensuite, on s'intéresse à la quantification des traits langagiers. Finalement, on présentera les discours d'Internet provenant du développement de la technologie informatique sous les angles suivants : les traits langagiers, les structures de formation et la constitution du corpus à partir de celui-ci.

1.1.1. Représentativité

La représentativité du corpus est directement liée à la génération des résultats de l'analyse ou du projet dans le traitement automatique des langues. En fonction des différents objectifs de l'analyse ou du projet, la représentativité du corpus se pose différemment. Pour l'élaboration des grammaires ou des dictionnaires, le corpus utilisé doit être représentatif pour des usages spécifiques d'une langue (Dugas, 2010). Dans ce cas, la quantité des données et la couverture de tous les registres sont déterminantes. Dans le cas où le corpus est constitué pour une analyse de langue de spécialité, un corpus de textes de spécialité dans un domaine spécifique doit nécessairement être pris en compte. Si un corpus est constitué afin d'étudier un fonctionnement linguistique ou un phénomène linguistique particulier (syntaxique, lexical ou discursif), un corpus contenant assez d'occurrences de ce phénomène ou de ce fonctionnement est obligatoire.

Anne Condamines (2005 : 19) a présenté trois cas en rapport avec la représentativité du corpus : le corpus existe préalablement à l'analyse qu'en fait le linguiste ; le corpus est constitué pour représenter une langue ou un état de langue et pour la description d'un phénomène linguistique ou celle d'un phénomène de connaissance au sens large. Il va de soi que la sélection du corpus dépend de l'objectif de l'analyse ou du projet.

1.1.2. Clôture

La clôture du corpus est une question associée à la question de la représentativité. La question de clôture est de savoir jusqu'où le linguiste peut faire intervenir ses connaissances pour construire l'interprétation (Condamines, 2005 : 22). Par exemple, dans un corpus médical, on trouve les termes *lésion*, *obstruction*, *sténose*, *occlusion*, *réocclusion* et les composés *artère lésée*, *artère sténosée*, *artère occluse*, mais il n'y a ni *artère obstruée* ni *artère réoccluse*. On se demande alors : est-ce qu'on doit douter de la représentativité du corpus ou est-ce qu'on doit conclure que ces termes n'existent pas puisqu'ils sont absents dans le corpus ? Ce phénomène peut avoir lieu dans toutes les sortes de corpus et pour tous les phénomènes linguistiques étudiés. Cependant, une analyse linguistique prend souvent en compte non seulement les données présentes dans un corpus mais aussi les données absentes qu'un analyste croit concernées.

1.1.3. Genre, registre et type de textes

Le concept de genre textuel implique que les textes soient caractérisés en fonction des caractéristiques linguistiques et métalinguistiques qu'ils partagent. Un texte devient ainsi représentatif d'un ensemble d'autres textes identiques selon cette hypothèse. Ainsi, la description d'un phénomène dans un texte est valable pour tous les autres textes du même genre. La notion de genre constitue une façon de s'inscrire socialement et linguistiquement dans une communauté qui existe déjà (Condamines, 2005 : 25). Les genres ou les registres sont les catégories intuitives qu'utilisent les locuteurs pour répartir les productions langagières, par exemple sous la forme d'un journal, d'un forum, d'un livre, etc. Les textes peuvent aussi être classés en fonction du sujet et du domaine.

Il existe également une classification de textes à partir d'un traitement statistique fondé sur l'étiquetage des textes. Elle permet de définir les corrélations de traits linguistiques (Habert, Fabre et Issac, 1998 : 40-42). D. Biber (1994) distingue clairement les types de textes, qui relèvent de l'analyse linguistique (l'écrit, l'oral, scientifique, etc.), et les registres ou genres qui correspondent à une catégorisation sociale (journal, roman, forum, etc.). Pour les textes écrits, il faut savoir qu'en général, ils ne sont pas rédigés à destination des linguistes. Pour les textes oraux, la situation de communication, telle que le rôle de l'intonation et de la prosodie, ou celui du statut des locuteurs, etc. doit être pris en compte. L'étude des gestes, des regards ou des postures des locuteurs et des interlocuteurs font aussi sens aux communications verbales. Le discours oral a été figé à un moment de transcription mais il garde toujours un caractère de processus en raison de la chronologie des communications.

1.1.4. Typologie de corpus

Les types de corpus envisagés varient en fonction de l'objectif de l'analyse ou de l'application à partir de ces corpus. Par exemple, pour les analyses contrastives diachroniques, un corpus diachronique est nécessaire ; pour les analyses syntaxiques ou certaines applications de TAL basées sur les analyses syntaxiques, un corpus annoté qui indique les relations syntaxiques est potentiellement requis ; pour une analyse d'une langue à une autre, un corpus aligné est très utile. Dans ce qui suit, on présentera les quatre types de corpus les plus utilisés en linguistique ainsi qu'en TAL.

Les corpus arborés sont les corpus associés avec des annotations syntaxiques de manière entièrement ou partiellement manuelles ou automatiques. Les annotations syntaxiques sont représentées par les arbres qui sont le dispositif habituel pour noter les relations syntaxiques. Les arbres syntaxiques sont composés de nœuds terminaux, de feuilles et d'autres nœuds non-terminaux. Les nœuds non-terminaux dominent directement les feuilles. Par exemple, pour représenter la phrase *Je tire un oiseau au vol en pyjama* , on trouve la forme d'arbre suivant :

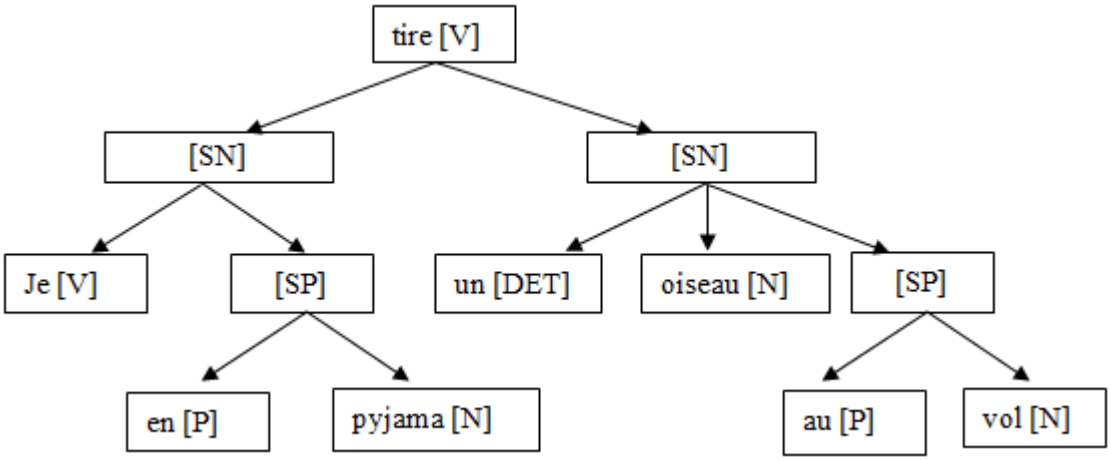

Figure 14 Corpus arborés

Dans ce schéma, SN indique le syntagme nominal, le SP réfère au syntagme prépositionnel, P signifie la préposition, DET indique le déterminant et N désigne le nom. Si l'on utilise les "[" et "]" pour grouper les constituants syntaxiques et si l'on fait un étiquetage pour chaque constituant au sein des parenthèses simples, on peut obtenir l'annotation syntaxique comme ce qui suit :

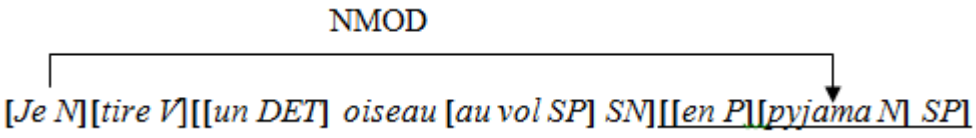

Figure 15 Annotation syntaxique

NMOD signifie N+Modifieur.

Le corpus diachronique est plutôt destiné aux études sur l'évolution de la langue. Le Trésor de la Langue Française s'appuie sur une base de textes s'étalant du XVIe au XXe siècle. Le corpus d'Helsinki est un corpus diachronique en anglais qui contient les textes de la période datant de 750 allant jusqu'à 1700. Cependant, la constitution et l'annotation d'un corpus diachronique se heurte à plusieurs obstacles: premièrement, les définitions et les distinctions de registre des textes diffèrent l'une de l'autre au fil du temps ; deuxièmement, les corpus historiques sur l'ancienne langue sont très petits et peu annotés ; troisièmement, les étiqueteurs ou parseurs automatiques développés aujourd'hui se limitent à l'ordre des mots de la langue actuelle, alors que la plupart des langues anciennes connaissent une variation importante dans l'ordre des mots.

« On appelle textes alignés des couples de textes dont l'un est une traduction de l'autre et pour lesquels il existe un système de mise en relation entre segments du texte de "grain équivalent" : sections, paragraphes, phrases » (Habert, 1997 : 135). Par exemple, le corpus HANSARD (documents du Parlement canadien) est un corpus bilingue anglais-français qui contient les discours législatifs de 1995, 1996 et 1997. Il comprend 2,87 millions de paires de phrases alignées ; le corpus EUROPARL (débat du Parlement européen) est aussi un corpus aligné. Il comprend onze langues et 20 millions de mots par langue. Le corpus aligné est souvent utilisé dans les études de traduction automatique ou dans l'élaboration des ressources lexicales multilingues. La reconstitution des correspondances traductionnelles unit les segments d'un texte source et ceux de sa traduction (Isabelle et Warwick-Amstrong, 1993 : 288). L'alignement peut s'effectuer aux différents niveaux de structuration de l'énoncé : les mots, les phrases, les paragraphes et les sections (Habert, 1997 : 138).

Un corpus de référence est constitué comme un échantillon représentatif de la langue traitée. Il vise à être suffisamment représentatif de toutes les variétés pertinentes de cette langue et de son vocabulaire caractéristique, de manière à pouvoir servir de base à des grammaires, des dictionnaires et d'autres ouvrages usuels fiables (Sinclair, 1996 : p.10). Par exemple, le corpus BNC (British National Corpus) est un corpus de référence qui contient 100 millions de mots étiquetés ; le corpus Frantext est aussi un corpus de référence rassemblant les textes littéraires et une petite proportion de textes scientifiques datant de XVIe siècle jusqu'à aujourd'hui.

Le corpus de textes de spécialité prend son origine dans les langues de spécialité. Les textes de spécialité sont les productions linguistiques, orales ou écrites, qui se manifestent dans le cadre des communications professionnelles et dont la finalité est exclusivement professionnelle (Cabré, 2007 : 38). Selon Harris Z. (2007 : 47), les sous-langages sont les langues de disciplines scientifiques ou techniques, ou les métalangues (comme celles de la grammaire ou de la linguistique) ; les sous-langages se caractérisent par un lexique limité et par l'existence de schémas de phrases en nombre fini. La langue de spécialité peut être définie comme un sous-langage d'un domaine spécifique selon Harris (2007 : 56-58). Les corpus

spécialisés réunissent des données linguistiques relatives à une dimension particulière : un domaine, un thème ou une situation de communication (Habert, 1997 : 144). Néanmoins, il faut remarquer que les énoncés dans un domaine particulier n'excluent pas les traits langagiers conclus dans la langue générale (opposé à la langue de spécialité) et ne comportent pas tous les phénomènes de la langue spécialisée de ce domaine.

1.1.5. Quantification des traits langagiers

La quantification des traits langagiers a pour objet d'étudier la variation de quantité de certains traits langagiers dans un ou différents corpus. La quantification des traits langagiers est souvent utilisée dans l'analyse du discours et surtout dans la statistique textuelle. Par exemple, dans l'ancienne langue française, on ne voit jamais le mot *footeux* (qui signifie joueur de football ou simple amateur et qui est parfois péjoratif), alors qu'à partir de l'année 2011, *footeux* est recensé par le dictionnaire Larousse et dans les textes d'aujourd'hui, surtout dans le discours d'Internet, *footeux* est utilisé de plus en plus fréquemment. Un autre exemple, dans les discussions à un forum thématique, les pronoms personnels de la première personne sont plus nombreux par rapport à ceux des actualités de ce forum. La quantification des traits langagiers peut s'effectuer sur les unités lexicales, sur l'axe de syntagmatique (par ex., les récurrences d'unités comme : *sécurité sociale de la France*, *niveau de vue en métropolitaine*, etc.) ou même sur l'axe paradigmatique. La comparaison des décomptes peut se dérouler au sein d'un corpus partitionné.

1.1.6. Corpus web

Avec le développement de l'informatique et des réseaux Internet, un nouveau genre de corpus apparaît--- le Corpus web. Néanmoins, une série de questions sont posées : est-ce que le corpus web correspond à des nouveaux modèles d'analyse ? Est-ce que les concepts et les notions utilisés dans les études et les analyses de corpus traditionnelles doivent être changés pour le corpus web et quels nouveaux concepts ou notions doivent être introduits pour cerner ses spécificités ? Quels moyens ou quels outils faudrait-t-il adopter pour constituer le corpus web ? Dans cette partie, on répondra à ces questions en exposant les structures, les traits

langagiers, et la constitution du corpus web. La discussion sur les traits langagiers et sur les structures des corpus web s'effectue par l'analyse d'un exemple de corpus web---le forum de discussion.

Un forum de discussion est une plateforme pour les échanges authentiques produits en l'absence de l'analyste qui les enregistre. La discussion sur un forum comprend les échanges qui ont eu lieu et qui sont en train d'avoir lieu. Les échanges qui auront lieu seront aussi enregistrés dans la discussion par la mise en mémoire du forum. Il s'agit ainsi d'un corpus homogène sans début ni fin. Les messages adressés à un forum sont classés par date, sujet et émetteur. Mourlhon-Dallies (2004 : 28-32) a proposé quatre modes de production de messages dans sa présentation sur la structure des échanges: à l'Internaute, un nouveau message peut être posté en position d'intervention initiative en ouvrant un nouveau fil de discussion ; il peut aussi être posté comme intervention réactive pour répondre à un message de forum ; un message peut aussi être envoyé directement à son interlocuteur par courrier électronique ; il peut également être à la fois posté comme intervention réactive et envoyé par courrier électronique à son interlocuteur. Pour envoyer un message à un forum, on demande à l'utilisateur de choisir le déplacement de son intervention dans la structuration, le destinataire à qui il veut envoyer le message et ce qu'il reste ou en sort après avoir posté le message. Néanmoins, on peut trouver beaucoup de questions qui n'ont jamais pu recevoir de réponses. De plus, il y a également beaucoup de réponses au message initial mais postées dans un message précédent. Les messages dans une discussion d'un thème sont enchaînés, c'est-à-dire que chaque intervention reprend une information ou une unité linguistique liée à la précédente dans l'échange.

Dans les discussions sur les forums, on peut trouver beaucoup de langages SMS (tels que *2m1* (demain), *bi1* (bien), *koi 2 9*(quoi de neuf), etc.) et langages Internet (tels que les codes d'abréviation comme *geeks* (gens passionnés et obsédés par l'informatique), *HS* (hors sujet), *xdr/xpdr/xpldr* (explosé de rire), *a+/ ++/@+* (à plus tard), *rpg* (Role Player Game), etc., et les émoticônes (ou smileys) comme *:S/é_è* (est confus), *:p/:P* (tire la langue), *<3* (coeur qui exprime l'amour, la tendresse), *:x/:-x* (est muet), *:'(* (pleure), etc.). Le langage SMS

diffère du langage Internet par plusieurs aspects : l'objectif de langage SMS est de mettre les mots en phonétique avec les caractères plus rapides à accéder, par exemple, *moi* est tapé comme *mwa* car *M*, *W* et *A* sont les lettres apparaissant en priorité en appuyant sur les touches correspondantes, alors que *O* et *I* sont troisièmes ; ce genre de transformation perd l'intérêt pour l'écriture sur l'ordinateur car *O* et *I* sont plus accessibles que *W* sur le clavier AZERTY ; Le langage Internet est plutôt utilisé pour une répartie, telle que *lol* (laughing out loud), *dtc* (dans ton cul) ou *ctb* (comme ta bite) ; les émoticônes sont peu utilisés pour les SMS en raison de la difficulté de saisie par les touches de téléphone. Les unités lexicales utilisées dans la discussion d'Internet comprennent de nombreux emprunts ou beaucoup de cas d'alternance codique, par exemple, dans un forum du thème informatique, on trouve une réponse comme : *Il faut delete ce fichier*. Le mot anglais *delete* est directement utilisé dans une phrase en français. En réalité, l'Internet est une grande source pour la naissance des néologismes, car les messages et les informations électroniques sont renouvelés plus rapidement, la discussion ou de reportage s'organise toujours à propos des thèmes les plus nouveaux, la langue là-dessus est toujours la plus récente et les échanges multilingues sont beaucoup plus facilités. De plus, dans la discussion sur le forum, la source de laquelle les informations sont transférées n'est pas toujours indiquée. Le discours rapporté non marqué affiche une présence importante. Par exemple, le discours *Même si ça continue à être très tendance* ***d'interdire d'interdire***. vient d'une légère transformation du slogan de 1968 *Il est interdit d'interdire*. Ce discours est de la modalisation autonymique d'emprunt et il est non marqué mais identifiable par interprétation (Mourlhon-Daillies, 2007 : 92). Finalement, il faudrait souligner que les phrases dans la discussion sur le forum sont plus simples mais moins structurées. Il existe souvent des fautes d'orthographes ou des fautes grammaticales qui échappent aux interlocuteurs.

Il existe également beaucoup d'autres genres de corpus web, tels que les journaux, les pages personnelles, les blogs, les commentaires sur les sites de vente, etc. Les différents genres de corpus web sont associés aux différentes caractéristiques linguistiques et métalinguistiques, mais ils partagent quand même certaines caractéristiques communes en raison de leur nature électronique et informatique. Les corpus web sont renouvelés plus rapidement et s'enrichissent de nouveaux phénomènes langagiers (tels que les néologismes,

les emprunts, les langages Internet, etc.). Les corpus web nous permettent d'étudier l'évolution de certains phénomènes langagiers (sur l'axe de termes, de syntagmes, de phrases ou d'énoncés) sur de très courtes durées. On peut avoir une comparaison entre les différents états de langue et déceler les changements de langue au fil du temps.

La constitution du corpus web exige deux processus importants : la documentation et la normalisation. La documentation est nécessaire pour la construction de tous les types de corpus. Un corpus sans la documentation perd sa valeur d'analyse. Benoît Habert (1997 : 156) a indiqué : « la documentation du corpus doit couvrir deux volets distincts : les sources utilisées et la responsabilité éditoriale de constitution du corpus d'une part, les conventions d'annotation d'autre part ». Pour documenter un corpus, D. Biber (1994 : 380-385) fournit une série de paramètres : canal (écrit, parlé ou écrit lu), format (publié ou non publié), cadre (institutionnel, autre cadre public ou privé interpersonnel), destinataire (pluralité : non compté, pluriel, individuel ou soi-même, présence : présent ou absent, interaction : aucun, peu ou beaucoup, connaissances partagées : générales, spécialisées ou personnelles), destinateur (variation démographique : sexe, âge, profession, etc., statut : individu/institution dont l'identité est connue, factualité : informatif factuel, intermédiaire ou imaginaire), objectifs (destinés à persuader, amuser, édifier, informer, expliquer, donner des consignes, raconter, décrire, enregistrer, se révéler, améliorer les relations interpersonnelles, etc.) et thèmes. La normalisation du corpus d'Internet s'effectue par deux aspects : codage et définition de type de document. ASCII-ISO 646 est un code de caractères sur 7 bits. On peut coder 128 caractères différents avec ASCII-ISO 646. Il constitue le seul codage de caractères universels. Le code ISO-LATIN-n-ISO 8859-n ne permet de coder que des textes en anglais, alors que le codage ISO-LATIN-n permet de définir d'autres langues européennes. ISO-LATIN-n est un codage sur 8 bits et il permet de représenter 256 caractères différents (2^8). La définition de type de document (DTD) est un document permettant de définir une structure de document. La structure de document est décrite au niveau de la structure logique et au niveau de la structure physique. La structure logique peut indiquer le nom, le nombre d'occurrences et l'ordre d'apparition des éléments, des sous-éléments ou des attributs associés qui peuvent apparaître. La structure physique peut définir les entités générales, telles qu'une abréviation

pour un fragment de texte répétitif, un renvoi à un autre fichier ou un synonyme composé de caractère permettant d'identifier les références par nom au lieu d'un code numérique. SGML est un métalangage pour définir les types de documents. Il permet d'étiqueter les textes avec le balisage logique qui vise à expliciter la structure d'un document. DSSL (Document style Semantics and specification Language) est une norme internationale (ISO-10179) visant à associer de manière normalisée des traitements à des documents balisés avec SGML. XML est une version simplifiée de la norme SGML (un sous-ensemble) et HTML (Hyper Text Markup Language) est un langage de description d'hypertextes pour le réseau Internet obéissant à une grammaire (une DTD) SGML.

1.2. Corpus pour la méthode distributionnelle

Le corpus constitué pour la méthode distributionnelle doit permettre de fournir les informations et le moyen d'exploitation des informations requis par cette méthode. La méthode distributionnelle se base sur les études de la fonction argumentale et consiste à exploiter le modèle de données des trois fonctions primaires pour acquérir automatiquement les noms d'artefacts. Ainsi, le corpus doit comprendre des occurrences de termes d'étude et des structures prédicat-argument suffisantes. Une variété de thèmes doit être garantie et la possibilité de considérer les nouveaux phénomènes langagiers (concernant les termes) est préférable. En fonction de ces critères, on propose quatre sources et certains genres de textes provenant de ces sources proposées pour constituer nos corpus dans le cadre de la méthode distributionnelle. La discussion sur les genres des textes se déroule sous deux angles : le contenu des textes et les traits langagiers des textes (y compris, la structure de la langue, la richesse des termes, etc.).

1.2.1. Sources : blog, forum, communauté et site de vente

Les blogs sont les plateformes électroniques permettant de publier les articles régulièrement en rendant compte des actualités sur le sujet donné. Ces articles sont postés avec les dates et les signatures. Ils sont publiés successivement dans un ordre antéchronologique (du plus récent au plus ancien). Un blog peut aussi catégoriser ses articles

selon les sous-sujets, par exemple, pour un blog du sujet jardinage, les articles là-dessus peuvent être catégorisés en fonction des sous-sujets : plantes vertes, fleurs, outils de jardinage, écologie, etc. Une catégorie pour les articles les plus lus peut aussi être proposée sur un blog. De plus, un espace de discussion après chaque article sur les blogs est souvent mis en place pour les interlocuteurs.

Les forums sont les plateformes informatiques virtuelles permettant de publier les discussions librement sur divers sujets. Il permet les échanges à distance. L'archivage des discussions sur le forum rend aussi les communications asynchrones possibles. Les forums de discussion permettent des conversations discontinues. (Mourlhon-Dailles, 2007 : 28). En général, un forum est créé selon une thématique et les différents espaces de discussions au sein du forum sont mis en place en fonction des différents sous-sujets. Dans chaque espace de discussion de sous-sujet, une ou plusieurs questions peuvent être postées et les réponses correspondantes suivent leur question. Les réponses et les questions sont datées et signées et elles sont aussi affichées dans un ordre antéchronologique.

Beaucoup de sites web combinent la fonction du forum et la fonction du blog. Ils permettent à la fois de publier les actualités et les articles de la thématique et de publier les discussions librement dans un forum intégré à ces sites web. Un espace de discussion ou un champ de commentaires est aussi mis en place après chaque article ou chaque actualité. On appelle ce genre de sites web les communautés thématiques.

Le site de vente s'appelle aussi la boutique en ligne. Il groupe des catalogues d'articles vendus et les services fournis associés aux articles. Pour chaque article, il y une description sur le produit, une présentation sur les services associés (par ex., le paiement, la livraison, le retour, etc.) et un espace de commentaires. Un hyperlien associé à un blog thématique sur les produits vendus aussi peut être souvent trouvé sur le site de vente. La partie la plus dynamique du site de vente est son espace de commentaires associé à chaque produit. Les clients publient librement leurs avis sur la qualité, leur satisfaction ou leurs expériences d'utilisation des produits qu'ils ont achetés dans cette boutique en ligne. À l'espace de commentaires, la plupart des messages envoyés sont des commentaires sur les produits, alors

qu'il est aussi probable que certaines discussions parmi un petit groupe d'internautes aient lieu.

Il faut également remarquer que certains sites web combinent non seulement la fonction du blog et du forum mais aussi la fonction du site de vente. Ils sont créés en fonction d'une thématique et fournissent toutes les possibilités d'échanges. Ils permettent de publier toutes les informations associées au thème, telles que les articles, les actualités, les discussions, les commentaires et même les vidéos. Un site de vente intégré nous permet de trouver tous les produits associés au thème.

1.2.2. Genres de textes

Tant sur les blogs, les forums, les communautés thématiques que sur les sites de vente, les genres de textes archivés sont variés. Par exemple, sur certains blogs de cuisine, en plus des articles qui rendent en compte des actualités thématiques, un autre genre de textes existent aussi---les recettes ; sur le site de communauté, les genres de textes archivés sont plus nombreux : la discussion dans le forum, les astuces sur une thématique, les actualités, les dossiers (tels que les manuels, les modes d'emploi, les présentations scientifiques, etc.) …… Dans ce qui suit, on présentera en premier lieu certains genres de textes, tels que les guides d'achat dans les sites de vente, les discussions au forum, les recettes sur les blogs de cuisine…du point de vue de leurs contenus et leurs traits langagiers. Ensuite, on présentera les genres de textes qu'on choisit et les démarches de l'établissement d'une liste d'URL (Uniforme Resource Locator : adresse électronique qui permet de localiser un site ou un document sur internet) pour la constitution du corpus dans le cadre du projet.

Les guides d'achat et les Astuces sont les genres de textes souvent trouvés dans les communautés thématiques ou dans les sites de vente. Les guides d'achat sont des articles rédigés pour guider les clients ou les consommateurs dans le choix d'achat. En général, dans un guide d'achat, on fait la comparaison du prix et de la performance des produits. On conseille également les critères à considérer avant de faire les achats. Les conseils donnés correspondent aux différents besoins de différentes personnes. Les Astuces sont les articles

qui présentent les astuces d'utilisation du produit pour la protection, l'économie ou une technique de manipulation. Le thème des Astuces ne se limite pas aux utilisations des produits. Il peut concerner tous les aspects de la vie quotidienne, tels que l'investissement, la relation interpersonnelle, les réseaux sociaux des entreprises, la création du blog personnel, l'établissement de différents types de contrats, etc. En ce qui concerne les traits langagiers des Astuces, les unités lexicales décrivant l'action associée à l'objet de discussion sont fréquentes. Dans les Guides d'achat, les unités lexicales décrivant les caractéristiques ou les fonctionnements des produits sont plus nombreux. De plus, les chiffres, l'adjectif *certain* (s), le syntagme *tous les +N,* le superlatif *le plus* ou *le moins* et l'adverbe relatif *plus* ou *moins* ont aussi une grande fréquence d'occurrence dans les guides d'achat. La langue dans les guides d'achat et dans les Astuces est plus structurée et plus riche par rapport à la langue des Tests. Les Tests sont plutôt les fiches de tableau avec une description très brève. Beaucoup de descriptions dans les Tests sont même générées par un robot.

Les actualités du sujet donné peuvent être trouvées dans les blogs, dans les sites de vente ou dans les sites de communauté thématique. Par exemple, pour une communauté informatique, dans le catalogue d'actualités, on peut trouver toutes les actualités concernant l'informatique, telles que le développement de la nouvelle technologie, le nouveau produit informatique lancé par les entreprises, le chiffre d'affaires trimestriel des entreprises informatiques, etc. Ce genre de textes décrit plutôt un évènement identifié par la date, le lieu et les personnes concernées. Le sujet de l'évènement correspond à la thématique du site web. Cependant, il existe une contradiction, c'est que les actualités sur les blogs du sujet d'artefact ne comprennent pas beaucoup d'occurrences de noms d'artefacts. Différents des journaux, les actualités sur les blogs, sur les sites de vente ou sur les sites de communauté ne diffusent, parfois, qu'un message pour informer un appel (par ex., un appel d'entretien, de participation à un concours thématique, de communication, etc.) ou un évènement qui a eu lieu ou qui aura lieu. Les contenus des actualités sur les blogs, sur les sites de vente ou sur les communautés thématiques peuvent concerner la vie professionnelle, une expérience personnelle associée à la thématique, une solution proposée pour le ciblage publicitaire, etc. La langue des actualités est plus structurée et comprend peu de fautes grammaticales ou de fautes d'orthographe. Les

phrases composées de plus d'une vingtaine de mots dans les actualités sont plus nombreuses par rapport aux guides d'achat ou des Astuces.

Les commentaires sur les sites de vente sont les avis publiés librement par les clients à propos des articles vendus au site. Les avis publiés peuvent concerner la qualité du produit, les caractéristiques du produit (telles que la couleur, la taille, la matière, la forme, etc.), la performance du produit (telle que la vitesse, le stockage, le confort, etc.) ou le prix du produit. Il peut également s'agir d'une expérience d'utilisation du produit ou d'une comparaison avec d'autres modèles ou avec les mêmes modèles de différentes marques. Les avis peuvent être négatifs ou positifs. Ils comprennent souvent les plaintes concernant les défauts du produit, les problèmes qu'on éprouve au cours de l'utilisation du produit ou les problèmes concernant la livraison du produit (par ex., l'emballage déchiré, la perte du colis, l'envoi en retard, etc.). Certains conseils ou astuces sont aussi donnés par certains clients dans les commentaires et parfois, une discussion parmi un petit groupe d'internautes peut avoir lieu. Les publicités sont souvent postées comme des commentaires des clients. Certains messages envoyés dans l'espace de commentaires pour les produits sont même générés et publiés par les robots ou par les personnes embauchées des entreprises. L'objectif de ces messages vise à fournir une critique positive du produit pour favoriser le bénéfice commercial. Ils comprennent souvent les structures de phrase similaires et de nombreux termes positifs. La langue des commentaires est plus courte et orale par rapport aux langues dans les guides d'achat, dans les Actualités, ou dans les Astuces. Beaucoup de fautes d'orthographe et grammaticales existent dans les commentaires. Les langages SMS et les langages Internet sont fréquemment utilisés pour critiquer en ligne. Il existe également de nombreux termes et syntagmes de critique, tels que *attirant, vraiment, ce sont des pros, à ma grande surprise, j'ai été étonné, j'ai été très surprise, ça m'a déçu, je le trouve très sympa, j'ai été stupéfait*, etc.

Les discussions sur le forum sont une série de questions/réponses à propos du sujet donné. Dans la discussion, on partage les expériences, les arguments à l'appui d'une idée, des solutions pour un problème, des conseils, les informations associées, etc. On peut aussi trouver de nombreux hyperliens liés aux sites externes (ou internes) proposé par les

internautes sur l'objet de discussion. La discussion de forum comprend aussi de nombreux langages SMS et langages Internet. Pareillement, la langue de discussion au forum est moins structurée, plus orale et plus courte. Il existe également beaucoup de fautes d'orthographe et grammaticales. La distribution du lexique est autour du sujet de discussion. Cependant, les termes concernant les autres sujets associés sont souvent introduits par les interlocuteurs. Ainsi, les messages au forum sur les artefacts sont riches du vocabulaire des noms d'artefacts.

Les recettes sont souvent publiées sur les blogs ou sur les communautés de cuisine. Une recette est normalement composée de deux parties : la liste d'ingrédients et les phases techniques de la recette. La liste d'ingrédients comprend tous les ingrédients nécessaires et la quantité nécessaire pour chaque ingrédient, par exemple, 1 tomate rouge, 1 botte de basilic frais, 15 cl de lait, 2 œufs entiers et 200 g de chapelure fine. Les phases techniques de la recette sont plutôt décrites par les syntagmes verbaux, tels que *Couper en tranches fines et régulières les tomates avec un couteau d'office, Disposer sur le fond de 4 grandes assiettes en alternant les couleurs, Faire frire ces beignets à la dernière minute à la friteuse très chaude, Assaisonner la tomate avec une pincée de sel, de poivre du moulin et un filet d'huile d'olive, etc.* La structure de la langue dans la recette correspond bien à la structure prédicat-argument. Le lexique dans la recette concerne non seulement les ingrédients listés, les actions pour faire la cuisine mais aussi beaucoup d'outils de cuisson introduits dans la description des phases techniques de recette. Cependant, ce genre de textes ont un sujet très restreint. La recette est un genre de textes spécifique pour la cuisine. Elle est riche de noms d'artefacts de la classe sémantique d'Appareils de cuisson, mais pour les noms d'artefacts des autres classes sémantiques, la recette ne contient presque aucun terme spécifique.

Pour les noms d'artefacts, on élabore un corpus qui concerne les thèmes principaux de la vie quotidienne, tels que le moyen de transport, la beauté et la santé, le bricolage et la décoration, le ménage, l'informatique, etc. On a choisi six genres de textes sur les thèmes cités ci-dessus : les commentaires dans les sites de vente, les actualités sur les blogs ou dans les communautés thématiques, les discussions aux forums, les guides d'achat et les astuces dans les sites de vente ou dans les communautés thématiques. Les textes choisis pour

constituer le corpus proviennent d'une dizaine de sites web (cf. Tableau 4). Le volume de corpus sur les noms d'artefacts atteint 22,858 Ko. Il comprend 3,754,334 mots, à savoir 19,099,378 caractères (espaces non compris) ou 22,808,096 caractères (espaces compris). Les textes de différents thèmes et de différents genres occupent environ la même proportion dans le corpus.

Informations / Sites	**Genre de textes**	**Thème**	**URL**
Comment ça marche	Actualités	Informatique	http://www.commentcamarche.net/news/
Comment ça marche	Forum	Informatique	http://www.commentcamarche.net/forum/
Comment ça marche	Astuces	Informatique	http://www.commentcamarche.net/faq/
Cnet France	Guides d'achat	Informatique	http://www.cnetfrance.fr/produits/guides/
Cent France	Actualités	Informatique	http://www.cnetfrance.fr/news/
Meilleur du chef	Forum	Cuisine	http://www.meilleurduchef.com/cgi/mdc/forum/fr
Meilleur du chef	Recettes	Cuisine	http://www.meilleurduchef.com/cgi/mdc/l/fr/recette/index.html
Geek Food	Blog	Cuisine	http://geekandfood.fr/blog/
Internaute	Forum	Bricolage	http://bricolage.linternaute.com/forum/electromenager-14
Auto Cara	Forum	Automobile	http://www.forum-auto.com/marques/index.htm
Futura	Forum	Bricolage et décoration	http://forums.futura-sciences.com/bricolage-decoration/
Doctissimo	Forum	Vie (santé, beauté, mode, etc.)	http://forum.doctissimo.fr/
Au féminin	Forum	Vie (santé, beauté, mode, etc.)	http://www.aufeminin.com/world/communaute/forum/forum0.asp
Ciao	Commentaires	Bureautique	http://www.ciao.fr/sr/q-bureautique
Ciao	Commentaires	Electromenager	http://www.ciao.fr/sr/q-electromenager
Ciao	Commentaires	Automobile	http://www.ciao.fr/sr/q-automobile
Blogautomobile	Blog	Automobile	http://blogautomobile.fr/#axzz32I3cV5UC

Tableau 4 Liste de sites web à aspirer pour constituer le corpus de noms d'artefacts

Le corpus dans le cadre de la méthode distributionnelle est élaboré de façon automatique. Un aspirateur de site web est développé dans le but de faciliter la construction automatique du corpus du projet. Cet outil prend en compte les problématiques classiques dans l'aspiration du web et la constitution du corpus. Dans ce qui suit, on en fera une présentation en détail.

1.3. Outil de constitution de corpus dans le projet

Les méthodes pour construire des corpus issus du web ont été bien étudiées au cours de la dernière décennie. Les aspirateurs de sites web ont une histoire aussi longue que celle de l'Internet. Ils permettent de récupérer les pages web associées à un ensemble d'URL données à l'avance et les pages web associées à tous les hyperliens contenus dans les pages web téléchargées à partir des URL initiales. Le processus agit de façon récursive. Un aspirateur de sites web est un outil très important pour collecter les pages web et construire de grands corpus informatisés. Dans cette section, on expose premièrement ce qui en est des aspirateurs de site web et la problématique qui existe actuellement pour la constitution automatique d'un corpus issu du web. Ensuite, on explique l'architecture de l'outil d'aspiration de web RENE qu'on a développé dans le but de contourner les défauts des aspirateurs de web actuels et de faciliter la construction de notre corpus dans le cadre du projet. On présente ensuite le principe de chaque module de notre aspirateur de site web et finalement le mode d'emploi de l'outil.

1.3.1. État de l'art à propos des aspirateurs de site web

Au fur et à mesure du développement de l'informatique, de plus en plus de recherches linguistiques utilisent les données provenant des sites web comme corpus. Les aspirateurs de sites web sont ainsi développés pour réaliser la constitution de corpus informatisés en grande quantité. Les aspirateurs de sites web permettent de récupérer les pages web associées à un ensemble d'URL données à l'avance et les pages web associées à tous les hyperliens contenus dans les pages web téléchargées à partir des URL initiales. Le processus agit de façon récursive. En fonction des différents objectifs de la construction du corpus, les aspirateurs sont construits de différentes façons. Par exemple, un corpus pour la recherche de néologismes doit toujours être le plus récent. Cela exige des mises à jour régulières du corpus. Dans ce cas-là, nous avons besoin d'un aspirateur de sites web dont la performance peut facilement renouveler le corpus régulièrement et efficacement. Il existe également plusieurs aspirateurs de sites web en licence libre offerts au public, tels que HTTrack er BootCat.

Dans ce qui suit, on présentera trois outils pour aspirer les sites web : HTTrack, BootCat et Telanaute. Puis, on les comparera entre eux. Ensuite, on fera état de trois outils établis dans le but de la construction des différents corpus : il y a premièrement un outil de construction du corpus permettant d'éliminer les doublons des textes et les parties insignifiantes pour la recherche visée ; on trouve ensuite un outil incrémental qui répond aux besoins de renouvellement régulier et efficace du corpus ; enfin, on obtient un outil qui a pour objet de récupérer les ressources provenant de façon aléatoire de domaines variables.

1.3.1.1. Trois outils pour aspirer le site web

HTTrack est un aspirateur de sites web robuste et en licence libre qui permet de télécharger un site web entier automatiquement ou interactivement, de télécharger des fichiers spécifiques et d'aspirer tous les sites web dans les pages. De plus, il est doté de la possibilité de renouveler le corpus téléchargé à la demande. HTTrack est un aspirateur de sites web configurable. On peut paramétrer les règles de filtrage (qui permettent d'exclure ou d'inclure les différents types de fichiers des différents sites), la profondeur de téléchargement, la taille maximale des fichiers, le temps de téléchargement maximal, le navigateur utilisé par défaut, la structure d'organisation des fichiers téléchargés, etc. Le site web copié est enregistré par HTTrack sur le disque dur de l'ordinateur en construisant les répertoires récupérant les html, les images et d'autres types de fichiers.

La première étape est de nommer le projet et de configurer le chemin de base où l'on enregistre le site web copié. Par défaut, HTTrack enregistre le site web copié dans C:\Mes Sites Web. On peut modifier le chemin de base pour répondre aux différents besoins. La deuxième étape a pour objet de définir une action en fournissant un ou plusieurs URL et de configurer les paramètres du téléchargement de chaque site web. Les options d'action sont données à l'avance par HTTrack : Copie automatique de site(s) Web, Copie interactive de site(s) Web (questions), Télécharger les fichiers spécifiques, Aspirer tous les sites dans les pages (miroirs multiples), Tester les liens dans les pages (tests de signet), Reprendre une copie interrompue et Mettre à jour une copie existante. La "Copie automatique de sites Web" permet de télécharger directement tout le site web sans avoir à présenter une demande au

milieu de téléchargement. Au contraire, la "Copie interactive de site(s) Web (questions)" exige la permission de l'utilisateur chaque fois qu'on a détecté un lien au sein du site web. Cela permet à l'utilisateur d'exclure certaines pages d'un site qui sont insignifiantes en fonction des différents besoins. De plus, dans le processus de configuration des paramètres, on peut aussi exclure certains types de fichiers (telles que les images *.png, *.gif, *.jpg ; les scripts *.css, *.js, etc.) dans le cadre " Règles de filtrage " ou exclure encore certaines pages web dans le cadre "Limites" où l'on peut configurer la profondeur d'aspiration, la taille maximale des fichiers html qu'on récupère, le temps maximal d'aspiration, etc. On peut aussi définir une structure d'organisation des fichiers copiés dans le même temps. Enfin, après avoir précisé les options de connexion (par exemple, la suspension de connexion d'internet à la fin de l'opération, la fermeture du PC après l'opération, etc.), on peut enfoncer la touche "Terminer" pour lancer le téléchargement.

L'avantage de HTTrack est qu'il est très efficace pour copier un site entier. Il permet aussi de configurer la profondeur et la largeur d'aspiration. Cependant, la flexibilité pour sélectionner les liens au sein des sites web apparaît un peu moins performante. Il permet soit de copier tout le site web, soit de copier une partie de ce site web selon la profondeur prédéfinie. Si l'on veut télécharger seulement certains liens dans la page web au lieu de tous, HTTrack est inadéquat. De plus, HTTrack ne permet pas le nettoyage des pages web récupérées. Dans certains cas, le corpus construit par HTTrack ne peut pas être utilisé directement. Il faut quand même faire un nettoyage ou un prétraitement avant de l'utiliser.

BootCat est un autre aspirateur de sites web en licence libre qui permet de récupérer les URL à l'aide de mots clés à travers le moteur de recherche. Avec les mots clés saisis par l'utilisateur, BootCat peut faire une combinaison de ces mots clés pour former les requêtes utilisées dans le moteur de recherche. On peut définir la longueur de la combinaison et le nombre maximal des combinaisons. Par exemple, on a une liste de mots clés : frigo, pain, boîte, confiture, hygiène et poêle. Si l'on définit la longueur de la combinaison comme 2, chaque combinaison sera composée de deux mots clés et on aura donc 15 ($C_n^i = \frac{n \times (n-1) \times ... \times (n-(i-1))}{i}$) possibilités de combinaison. Si le nombre maximal de

combinaisons est évalué à 10, les cinq dernières combinaisons seront ignorées. Si la longueur de la combinaison et le nombre maximal de combinaisons sont respectivement de 3 et 8, on aura 10 possibilités de combinaisons et les deux dernières combinaisons sont ignorées. Ensuite, BootCat utilise les combinaisons de mots clés comme requêtes et les entre dans le moteur de recherche pour récupérer tous les URL correspondantes. Pour la récupération des URL, on peut également configurer les limites, telles que la limite de domaines de sites (ex, .edu, .org, etc.) où l'on fait la recherche, la limite du nombre maximal d'URL récupérées, etc. Ensuite, on lance la collection des URL et BootCat va afficher tous les liens récupérés dans l'interface après avoir fini la récupération. On peut choisir les liens qui nous intéressent et exclure manuellement ceux qui ne le sont pas à partir de la liste d'URL récupérées. Enfin, après avoir téléchargé les pages web en fonction de la liste d'URL fournie à la fin, BootCat fait un nettoyage basique automatiquement. Il enlève les codes html, les menues, les barres de navigation, les messages d'erreurs générés automatiquement, les scripts, etc. Avec cette méthode d'aspirer qui se base sur les mots clés, on peut obtenir un corpus qui couvre des domaines variables.

Cependant, le nettoyage fourni par BootCat est un nettoyage très basique. Il est possible qu'il reste quand même certaines parties des textes qui soient insignifiantes pour les analyses, car les différents contenus dans la page web ne sont pas identifiés et distingués. Or, le plus souvent, selon les différents objectifs d'analyses spécifiques, on a besoin des différents discours. Par exemple, pour analyser le discours politique à partir du corpus récupéré du site web "www.lemonde.fr", les publicités et les commentaires qui illustrent les contenus des articles deviennent le bruit. Néanmoins, si l'objectif de notre travail est d'analyser le discours des commentaires dans le site web "www.lemonde.fr", tous les contenus des articles dans le site deviennent en revanche les composantes de bruit. Un autre inconvénient de BootCat est que l'on doit filtrer les liens manuellement chaque fois. Si le nombre d'URL récupérées par BootCat est colossal, le travail de filtre devient difficile.

Telanaute est un aspirateur de sites web développé au sein du laboratoire LDI (Lexicologie Dictionnaires Informatique) par Fabrice Issac (2007). Il est développé en se

basant sur le concept de module et permet d'intégrer le module de traitement (appelé "handler") développé ou personnalisé par l'utilisateur. Telanaute est composé de deux modules. Dans le premier module "arpenteur", on met un ensemble d'URL dans la liste "URL à traiter" et on récupère les pages webs associées aux URL dans la liste "URL à traiter". On enregistre après, respectivement la page html récupérée et les hyperliens dans cette page web découverte à partir de la liste de "URL à traiter" et on continue le processus récursivement. La profondeur d'aspiration est configurable. On peut également faire un filtrage sur les pages html à enregistrer et les hyperliens à utiliser pour l'aspiration suivante. En même temps, dans le deuxième module, dès qu'une page web est récupérée, on exécute le traitement des données de cette page. On peut définir le filtre d'information pour chaque page web. Par exemple, dans une page html contenant les articles, les commentaires, les publicités, les scripts, les liens, etc., on peut exclure les informations insignifiantes (par exemple, les publicités, les commentaires, ou les scripts) par le filtre d'information. L'utilisateur peut également définir son traitement de données spécifiques par la création de greffons en utilisant l'interface de programmation appliquée (API). Dans la partie suivante, on présente chaque module avec plus de détails.

Le premier module a pour objet de récupérer les pages web et les pages web associées aux hyperliens contenues dans ces pages web. Le processus se poursuit récursivement jusqu'à ce qu'on ait téléchargé toutes les pages web au sein du site indiqué. Cependant, toutes les pages web récupérées ne sont pas intéressantes pour l'analyse. La possibilité de faire un filtre sur ces pages devient, ainsi, très importante. Telanaute permet de faire un filtre par les trois types de paramètres : la syntaxe URL, la méta-information et la source (le contenu de la page html) des pages récupérées avant de les passer dans le deuxième module. Avec chaque paramètre, on renvoie une valeur booléenne indiquant si la page est acceptée ou refusée. Le premier paramètre permet de filtrer les URL qui ne sont pas intéressantes et restreindre le robot à un ou certains sites web. Le deuxième paramètre concerne les méta-informations renvoyées par le serveur et cela permet d'éliminer les pages dont les méta-informations ne respectent pas la configuration. Par exemple, si l'on limite la valeur d'attribut "content-length" inférieure à 100, les pages web dont la taille est supérieure à 100 seront toutes filtrées. Le

troisième paramètre s'applique au corps du texte et les balises html. On peut réaliser tous les types de traitements du texte, tels que la sélection des pages web d'une langue particulière, la sélection des pages web contenant certains mots clés ou certains groupes de mots, etc. Les filtres sont réalisés par Python placés dans le répertoire v0.6/greffons/ du répertoire de l'utilisateur Telanaute. La capture d'écran suivant montre la configuration du paramètre de source dans le fichier " AutoFiltre.py " :

```
1   import re
2   class Filtre:
3       def __init__(self):
4           #init
5           self.nom='FiltreAuto'
6           seil.portailAuto=re.compile(r'citroëne|peu',re.IGNORECASE)
7       def test(self,texte):
8           if self.portailAuto.search(texte):
9               return True
0           else:
1               return False
```

Figure 16 "AutoFiltre.py" dans Telanaute

Dans le fichier " AutoFiltre.py ", on utilise les expressions régulières pour repérer les mots clés définis dans le code source de la page html. Les pages web dont les contenus ne comprennent pas les mots clés sont éliminées par le filtre. Les pages web html sélectionnées finalement sont enregistrées dans le disque et les URL sélectionnées sont enregistrés dans une base de données.

Synchroniquement, le deuxième module qui s'appelle "handler" fournit une chaîne de traitements. L'utilisateur peut aussi développer son propre traitement et l'intégrer dans le module "handler". On nettoie la page et on fait une segmentation des phrases. Dans ce module, on garde la trace de chaque transformation du fichier. Ainsi, selon la demande, on peut réutiliser les fichiers intermédiaires. Le schéma ci-dessous illustre le fonctionnement des deux modules dans Telanaute :

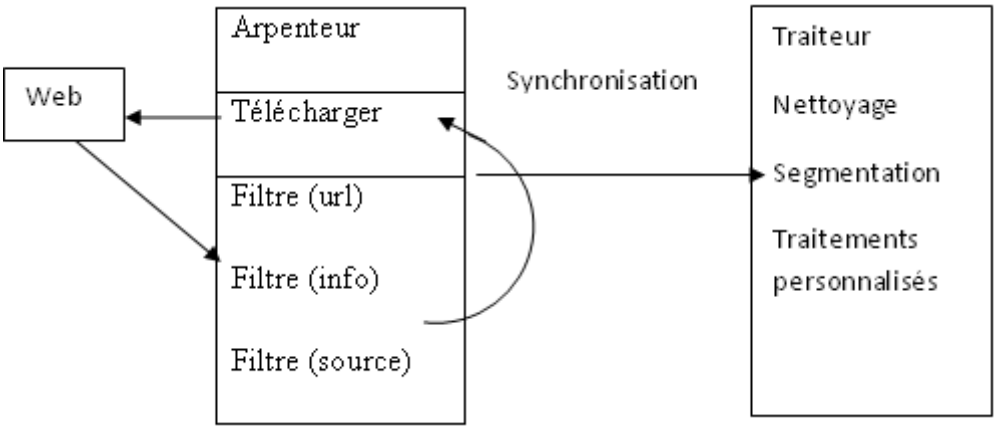

Figure 17 Architecture de Telanaute

Telanaute peut être utilisé directement en ligne de commande. Il y a aussi une interface Web utilisateur qui regroupe toutes les fonctions des deux modules “arpenteur” et “traiteur”. Le script "interface.py" situé dans le dossier v0.6/interface/ du répertoire de l’utilisateur Telanaute (accessible à http://telanaute.univ-paris13.fr:12357/index.tp) permet de lancer l’interface web.

Par rapport aux HTTrack et BootCat, Telanaute est beaucoup plus configurable, mais il est un peu compliqué du point de vue de sa manipulation. Chaque fois qu’on veut lancer une aspiration, il faut effacer l’ancienne table créée dans la base de données tout d’abord, réaffiner les filtres, initialiser les URL, lancer la commande pour se connecter à la base de données où l’on enregistre les URL filtrées. Si l’on veut lancer l’aspiration plusieurs fois, on doit nécessairement effectuer plusieurs opérations et passer entre les fichiers et le terminal ou l’interface Web plusieurs fois. Cependant, Telanaute est d’une grande flexibilité de configuration et une possibilité d’interaction avec l’utilisateur. Il permet aussi de créer le module de traitement personnalisé de l’utilisateur.

1.3.1.2. Outils pour les constructions du corpus Web

La construction d’un corpus Web est une étape très importante dans les analyses linguistiques, la fouille de données et le traitement automatique des langues naturelles. La qualité de corpus a une grande influence sur la qualité des analyses ou sur le résultat de traitement. Dans l’introduction, on a présenté la variabilité des aspirateurs. Dans cette section,

on présentera trois outils de construction du corpus développés dans le but de différentes applications.

Dans un premier temps, on présentera un outil de construction du corpus qui permet d'extraire les parties pertinentes à partir des pages web téléchargées et d'assurer que la plupart des textes ne sont présents qu'une fois dans le corpus. Dans un second temps, on expliquera le fonctionnement d'un outil qui permet la mise à jour régulière du corpus régulièrement avec une grande efficacité. En définitive, on fera état d'un aspirateur de sites web qui se concentre sur le problème de la variété, de la richesse et des productions au hasard du corpus.

A. Building Large Corpora from the Web using a New Efficient Tool Chain (Roland Schäfer, Felix Bildhauer, 2012)

Schäfer et Bildhauer (2012) ont proposé une chaîne d'outils pour construire un corpus web multilingue très large. Cet outil permet d'extraire les paragraphes de textes concernés et d'éliminer les textes répétitifs. Cet outil suit cinq étapes de traitements : aspirer les ressources du site web, éliminer les parties de bruit, faire le nettoyage, détecter les textes connectés et enlever le duplicata du corpus.

L'aspirateur de sites web Heritrix 1.4 permet de récupérer les pages web associées avec des milliers d'URL initiales et couvre des milliers d'URL. Heritrix est un aspirateur de sites web très robuste qui a été développé en Java par Internet Archive (IA). Heritrix est modulaire, "multithread" et capable de traiter les grandes aspirations. Il se divise en trois parties : Scope, Frontière et Processeur de chaînes. "Scope" détermine quelles URL découvertes seront utilisées pour l'aspiration et limite les URL à certains domaines ou sous-domaines. "Scope" peut utiliser les expressions régulières très complexes pour filtrer et refuser les fichiers contenant plus d'un certain nombre de liens. "Frontière" maintient l'état interne de l'aspiration et surveille tout le temps quels URL ont déjà été récupérées, lesquelles doivent être téléchargées et lesquelles sont en train d'être récupérées. "Processeur de chaînes" a pour objet de normaliser les URL, de filtrer après les URL selon la configuration, de récupérer les pages web associées avec les URL, de les enregistrer sur le disque et de fournir les

informations sur l'état d'aspiration. Cependant, Heritrix ne permet pas de renouveler le corpus régulièrement comme un aspirateur incrémental. Pour lancer Heritrix, on doit fournir un ensemble d'URL qui peuvent être récupérés par ODP (Open Directory Project). ODP est un répertoire de sites Web géré par une vaste communauté d'éditeurs bénévoles provenant de tous les milieux. Des millions d'URL sont catégorisées en fonction de critères définis. Toutes les données d'ODP sont accessibles gratuitement.

On utilise, ensuite, la méthode WaCky (Baroni et al., 2009) pour détecter les textes destinés à des analyses linguistiques et enlever toutes les autres parties (les parties de bruit). L'idée est qu'on sélectionne la fenêtre du texte à partir du document récupéré pour calculer le ratio de caractères textuels contre les caractères de balise. Si le ratio du texte est maximal par rapport aux ratios des autres blocs des textes dans le document, on considère que cette partie est pertinente et on l'extrait directement. Cette méthode se base sur l'observation que la section d'une page qui est riche de contenus a une moins grande densité de balises html et que les parties de bruit comportent un grand nombre de balises. Cependant, cette méthode ne peut servir à extraire les segments de textes connectés discontinus. Par exemple, avec une partie de bruit au milieu d'un texte d'une page, un texte est divisé en deux segments. Or, avec cette méthode, on va extraire soit un de ces deux segments soit toute la partie y compris la partie de bruit au milieu. De plus, il y a aussi le problème concernant les marges de la section extraite. Quelques fois, à un bout de la section extraite, une partie de bruit est incluse ou à un autre bout de la section extraite, on a enlevé certains textes connectés. Actuellement, de plus en plus de méthodes supervisées ou semi-supervisées sont utilisées. Le principe de ces méthodes est d'entraîner un classifieur statistique à partir d'un ensemble de documents annotés par les experts. Avec le classifieur entraîné, on peut identifier automatiquement les sections riches de contenus dans les nouveaux documents.

Troisièmement, on fait un nettoyage basique qui a pour objet d'éliminer tous les scripts et les balises dans les pages html. De plus, puisque la chaîne d'outils est optimisée pour l'encodage ISO-8859, dans cette étape, on fait aussi une conversion des caractères UTF-8 en

encodage ISO-8859. En même temps, on convertit également les entités html en caractères normaux.

Quatrièmement, on utilise un outil (Texprof) pour détecter les textes connectés. Les textes non connectés indiquent les textes dans lesquels il existe les matériels non phrastiques. Pour détecter les textes connectés, on adopte la méthode dans laquelle on tokenize un ensemble de documents entraînés et on calcule les fréquences relatives (fréquence calculée par rapport à la longueur du document) des types de tokens les plus fréquents pour entraîner un classifieur statistique. Avec ce classificateur, on détecte les textes connectés et on élimine tous les textes non connectés.

Finalement, pour éviter de mettre les documents répétitifs dans le corpus final, on fait une suppression des duplicatas. Il y a deux types de suppressions de duplicata : la suppression de duplicata parfaite et la suppression de duplicata proche. Pour la suppression de duplicata parfaite, l'outil Texrex fonctionne très bien. Il crée un tableau de 128 caractères pour chaque document. Les 128 caractères sont uniformément distribués sur tout le document. Les tableaux établis pour chaque document sont enregistrés dans un tableau associatif. Si les nouveaux documents entrés correspondent à un tableau enregistré, ils sont considérés comme duplicata et sont retirés. En ce qui concerne la suppression de duplicata proche, on a utilisé une chaîne d'outils. Tout d'abord, on utilise l'outil Teshi pour tokenizer les documents et former un ensemble de tokens de n-grammes (qui s'appellent aussi w-galets) pour chaque document. Ensuite, l'outil Tender calcule les n-grammes partagés entre chaque document. Pour chaque paire de documents avec un nombre de n-grammes partagés, si le nombre de ne-grammes partagés par deux documents est supérieur à un seuil, le document moins long entre les deux sera éliminé à la fin.

Pour l'expérimentation de la chaîne d'outils, les URL du Yahoo ont été récupérées jusqu'à ce que l'accès d'API libre soit discontinu comme les URL initiales du projet. Ensuite, Heritrix 1.4 a été utilisé pour découvrir plus d'URL à travers le moteur de recherche Microsoft Bing et récupérer les pages web en excluant les fichiers non textuels et les fichiers des autres langues. Néanmoins, Schäfer et Bildhauer (2012 : 488) ont indiqué, ici, que la récupération des URL

dépendait complètement du moteur de recherche, ce qui est une stratégie dangereuse. La discontinuation de Yahoo peut bien prouver le danger de la stratégie dépendant du moteur de recherche. À travers leurs expérimentations, l'équipe de Roland (2012) a découvert qu'un grand nombre de serveurs qui n'apparaissent jamais dans les résultats d'un moteur de recherche peuvent être détectés par l'aspiration de terme long. Cependant, la récupération des URL à travers le moteur de recherche peut garantir l'effet " hasard " du corpus construit.

En résumé, le corpus construit avec cette chaîne d'outils est un ensemble de documents bien nettoyé avec peu de répétitions. En calculant la densité de caractères de balises dans le document, on peut identifier rapidement les parties de bruit dans tous les documents. Cette méthode permet de traiter un grand nombre de documents sans contrainte. Cependant, l'identification des parties de bruit avec cette méthode n'atteint pas à un taux de précision à 100%. Il reste souvent quelques parties de bruit dans les textes ou certains textes connectés sont éliminés. Pareillement, il existe le même problème pour l'identification des duplicatas et des textes connectés.

B. The Evolution of the Web and Implications for an Incremental Crawler (Junghoo Cho, Hector Garcia-Molina, 1999)

Traditionnellement, un aspirateur de sites web consulte les pages web et les télécharge jusqu'à ce qu'on obtienne le nombre désiré de pages web. Ensuite, quand il est nécessaire de renouveler la collection, l'aspirateur de sites web relance le même processus pour faire une nouvelle collection et remplace l'ancienne. On appelle ce type d'aspirateur un aspirateur de sites web périodique. Si l'aspirateur de site web renouvelle la collection en remplaçant les pages moins importantes au fur et à mesure par les nouvelles pages qui sont plus importantes, on l'appelle l'aspirateur incrémental. En général, un aspirateur incrémental est plus efficace qu'un aspirateur périodique (Cho et Garcia-Molina, 1999 : 200). Par exemple, si l'on peut estimer la fréquence de changement d'une page web, l'aspirateur incrémental peut revisiter seulement les pages qui ont changé au lieu de refaire toute la collection. Dans cette partie, on va discuter la méthode proposée par Cho et Garcia-Molina (1999) pour construire un aspirateur incrémental. Tout d'abord, on présente l'expérimentation exécutée pour un corpus

de plus d'un demi-million de pages pour étudier l'évolution des pages web avec le temps. Ensuite, on présente l'architecture de l'aspirateur incrémental établi par Jungo et Hector.

Premièrement, on a collecté 720,000 pages web des 270 sites les plus populaires, y compris Yahoo (http://yahoo.com), Microsoft (http://microsoft.com), Stanford (http://www.stanford.edu), …… en utilisant l'aspirateur de site web Stanford WebBase. Pour chaque site, on a aspiré 3,000 pages. L'objectif d'expérimentation est de mesurer la popularité d'un site dans le temps pour estimer l'intervalle de changement moyen de toute la page web. Si une page web est liée par beaucoup d'autres pages web, on considère que cette page est populaire. La popularité d'une page P est notée comme PR(P) et elle est définie par l'équation suivante :

$$PR(P) = d + (1-d) \quad [PR(P_1)=c_1 + PR(P_2)=c_2 \ldots.. PR(P_n)=c_n] \qquad (27)$$

dans laquelle $P_1 \ldots\ldots P_n$ représentent les pages de différentes périodes indiquant une même page P, $c_1 \ldots\ldots c_n$ sont les nombres de liens dans les pages $P_1 \ldots\ldots P_n$ et d est un coefficient qui est défini comme 0,9 dans l'expérimentation. Ainsi, à chaque étape de calcul, une page web a une équation et la nouvelle valeur de PR(P) est calculée à partir des anciennes valeurs PR(Pi) jusqu'à ce que les valeurs convergent. À travers cette expérimentation, on a essayé de répondre à quelques questions sur le changement de pages web : quelle est la fréquence de changement d'une page web ? Quelle est la durée de vie d'une page web ? le changement d'une moitié d'un site prend combien de temps ? Est-ce qu'on décrit les changements de pages web par un modèle mathématique ? En résumé, les pages web varient rapidement et la fréquence de changement change d'un site à un autre. En vérifiant pendant combien de jours une page est accessible (sans considérer le changement du contenu) dans notre fenêtre, on trouve que plus de 70% de pages venant de tous les domaines restent dans la fenêtre plus d'un mois et plus de 50% de pages dans les domaines « edu » et « gov » restent plus de quatre mois. Les pages du domaine « com » ont une durée de vie plus courte et les pages des domaines « edu » et « gov » vivent le plus longtemps. Le changement de 50% de pages du domaine com prend seulement 11 jours, alors que le changement de 50% de pages des domaines

« gov » et « edu » prend 4 mois. Le modèle mathématique pour décrire le changement d'une page est donc possible.

L'aspirateur de site web incrémental établi par Junghoo et Hector permet de remplacer les anciennes pages et les pages moins importantes par de nouvelles pages dont certaines seront plus importantes. De plus, l'importance des pages existantes change avec le temps. Ainsi, si les pages existantes deviennent moins importantes que les pages ignorées précédemment, cet aspirateur permet aussi de remplacer les pages existantes par les pages ignorées précédemment. L'aspirateur incrémental de Junghoo et Hector est composé de trois modules : RankingModule, UpdateModule et CrawlModule et trois structures de données : AllURL, CollURL et Collection. AllURL enregistre toutes les URL découvertes et CollURL enregistre les URL dans la Collection qui maintient un nombre de pages fixe. RankingModule choisit les URL dans CollURL et évalue la popularité de toutes les URL constamment. Quand une page qui n'est pas dans CollURL devient plus importante qu'une page dans CollURL, RankingModule planifie le remplacement. UpdateModule extrait les entrées les plus hautes dans CollURL constamment et demande à CrawlModule de télécharger les pages web et remettre les URL aspirées dans CollURL. La position d'URL dans CollURL est déterminée par la fréquence de changement estimée et son importance. UpdateModule estime la fréquence de changement en enregistrant la somme de contrôle (l'histoire de changement) de la page. Ainsi, cet aspirateur peut renouveler le corpus sans remplacer tous les documents dans la collection.

C. Efficient Web Crawling for Large Text Corpora (Vít Suchomel, Jan Pomikálek, 2012)

La méthode de Suchomel et Pomikálek (2012) se focalise sur la façon de traiter les données inefficaces et comment aspirer les domaines riches de textes. L'outil SpiderLine développé par Vít et Jan permet de filtrer les données inefficaces et d'éviter de travailler sur des domaines non significatifs. Pour développer l'outil SpiderLine, on a fait une expérimentation avec Heritrix sur le taux d'utilisation des pages web téléchargées. Dans la

partie suivante, on va présenter, en premier lieu, l'expérimentation et le résultat d'analyses correspondants. En second lieu, on présente le fonctionnement de SpiderLine.

L'objectif de l'expérimentation est d'analyser combien de données téléchargées ne sont pas utilisées pour la construction du corpus. On a analysé le corpus du Portugais (la langue Européen) qui est d'un milliard de mots téléchargé du domaine « .pt » avec Heritrix. Pour chaque page web téléchargée, on calcule son taux de rendement selon la formule suivante :

$$\textit{Taux de rendement} = \frac{Donnée\ finale}{Donnée\ téléchargée} \tag{28}$$

dans laquelle la donnée finale indique le nombre de bits dans le texte qui contribue au corpus final ; la donnée téléchargée correspond au nombre de bits téléchargés. On a remarqué qu'il y a de nombreuses pages ayant un taux de rendement zéro, du fait qu'elles sont refusées par le classifieur de langue ou qu'elles contiennent des textes répétitifs. Ensuite, on groupe les données selon les domaines web et on calcule le taux de rendement de chaque domaine (qui est le taux de rendement moyen des pages contenues). De plus, on a choisi une série de seuils et pour chaque seuil, on calcule le nombre de domaines ayant le taux de rendement le plus élevé, la quantité de pages téléchargées et la quantité de pages contribuant au corpus final dans ces domaines. On a trouvé que la quantité de téléchargement baisse rapidement au fur et à mesure de l'augmentation du seuil alors qu'on a seulement perdu très peu de données finales. Cela signifie qu'un aspirateur de sites web peut gagner beaucoup de temps en évitant les domaines ayant un taux de rendement bas sans perdre trop de données finales.

Ainsi, l'aspirateur de site web SpiderLing est développé en se basant sur le principe que si le taux de rendement d'un domaine est inférieur à un seuil prédéfini, on cesse d'aspirer les pages de ce domaine. De cette façon, SpiderLing ne cherche que les ressources riches de textes et réduit la somme de téléchargement des documents indésirables. De plus, dans SpiderLine, on a intégré un outil jusText permettant d'éliminer le contenu de bruits, tels que les liens, les publicités, les têtes html, ……pour que seulement les paragraphes comprenant les phrases complètes soient préservés. Dans SpiderLine, on supprime les duplicata en

vérifiant le « checksum » (l'histoire de visite) des pages web. Les pages web ayant un checksum "visité précédemment" sont refusées. Quand on a testéSpiderLine avec 2570 URL initiales et par rapport à Heritrix, l'efficacité de l'aspiration est beaucoup plus élevée.

1.3.2. Problématique

Tous les outils présentés dans l'état de l'art récupèrent toutes les pages web associées aux hyperliens contenus dans les pages web téléchargées à partir des URL initiales sans pourvoir sélectionner automatiquement les hyperliens à aspirer selon la demande donnée à l'avance. Pour faire la sélection des pages web au cours de téléchargements, ils font une interrogation interactive et cela exige beaucoup d'intervention humaine au cours d'un processus d'aspiration de web. Si l'on a de nombreux URL initiales à aspirer, plusieurs niveaux à approfondir pour chaque URL et beaucoup de pages web à enlever à chaque niveau, l'intervention humaine pour faire la sélection devient un travail énorme. Dans la plupart des cas, les niveaux à approfondir et les hyperliens à choisir à chaque niveau pour chaque URL donnée à l'avance diffèrent l'une de l'autre. Cette problématique de largeur et longueur d'aspiration est incontournable pour le développement de tous les types d'outils d'aspiration de site web.

1.3.2.1. Identification des blocs de textes

L'édition des pages web se réalise par le langage HTML (Hyper Text Markup Language) qui obéit à une grammaire (une DTD) SGML. Un document HTML est composé d'une tête (<head><title></title>…</head>) dans laquelle on enregistre les métadonnées, telles que le titre, la date, l'auteur, l'encodage, le style,…et d'un corps ((<body></body>) dans lequel les textes sont enregistrés et mis en forme par les balises de séparations, telles que les balises de titre de section <H1></H1>, les balises de section , les balises de paragraphe , les balises de listes (par ex., <OL>, <UL>, etc.), balises liées à l'hypertexte <link href= " …"></link>, etc.

La rédaction d'une page web respecte la grammaire du langage HTML, mais les textes sur les différentes pages web sont structurés de manières différentes ; par exemple, les débuts des articles dans Le Monde en ligne sont marqués par la balise , alors que ceux dans Le Figaro sont marqués par la balise . Non seulement les balises marquant les débuts des textes mais aussi les balises marquant les fins des textes varient avec le changement de site web. Les façons de structurer les pages web varient non seulement avec le changement de sites web mais aussi au fil de temps. Les pages web du même site mais rédigées à différents temps peuvent être structurées de différentes manières. Il y a autant de façons de structurer des pages web qu'il y a de sites web. Les sites web créées dans différents buts (commercial, d'intérêt général, …) apparaissent tous les jours sans les compter. Ainsi, comment développer un outil permettant d'identifier les blocs de textes pertinents pour les analyses et enlever les textes de brut d'une façon automatique et pertinente à partir de tous les types de sites web est toujours une problématique classique dans l'aspiration de site web. C'est aussi pour cette raison que la plupart des aspirateurs de site web qui intègre le module de nettoyage ne permettent actuellement d'enlever que les balises, les liens et les scripts qui sont les parties communes de la plupart des sites web. Ils ne permettent qu'un nettoyage très basique, car la variété des structures des pages web nous amène beaucoup de difficultés pour identifier les blocs de textes pertinents d'une façon automatisée.

Certains spécialistes ont proposé une méthode statistique qui a pour objet de détecter les blocs de textes pertinents en calculant la densité de caractères balisés dans chaque bloc d'une page web. On considère que le bloc de textes qui a des caractères balisés les moins denses est le bloc de textes pertinents pour les analyses. Cependant, en fonction des différents objectifs d'analyse, le bloc pertinent peut avoir une densité de caractères balisés moins grande, aussi grande ou plus grande que les autres blocs de textes. Par exemple, dans une page web composée d'un article, d'un espace de commentaires et d'un ensemble de publicités, si l'objet d'étude est les commentaires au lieu de l'article, il est possible que le bloc de textes pertinents "commentaires" ait des caractères balisés plus denses que le bloc de textes "article"

et ait autant de densité que le bloc de textes "publicités". Par conséquent, cette méthode statistique donne quand même des bruits et des silences dans les applications parfois.

1.3.2.2. Encodage des pages web

Comme l'édition de pages web, l'encodage de page web en français varie également avec le changement de site web et au fil du temps. Les pages web de chaque site web est encodées de différentes manières. Les pages web au sein d'un même site mais rédigées à différents temps peuvent aussi être encodées différemment. La plupart des pages web en français sont encodées en ISO 8859-1 (qui s'appelle aussi Europe occidental ou latin-1) ou en UTF-8, mais les pages web français en ISO 8859-15 (souvent appelé latin-9), en Windows-1252 (parfois appelé AINSI) ou même en UTF-16 existent aussi.

Dans le français, ce sont les caractères accentués qui font l'objet d'une attention particulière. Pour afficher une page web contenant les caractères accentués correctement dans le navigateur sans problème d'encodage, l'encodage du navigateur et l'encodage du code source de la page web doivent s'accorder. Si l'on veut éviter de configurer l'encodage du navigateur toutes les fois et permettre d'afficher correctement la page web de tous les encodages dans les navigateurs, on peut utiliser les entités html pour représenter les caractères accentués et spéciaux. Les entités html sont utilisées dans le langage html pour incorporer les caractères spécifiques. Les entités html commencent toujours par une perluette « & » et finissent par le signe de ponctuation point-virgule « ; ». Il y a deux types d'entités html : les entités de type numérique composées d'un nombre précédé du caractère croisillon # ; les entités de type caractère composées d'une chaîne de caractères entre la perluette et le point-virgule. Par exemple, pour le caractère accentué en français « à », son entité html de type numérique est « à » et son entité html de type caractère est « à ». De plus, les entités html permettent également d'afficher des caractères qui ne sont pas accessibles depuis le clavier et des caractères en dehors du jeu de caractères déclaré en début de fichier html avec la balise <meta http-equiv="Content-Type" content="text/html" charset="iso-8859-1" > dans l'attribut « charset ». Ainsi, identifier l'encodage de pages web, encoder correctement les

textes extraits des pages web téléchargées et décoder des entités html là-dedans sont nécessaires pour la résolution d'encodage dans la constitution du corpus web.

1.3.3. Présentation d'outil

RENE (Récupération, extraction, nettoyage et encodage) est un outil d'aspiration de site web qu'on développe pour faciliter la constitution du corpus dans le cadre du projet de cette thèse. Il permet de télécharger les pages web, d'extraire les textes pertinents à partir des pages web téléchargées, de les nettoyer et de résoudre le problème d'encodage selon les paramètres configurés à l'avance. L'avantage de cet outil est qu'il permet de configurer la longueur et la largeur d'aspiration de web avant le démarrage de l'outil. Il permet de cibler précisément les pages web qu'on veut télécharger et les textes pertinents pour les analyses qu'on veut extraire à partir des pages web. Il évite trop d'interventions humaines au cours du téléchargement. RENE fait économiser beaucoup de temps à l'utilisateur. Le désavantage de cet outil est qu'il demande à l'utilisateur d'avoir au moins une connaissance de base du principe d'aspirateur de site web et des expressions régulières pour établir les paramètres d'aspiration. L'établissement de paramètres d'aspiration est un peu compliqué et il est aussi facile de commettre des erreurs au cours de paramétrage. Malgré tout, cet outil est très utile pour les linguistes ou les ingénieurs de TALN qui tentent d'avoir un corpus numérique provenant d'une série de sites web bien ciblés mais peu nombreux par une aspiration de plusieurs niveaux et qui ne veulent pas beaucoup de bruits dans leur corpus. Dans ce qui suit, on présentera en détail l'architecture de RENE, son principe de fonctionnement et son utilisation.

1.3.3.1. Architecture

RENE est composé de trois modules et de deux fichiers. Le module Récupération sert à télécharger les pages web à partir d'une URL donnée à l'avance en approfondissant et étendant l'aspiration en fonction des paramètres saisis. Le module Nettoyage&Extraction est à reconnaître le bloc de textes pertinents dans chaque page web selon des paramètres configurés à l'avance, pour les extraire et pour les nettoyer. Le module Encodage résout des problèmes

d'encodage. Les URL initiales et les paramètres de la longueur et de la largeur d'aspiration sont enregistrés dans un fichier. Les paramètres pour l'extraction des textes pertinents sont enregistrés dans l'autre. L'architecture de RENE correspond au schéma suivant :

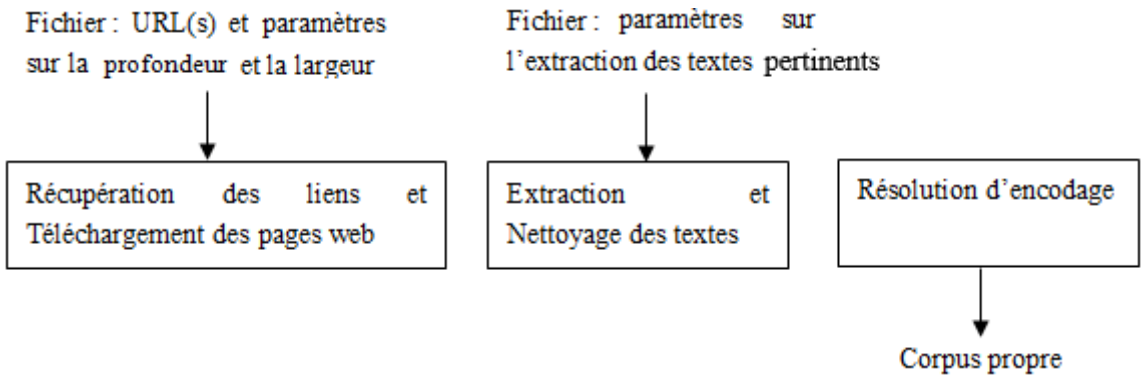

Figure 18 Architecture de RENE

Le premier module met en marche le téléchargement avec les pages web associées aux URL indiquées dans le fichier de paramètres. À partir de chaque page web téléchargée, toutes les pages du tourne-pages (s'il y en a) et les hyperliens de chaque page récupérée sont identifiés en fonction des paramètres donnés à l'avance. Les paramètres indiquant quels hyperliens choisir sont indiqués par les expressions régulières dans le fichier de paramètres. Par exemple, le paramètre d'hyperlien <a class="forumlink" href=\"(http:\/\/www\.le\-forum\-emploi\.com\/[^><]*?)\"> permet à l'aspirateur de récupérer tous les liens qui commencent par *http://www.le-forum-emploi.com* et qui sont enregistrés dans le troisième attribut (*href*) de la balise dont le nom est *a* et dont le deuxième attribut est *class="forumlink"*. À partir des URL récupérées au niveau précédent, l'outil continue à aspirer le tourne-pages et les hyperliens selon les paramètres et répète le processus itérativement jusqu'au dernier niveau d'aspiration. Le nombre de niveaux d'aspiration et le nombre de pages à récupérer dans le tourne-pages est également configuré à l'avance dans le fichier de paramètres. Toutes les URL récupérées au dernier niveau sont liées aux pages web qui doivent être téléchargées. Le schéma suivant montre ce processus itératif :

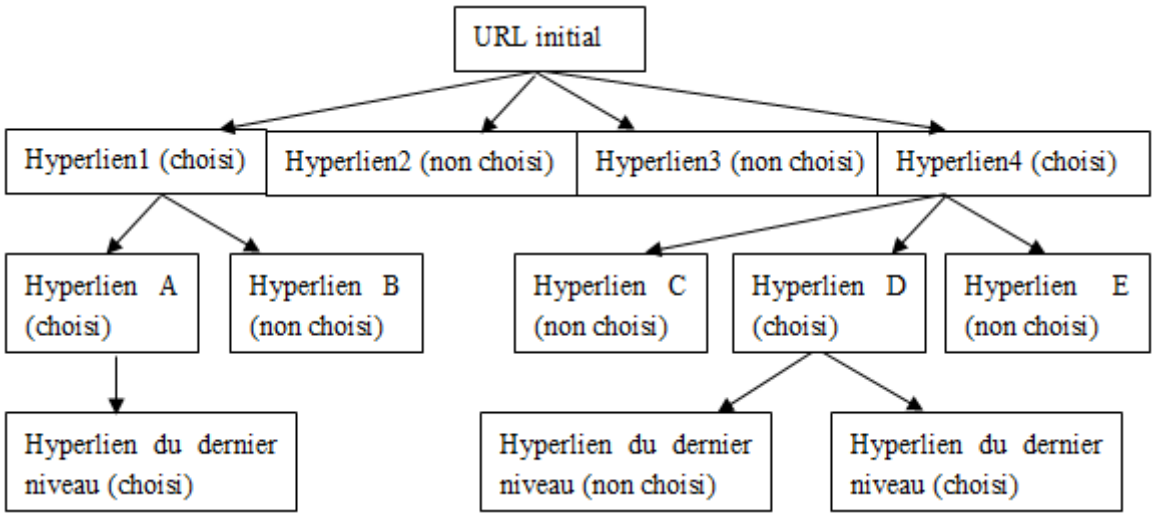

Figure 19 Hiérarchie des niveaux d'aspiration

Le module Récupération est réalisé en programmation orientée objet. Il fait appel à un module perl WWW::Mechanize qui permet de se connecter au serveur et de télécharger les pages web associées aux URL données. Pour accéder au web, Perl fournit une librairie LWP. Il s'agit d'un ensemble de modules permettant d'accéder au web, tels que LWP::UserAgent, WWW::Mechanize, LWP::Simple, LWP::RobotUA, etc. LWP est une librairie orientée objet. Il s'appuie sur le modèle de communication HTTP qui construit une requête, puis l'envoie au serveur et enfin reçoit une réponse du serveur. En se basant sur le protocole de communication HTTP, LWP permet de réaliser non seulement des requêtes HTTP mais aussi FTP, SMTP, des accès au système de fichier local, etc. Le module WWW::Mechanize est une classe héritée de LWP ::UserAgent. L'avantage de ce module est qu'il permet d'envoyer des requêtes de différents protocoles.

Les pages web extraites sont au format html. Elles comprennent des méta-informations, des balises, des scripts, des textes qui sont pertinents pour les analyses du projet et des autres textes. Le deuxième module est composé de deux sous-modules : le sous-module Extraction permet d'identifier le bloc de textes pertinents en fonction de la marque de début et la marque de fin paramétrée à l'avance et de l'extraire automatiquement ; le deuxième module a pour objet de nettoyer les textes extraits. La marque de début est la balise qui marque le début du bloc de textes qu'on veut extraire et la marque de fin est celle qui marque la fin du bloc de textes qu'on veut extraire. La marque de début ou la marque de fin

peut être une balise avec ou sans attributs, mais une paire de marques doit nécessairement permettre d'identifier le bloc de textes désiré. Pareillement, les méta-informations, telles que le titre, l'auteur, la date, etc. peuvent toutes être récupérées avec une paire de balises comme <titre></titre>, <auteur></auteur>, <date></date>, etc. Le nettoyage concerne plutôt l'élimination de certains espaces ou retours de chariot superflus dans les textes extraits. Le sous-module Extraction est programmé en faisant appel au module perl Parser::HTML. Parser::HTML est développé pour faciliter le traitement de fichiers html. Il permet de parcourir un fichier html ligne par ligne et d'enregistrer respectivement les informations de balises (telles que les noms de balises, les noms d'attributs ainsi que les valeurs d'attributs) rencontrées et les textes rencontrés dans les variables prédéfinies du Parser::HTML. Le sous-module Nettoyage est plutôt réalisé à l'aide d'expressions régulières. Par exemple, pour supprimer les espaces superflus dans les textes, on remplace plusieurs espaces par un seul en faisant appel aux expressions régulières s/\s+/\s/.

Les pages web provenant des différents sites web peuvent être encodées différemment. Deux encodages utilisés principalement pour les pages web en français sont utf8 et latin1. Une page web en utf8 doit être téléchargée en encodant en utf8 et une page web en latin1 doit être téléchargée en encodant en latin1. La résolution d'encodage est achevée à l'aide du module perl Encode. Pour décoder les entités html dans les pages web, nous pouvons faire un script en nous fondant sur un tableau d'entités html dans lequel chaque entité html est associée à son caractère spécifique correspondant. Il existe également de nombreux modules permettant de décoder les entités html. Nous avons adopté le module HTML::Entities en perl pour accomplir le décodage d'entités html.

1.3.3.2. Algorithme et programmation

L'outil qu'on développe aspire le web en extension et en profondeur. Du côté de l'extension, il s'agit de tourner les pages et de récupérer les hyperliens. Du côté de l'approfondissement, il s'agit du téléchargement des pages web à partir de celles qui ont été récupérées aux niveaux précédents. La largeur et la profondeur d'aspiration sont configurées dans un fichier de paramètres. Cette aspiration horizontale et verticale optionnelle exige une

série de paramètres qui définissent une URL initiale, les expressions régulières qui indique la modification d'URL en raison du tourne-pages, le nombre de pages à récupérer, le nombre de niveaux d'aspiration et les hyperliens à récupérer à chaque niveau.

Le programme développé pour réaliser le module Récupération est une classe nommée Crawler : elle est composée d'un constructeur, d'un ensemble d'attributs et de trois méthodes : $Page, $Fetch et Affiche. $Page et $Fetch sont deux méthodes privées qui ne peuvent pas être appelées par l'objet à l'extérieur, alors qu'Affiche est une méthode d'objet permettant la communication entre l'intérieur et l'extérieur de l'objet. La méthode $Page est chargée de récupérer les URL associées au tourne-pages. L'URL associée à chaque page connaît souvent une modification sur le segment numérique de l'URL au fil du tourne-pages. Par exemple, l'URL de la première page du forum bricolage est *http://bricolage.linternaute.com/forum/electromenager-14,* celui de la deuxième page est *http://bricolage.linternaute.com/forum/electromenager-14?page=2* et celui de la troisième page est *http://bricolage.linternaute.com/forum/electromenager-14?page=3,* ainsi de suite. La partie *?page=2* surajoutée sur l'URL initiale est appelée le chemin du tourne-pages. L'écart entre les parties numériques dans le chemin du tourne-pages varie avec différents sites. Il peut être de 1, 5, 10, 15, ou même 30, 100, etc. La méthode $Page prend les paramètres suivants : le chemin du tourne-pages, l'écart de modification numérique, le nombre de pages à récupérer et l'URL initiale. $Page récupère tous les URL associées aux pages qui sont tournées en fonction des paramètres saisis s'il trouve un tourne-pages sur le site ; sinon, $Page ne fait que renvoyer l'URL initiale. Dans le dernier cas, le chemin du tourne-pages doit être une chaîne de caractères vide et l'écart ainsi que le nombre de pages doivent tous être configurés sous 0. La méthode $Fetch a pour objet de récupérer tous les hyperliens requis dans les pages web entrées. Cette méthode privée nécessite trois paramètres : une URL de racine, une expression régulière permettant de repérer un hyperlien, un ensemble d'URL associées aux pages web dans lesquelles on recherche l'hyperlien indiqué. Une URL est nécessaire au cas où l'hyperlien enregistré n'est pas une URL complète. Par exemple, les hyperliens associés aux articles contenus dans la page web *http://www.commentcamarche.net/news/* sont enregistrés dans la balise *<a href="/news/5865844-noel-2014-achats-en-ligne-a-bas-prix">* sous forme

incomplète. Cet hyperlien ne peut être complet qu'en ajoutant une URL de racine : *http://www.commentcamarche.net.* La méthode Affiche permet d'approfondir l'aspiration en faisant appel aux méthodes privées $Page et $Fetch à chaque niveau. Elle aspire toutes les pages qui sont tournées, récupère ensuite les hyperliens indiqués dans chaque page obtenue et répète cet ensemble de processus itérativement pour approfondir l'aspiration. La méthode Affiche prend les paramètres saisis dans le fichier de paramètres. Il s'agit du nom de répertoire à créer pour enregistrer les pages web téléchargées, une URL initiale à partir duquel l'aspiration sera effectuée, le nombre de niveaux d'aspiration, les informations sur le tourne-pages et les informations sur les hyperliens à récupérer à chaque niveau.

L'extraction de textes pertinents est réalisée à l'aide du module perl HTML::Parser. C'est un parseur qui analyse un fichier html ligne par ligne. Chaque fois qu'il trouve une paire de balises (balise ouvrante et balise fermante correspondante) et le texte encadré entre eux, il enregistre respectivement le nom de balise ouvrante, les noms d'attributs ainsi que les valeurs d'attributs correspondants (s'il y a des attributs dans la balise ouvrante), le texte encadré entre la balise ouvrante et la balise fermante, et le nom de la balise fermante dans une série de variables prédéfinies. Les informations concernant la balise ouvrante, le texte et la balise fermante peuvent être interrogées respectivement par trois méthodes : start, text et end. Si les informations interrogées sur la balise ouvrante et la balise fermante correspondent à celles qu'on veut, on peut ordonner à la méthode text d'afficher le texte enregistré dans la variable prédéfinie. Le sous-module Extraction de RENE est un programme permettant de repérer les informations que l'on veut à l'aide du module HTML::Parser. Il prend un ensemble de paires de balises comme paramètres. Chaque paire de balises comprend une balise marquant le début du texte qu'on cherche et une balise marquant la fin de ce texte. Quand le parseur trouve les balises paramétrées, la méthode text affiche les textes qui y sont encadrés. Le sous-module Nettoyage est un script perl qui a pour objet d'éliminer les espaces ou les retours de chariot superflus à l'aide d'expressions régulières.

Le troisième module Encodage est en fait intégré dans le module Récupération et le module Extraction&Nettoyage. Dans le module Récupération, les pages web téléchargés sont

encodées en utf8. Dans le sous-module Extraction, les entités html sont décodées à l'aide du module perl HTLM::Entities et les textes extraits à partir des pages web rédigées en utf8 sont encodés en utf8, alors que les textes extraits des pages web rédigées en latin1 sont encodés en latin1 à l'aide du module perl Encode. L'encodage des pages web est détecté par le script en vérifiant l'information d'encodage enregistrée dans l'attribut "charset=".

1.3.3.3. Utilisation

L'outil de constitution de corpus que nous développons est alimenté par deux fichiers de paramètres. Le premier fichier sert au module Récupération et le deuxième est pris par le module Extraction&Nettoyage. Il suffit à l'utilisateur de saisir les paramètres nécessaires en respectant le format requis dans ces deux fichiers de paramètres et de lancer respectivement le programme objet_crawler.pl, extrait_utf8.pl (ou extrait_latin1.pl) et cleaner.pl pour constituer un corpus numérique provenant du web. Chaque programme est lancé à la suite du précédent, car les résultats obtenus dans l'étape précédente serviront de fichiers d'entrée pour le programme suivant. Les difficultés d'utilisation de différents outils proviennent de la saisie de paramètres. Dans ce qui suit, on présentera en détail la construction des deux fichiers de paramètres.

Dans le premier fichier de paramètres, chaque ligne correspond à un site web à aspirer et est composée de quatre colonnes séparées par « ; ». Dans la première colonne, il faut saisir le nom du répertoire à créer pour enregistrer les pages web téléchargées. Dans la deuxième colonne, nous devons fournir une URL initiale à partir duquel l'aspiration sera effectuée. La troisième colonne contient une valeur numérique qui indique le nombre de niveaux d'aspiration. La valeur dans la quatrième colonne est un tableau dans lequel chaque élément est un sous-tableau et est séparé par « # ». À chaque sous-tableau comprend une série d'informations sur le tourne-pages ainsi que les hyperliens d'un niveau d'aspiration. Il y a autant de sous-tableaux dans la quatrième colonne que le nombre de niveaux d'aspiration. Les informations contenues dans chaque sous-tableau sont séparées par « @ » et elles sont respectivement : le chemin du tourne-pages, l'écart du tourne-pages, le nombre de pages à récupérer, une URL de racine et un ensemble d'hyperliens à récupérer à ce niveau. Tous les

hyperliens à récupérer doivent être séparés par « ! » et ils doivent être donnés par des expressions régulières qui permettent de les identifier dans des fichiers html, par exemple, <a class=\"userReviews\"[^><]*?href=\"(http:\/\/www\.ciao\.fr\/Avis[^><]*?)\"> est un paramètre d'hyperlien correspondant au format requis dans RENE. Cette expression régulière sur l'hyperlien à récupérer est encadrée par les parenthèses. Le paramètre sur le chemin du tourne-pages est aussi une expression régulière. C'est une expression régulière divisée en quatre groupes par quatre paires de parenthèses. Le premier groupe permet de trouver dans l'URL initiale une chaîne de caractères marquant le début de l'insertion de la modification sur l'URL initiale pour le tourne-pages et le dernier groupe repère celle qui marque la fin. Le deuxième groupe permet d'ajouter la partie textuelle surajoutée à l'URL initiale dans la modification d'URL et le troisième groupe permet d'ajouter la partie numérique pour la modification d'URL. Le nombre indiqué dans le troisième groupe peut être le numéro de la première page à récupérer ou le numéro de la page à partir de laquelle on veut commencer l'aspiration. Par exemple, pour la page web *http://www.ciao.fr/sr/q-automobile* qui contient un tourne-pages, son paramètre sur le chemin du tourne-pages est configuré comme (.*?automobile)(,p-)(1)(). Les URL associées à la première page (*http://www.ciao.fr/sr/q-automobile,p-1*), à la deuxième page (*http://www.ciao.fr/sr/q-automobile,p-2*), à la troisième page (*http://www.ciao.fr/sr/q-automobile,p-3*), …… seront toutes récupérés à l'aide de cette expression régulière. La hiérarchie des paramètres saisis pour chaque site à chaque ligne peut être représentée par le schéma dans la Figure 20.

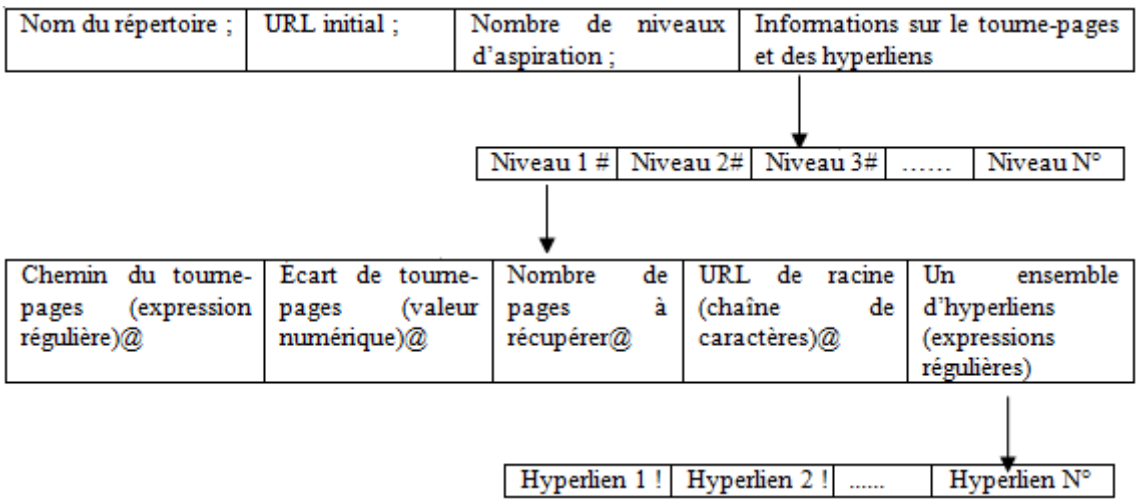

Figure 20 Hiérarchie des paramètres

Dans le deuxième fichier de paramètres, les informations concernant l'extraction de textes pertinents pour différents sites web se trouvent également à différentes lignes. Chaque ligne comprend une série d'informations : le nom du répertoire à partir duquel le programme extrait_utf8.pl (ou extrait_latin1.pl) prend les fichiers d'entrée, une balise qui permet d'identifier le titre du texte et une paire de balises permettant de repérer la partie de textes pertinents. Ces trois éléments sont séparés par le point-virgule. Une paire de balises dans la troisième colonne peut être composée d'une balise ouvrante qui marque le début du texte qu'on cherche et d'une autre balise ouvrante marquant le début d'une autre partie de textes qui n'appartient plus à ce qu'on veut. Cette paire de balises peut également consister en une balise ouvrante et la balise fermante.

2. Méthode

Dans cette section, on présente deux méthodes : la méthode distributionnelle supervisée et la méthode distributionnelle semi-supervisée. La méthode supervisée permet de récupérer un ensemble d'arguments sémantiquement homogènes à partir des prédicats appropriés donnés à l'avance. La méthode semi-supervisée est développée en se fondant sur la méthode précédente. Cette méthode consiste à récupérer plus d'arguments de la classe sémantique donnée à partir d'un ensemble d'arguments donné à l'avance.

2.1. Méthode distributionnelle supervisée

La méthode distributionnelle supervisée consiste à reconnaître automatiquement les noms d'artefacts à partir d'un ensemble de prédicats appropriés. Ces prédicats sont associés aux noms d'artefacts donnés à l'avance. Cette méthode est réalisée par la représentation des distributions syntactico-sémantiques des prédicats appropriés de la classe sémantique : ce sont les noms d'artefacts. Les distributions syntactico-sémantiques des prédicats appropriés sont représentées par les patrons syntaxiques. Dans cette méthode, un prétraitement est d'abord effectué pour augmenter la pertinence d'étiquetage morphosyntaxique. Ensuite, toutes les structures prédicat-argument sont identifiées en se basant sur les patrons syntaxiques à partir

des prédicats appropriés donnés à l'avance. Finalement, les arguments de la classe sémantique sont sélectionnés par l'intersection des arguments.

2.1.1. Prétraitement

L'étiquetage morphosyntaxique est crucial pour étiqueter les informations avec des grammaires locales, car ces grammaires exploitent la catégorie grammaticale des unités linguistiques. L'étiquetage des expressions multi-mots et la désambiguïsation morphosyntaxique sont les deux problématiques incontournables dans l'étiquetage morphosyntaxique. Les éléments constitutifs des expressions multi-mots sont souvent reconnus à tort comme un constituant d'un groupe nominal ou comme un constituant syntaxique de la phrase. L'exploitation des structures prédicat-argument est réalisée à l'aide de patrons syntaxiques. La pertinence de l'étiquetage morphosyntaxique a une influence directe sur la performance de la méthode. Le prétraitement se focalise sur le traitement des expressions multi-mots et la désambiguïsation morphosyntaxique des prédicats. Dans ce qui suit, on présente d'abord le traitement des expressions multi-mots qui a pour objet de minimiser les erreurs d'étiquetage morphosyntaxique. Ensuite, on détaille la méthode de désambiguïsation morphosyntaxique et la méthode d'étiquetage des prédicats.

2.1.1.1. Traitement des expressions multi-mots

Les constituants des expressions multi-mots sont souvent reconnus à tort comme les constituants syntaxiques de la phrase, par exemple, pour identifier tous les groupes nominaux N+A, la séquence *fois acheté* dans la phrase *Une fois acheté mon nouveau manteau, je rentre à la maison* est aussi extraite comme groupe nominal N+A. Un autre exemple, le prédicat *fermer* dans le sens « Faire cesser le fonctionnement d'un appareil à gaz, électrique, etc. » est le prédicat approprié des noms d'artefacts, alors que *les yeux fermés* dans l'expression *faire quelque chose les yeux fermés* peuvent également être mal reconnue comme la structure prédicat-argument par le patron syntaxique N+Vpp (ici, Vpp est *fermé*). Pareillement, dans la phrase *Ces pièces sont par la suite assemblées.*, il est possible que *suite*, élément constitutif de

l'expression multi-mots *par la suite*, soit mal analysé par la machine comme constituant de la structure prédicat-argument Arg+Pred(adjectif) (Ici, le prédicat adjectif renvoie à *assemblé*).

Pour résoudre le problème, on adopte la stratégie suivante : on fait appel au dictionnaire Delac dans Unitex pour étiqueter toutes les expressions multi-mots adjectivaux (par ex., *à but lucratif, à côté, de bas niveau, de bonne foi, en dehors*, etc.), adverbiaux (par ex., *les yeux fermés, une fois, en effet, au fond, à la fin, quand même, quelques temps, de suite, de manière*, etc.), verbaux (*prendre en compte, mettre en place, prendre place, laisser place, faire place,* etc.), prépositionnels (*au bout de, à angle droit avec, à bord de, à l'occasion de,* etc.) et conjonctives (*au moment où, autant que, autrement que, avant que,* etc.) ; les séquences étiquetées sont ensuite remplacées par une étiquette morphosyntaxique correspondante (comme <ADV>, <ADJ>, <V>, <PREP> et <CONJS>). Par exemple, la séquence *une fois acheté* devient *<ADV> acheté* après le remplacement des expressions multi-mots. L'application de ce traitement se limite aux expressions multi-mots adverbiales, adjectivales, verbales, prépositionnelles et conjonctives. Les expressions multi-mots nominales (i.e. les noms composés) ne sont pas traitées, puisqu'elles sont les unités lexicales à identifier par les méthodes développées dans cette thèse.

2.1.1.2. Désambiguïsation morphosyntaxique des prédicats

Les instances de certains prédicats peuvent avoir plusieurs interprétations morphosyntaxiques. Par exemple, *allume* peut-être le présent singulier de la troisième personne du verbe *allumer* et il peut aussi être un constituant d'un nom composé comme *allume-cigare*. Un autre exemple, *cuisine* peut-être un nom dans la phrase *Ces ingrédients sont pour la cuisine d'aujourd'hui* et une forme conjuguée du verbe *cuisiner* dans *Elle cuisine un plat.* De plus, certains prédicats nominaux ont la même forme qu'un argument élémentaire, par exemple, *prise* est un prédicat nominal dans la séquence *prise de conscience*, alors qu'il est un argument élémentaire dans *brancher la prise sur le circuit*. Pour la désambiguïsation morphosyntaxique, nous adoptons la stratégie suivante: pour les unités lexicales dont les parties du discours sont utilisées comme référence pour la désambiguïsation morphosyntaxique des autres unités lexicales, on sélectionne leur partie du discours la plus

utilisée et on supprime leurs autres possibilités morphosyntaxiques dans la ressource linguistique Morfetik transformé en Dela ; ensuite, on construit une série de graphes pour lever l'ambiguïté morphosyntaxique des prédicats appropriés en dépendant du contexte. Le contexte utilisé pour la désambiguïsation morphosyntaxique est un ensemble de codes grammaticaux (i.e. ceux qui indiquent les catégories grammaticales).

Les parties du discours de certaines unités lexicales sont utilisées comme référence pour la désambiguïsation morphosyntaxique des autres unités lexicales dans leurs contextes, telles que les mots grammaticaux (comme les déterminants, les prépositions, les adjectifs non qualificatifs et les conjonctions), les pronoms, certains verbes fréquemment utilisés (tels que *être, avoir, aller, savoir,* etc.) et certains adverbes (tels que *bien, ne, pas, plus*, etc.). Par exemple, pour la forme *lessive,* on décide souvent s'il s'agit d'un nom ou d'un verbe au présent en fonction de la partie du discours des unités lexicales qui la précèdent ou la suivent. Si l'unité lexicale devant *lessive* est un déterminant, *lessive* est plutôt un nom. En revanche, si ce qui suit *lessive* est un groupe nominal (ex., Dét+N+A), la forme *lessive* est plutôt interprétée comme un verbe au présent. Néanmoins, si la catégorie grammaticale des unités lexicales dans le contexte de l'unité lexicale à désambiguïser a aussi plusieurs possibilités d'interprétation, il devient difficile de lever l'ambiguïté morphosyntaxique pour l'unité lexicale ciblée. Par exemple, *a,* du verbe *avoir,* est aussi enregistré comme nom dans le dictionnaire Morfetik, cependant, le graphe pour distinguer le prédicat adjectif du participe passé (après le verbe auxiliaire) dépend de l'occurrence de *a*, verbe auxiliaire, devant les entités à désambiguïser. Si la partie du discours de *a* ne peut pas être décidé, il est difficile de déterminer si l'unité lexicale qui suit *a* est un participe passé (un prédicat verbal) ou un prédicat adjectif. Ainsi, on enlève les possibilités de parties du discours relativement moins utilisées de certaines unités lexicales. Les parties du discours de ces unités lexicales sont souvent utilisées comme référence pour désambiguïser les autres unités lexicales. Le Tableau 5 liste les unités lexicales auxquelles s'applique cette élimination.

Catégorie grammaticale	**Unités lexicales**	**Entrées à enlever**
PREP	*avant, avec, après, chez, contre, dans, devant, depuis, derrière, entre, envers, malgré, sauf, selon, sous, sur, sans, pour, par, pendant, en, à, de, vers, en*	N A ADV
DET	*la, le, un, une, les, deux, trois, …, cent, mille, million, milliard*	N, A
	force	DET
CONJS	*comme*	A
	mais	N, ADV
	que, quand, si,	ADV, N
	soit	ADV
A	*plein, bon, moins, plus, dur, minimum, maximum, mauvais, préféré, super, superbe, nouveau, différent, beau, grand, petit, son, mon, ton, ses, mes, nos, tes, vos, leur, leurs, tout, même, rien, moi, autre, premier, deuxième, second, troisième,……milliardième*	N
ADV	*bien,mal,arrière,derrière,dedans,dehors,intérieur,extérieur*	A, N
	(ne…) pas, (ne…) plus, à priori, à peu près	N
N	*menu, produit, porte*	A
V	*être, est, été, étant, avoir, aller, savoir, allant, faire, devoir, doit, fait, dû, aime, aimé*	N
PRON	*a, ç, c, d, l, m, n, s, t, y, j, il*	N
INTERJ	*combien, comment*	N

Tableau 5 Entrées à enlever pour la désambiguïsation morphosyntaxique

Une série de graphes pour l'identification des différents emplois des prédicats appropriés (emploi verbal, emploi adjectival et emploi nominal) est ensuite établie et la désambiguïsation morphosyntaxique est faite en même temps en se fondant sur le contexte. La désambiguïsation morphosyntaxique sur les prédicats appropriés se focalise sur la désambiguïsation pour le présent et le nom (par ex., la forme *clôture* est-elle un nom ou la conjugaison du verbe *clôturer* au présent de la troisième personne du singulier?), le prédicat verbal au participe passé et le prédicat adjectival (par ex., la forme *innové* dans la séquence *avoir innové* est le participe passé du verbe *innover* et elle est un prédicat adjectival dans la séquence comme N+innové), le participe présent et le nom (par ex., la forme *contenant* peut être un nom d'artefact ou un participe présent du verbe *contenir*), et le participe présent et le prédicat adjectival (par ex., la forme *brûlant* doit-t-elle être comprise comme un adjectif ou un participe présent du verbe *brûler* ?). La désambiguïsation est réalisée par la reconnaissance du contexte morphosyntaxique de chaque prédicat. On établit respectivement pour chaque prédicat un graphe permettant de repérer les syntagmes verbaux passifs (comme *avoir été +ADV+Vpp, se faire+V, aller être +ADV+Vpp, se+être+ADV+Vpp, ……*), un graphe qui reconnaît les syntagmes verbaux actifs (comme *avoir+Vpp, Vpp+Det+N, Vpr+DET+N, ……*),

un graphe qui identifie l'emploi nominal du prédicat et un graphe qui reconnaît l'emploi adjectival du prédicat. Les syntagmes verbaux passifs sont signalés par les étiquettes <Vpassif ></Vpassif>, les syntagmes verbaux actifs, par <V></V>, les prédicats nominaux, par <N ></N > et les prédicats adjectivaux, par <A></A>. Le Tableau 6 liste les structures des syntagmes verbaux à reconnaître par les graphes.

Syntagmes verbaux passifs
(ne/n') + (aller) + avoir + (ADV) + été + (ADV) + Vpp
(ne/n') + (aller) + être + (ADV) + Vpp
(ne/n') + (aller) + se + être + (ADV) + Vpp
(ne/n') + (aller) + se + être + (ADV) + fait (e, es, s) + Vinf
(ne/n') + (aller) + se fait + (ADV) + Vinf
(ne/n') + (aller) + se trouver/devenir/sentir/rester + (ADV) + Vpp
(ne/n') + (aller) + se + V
Syntagmes verbaux actifs
V
avoir + (ADV) + Vpp
en + Vpr + (ADV) + (Det) + N/GN
PRON/ne/N/GN + le/la/les/lui/leur + V

Tableau 6 Structures de syntagmes verbaux

2.1.2. Noms d'artefacts et leurs prédicats appropriés

« Les primitives conceptuelles introduites pour les noms d'artefacts sont celles d'entité artificielle, de production intentionnelle d'objets, de capacité à exercer un rôle dans des actions d'un type donné, de fonction et d'entité fonctionnelle » (Kassel, 2009 : p.121). Cette définition met l'accent sur trois aspects d'un artefact : « la production intentionnelle d'objets », « l'entité artificielle » et « l'entité fonctionnelle ». « La production intentionnelle d'objets » implique que les artefacts sont les objets produits par les actions intentionnelles. « L'entité artificielle » souligne que les objets produits intentionnellement sont des conséquences d'activités humaines. « L'entité fonctionnelle » définit les artefacts comme les objets produits en leur attribuant une fonction spécifique. La sémantique des prédicats appropriés des noms d'artefacts est étroitement liée à ces trois aspects de l'artefact. L'analyse sémantique des prédicats appropriés des noms d'artefacts se déroule en se basant sur cette définition de noms d'artefacts.

Les caractéristiques sémantiques des prédicats appropriés des noms d'artefacts concernent souvent les trois aspects (« la production intentionnelle d'objets », « l'entité artificielle » et « l'entité fonctionnelle ») indiqués dans la définition des noms d'artefacts, car les prédicats appropriés qui prédisent leurs classes sémantiques d'arguments contiennent certainement les concepts sémantiques associés à leurs classes sémantiques d'arguments correspondantes. En fonction de la définition de Kassel (2009) des noms d'artefacts et la prédictibilité appliquées aux classes sémantiques par les prédicats appropriés, nous pouvons avoir trois groupes de prédicats appropriés: les prédicats appropriés du premier groupe décrivent les actions intentionnelles des êtres humains, tels que *fabriquer (fabriquer une excavatrice/un ordinateur/un appareil……), produire (produire un réfrigérateur/un téléviseur/un grille-pain……), emboutir (emboutir une bassine en tôle/une boîte en acier……), creuser (creuser un puits/un tunnel……), articuler (articuler deux tiges/les échelles/ les pièces d'un meuble……), transformer (transformer la matière première en ustensiles, transformer le métal en voiture……*), etc. ; le deuxième groupe recense les prédicats appropriés dont la sémantique est associée aux propriétés des noms d'artefacts, tels que *brancher* (associé à la propriété électronique de certains artefacts, par ex.*, brancher un ordinateur/une lampe/un bouilloire/la prise/le réfrigérateur/le chauffage électrique*……), *garer* (associé à la propriété de parking du véhicule, par ex., *garer ma moto/sa voiture/le camion/un train/une automobile/ un bus……*), *réparer* (associé à la propriété de réparation des artefacts, par ex., *réparer une voiture/une repasseuse/une machine à laver/une paire de chaussures/les tubes/la porte*……), etc. ; les prédicats appropriés du troisième groupe décrivent la fonction des noms d'artefacts, par exemple, *fumer (fumer une cigarette), allumer (allumer une lampe), couper (couper des branches avec les ciseaux), aspirer (l'aspirateur peut aspirer des poussières)*, *lessiver (lessiver des vêtements)*, etc.

Les prédicats appropriés du second groupe (le groupe de prédicats appropriés dont la sémantique est associée aux propriétés des artefacts) peuvent appartenir aussi, parfois, au troisième groupe (dont les prédicats appropriés décrivent la fonction des artefacts). Par exemple, *décorer* peut être considéré comme un prédicat décrivant la fonction des artefacts ou un prédicat dont la sémantique est associée aux propriétés des artefacts. Dans la phrase *La*

poupée japonaise/le cadre de photo/un morceau de ruban peut décorer la maison d'enfant., *décorer* décrit la fonction des artefacts : poupée japonaise, cadre de photo et ruban, alors que dans les séquences *décorer une maison/un bureau/une table*, etc., les artefacts maison, bureau et table possèdent la propriété : « DECORATION ». Le Tableau 7 présente la liste des exemples concernant le classement sémantique des prédicats appropriés des noms d'artefacts. Pour chaque groupe, on liste cinq prédicats et certains exemples correspondants. Dans la deuxième colonne du tableau, on liste le concept sémantique décrit par chaque prédicat.

Prédicats	**Concept sémantique décrit**	**Exemples**
Groupe1	**Action intentionnelle humaine**	
forger *transformer* *fabriquer* *produire* *tremper* ……	FORGEAGE TRANSFORMATION FABRICATION PRODUCTION TREMPE (technologie industrielle) ……	*forger un fer à cheval/une pièce* *transformer le fer en véhicule* *fabriquer une automobile/pince* *produire des jouets d'enfant* *tremper une lame* ……
Groupe2	**Propriété des artefacts**	
démarrer *réparer* *garer* *brancher* *tirer* ……	DÉMARRAGE RÉPARATION PARKING ÉLECTRONIQUE TIRAGE ……	*Démarrer le moteur/la voiture* *réparer une montre/la route* *garer le bus/un camion* *brancher l'ordinateur/la bouilloire* *tirer une corde/la sonnette d'alarme* ……
Groupe3	**Fonction**	
maquiller *afficher* *nettoyer* *coudre* *abouter* ……	MAQUILLAGE AFFICHAGE NETTOYAGE COUTURE ABOUTEMENT ……	*maquiller avec bb crème/pinceaux* *afficher sur l'écran/le panneau* *nettoyer avec essuie-tout/balai* *coudre des pantalons/un sac* *abouter les câbles/les panneaux* ……

Tableau 7 Classement sémantique des prédicats appropriés des noms d'artefacts

2.1.3. Distributions syntactico-sémantiques des prédicats appropriés

Les prédicats appropriés des noms d'artefacts peuvent être divisés en quatre classes en fonction de leurs distributions syntactico-sémantiques. La Classe1 regroupe les prédicats dont le complément d'objet (direct ou indirect) est toujours un nom d'artefact, par exemple, *éteindre*, dans son sens « Faire cesser le fonctionnement d'un appareil à gaz, électrique, etc », son complément d'objet direct est toujours un nom d'artefact ; *jouer*, dans le sens « s'amuser avec un jeu ou jouet », son complément d'objet indirect est aussi toujours un nom d'artefact. La distribution syntactico-sémantique des prédicats appropriés de la Classe1 peut être représentée par la structure V+C.O.(NAF) (C.O. désigne le complément d'objet et NAF

signifie le nom d'artefact). La Classe2 comprend les prédicats qui sont souvent utilisés avec un complément de moyen formé à partir d'un nom d'artefact. Néanmoins, le complément d'objet des prédicats de la Classe2 n'est toujours pas un nom d'artefact. Il peut être un nom d'artefact ou un nom d'une autre classe sémantique. Les prédicats de la Classe2 ont un champ sémantique plus large à la position de complément d'objet direct, par exemple, pour le prédicat approprié *découper,* son complément de moyen est toujours un nom d'artefact comme *découper avec des ciseaux, découper avec le couteau*……, mais son complément d'objet direct peut être un nom d'artefact ou un nom d'autres types comme *découper les papiers/les branches/les pierres*….. La distribution syntactico-sémantique de la Classe2 peut être représentée par le patron V+C.O.D.(NAF ou Nc/NAF)+C.M.(Prep+NAF) (Nc désigne le nom des autres classes sémantiques, C.M. signifie le complément de moyen et Prep réfère à la préposition). La distribution syntaxique de la Classe3 correspond à la forme suivante : V+C.O.D.(NAF/Nc)+C.C.(Prep+NAF) (C.C. désigne le complément circonstanciel de lieu).

Les prédicats de la Classe3 sont souvent utilisés ensemble avec un complément de lieu nominal introduit par une préposition et le C.C. des prédicats de la Classe3 est souvent un nom d'artefact, par exemple, *ranger* (*Luc range les bouquins dans les étagères), coincer (un dossier est coincé entre des livres), coller (coller un timbre sur l'enveloppe),* etc. Pour les prédicats de la Classe4, leurs compléments d'objet ont un champ sémantique large. Par rapport à ceux des prédicats de la Classe3, les compléments d'objet des prédicats de la Classe4 ont moins de possibilités d'être des noms d'artefacts. Les prédicats de la Classe 4 apparaissent souvent avec un complément de lieu ou un complément de moyen qui est souvent un nom d'artefact. La distribution syntactico-sémantique de la Classe 4 est représentée comme ce qui suit : V+C.O.D.(Nc)+C.M./C.C.(Prep+NAF), par exemple, *transvaser->transvaser du vin dans un gobelet/verre, afficher->afficher une annonce sur un panneau/écran, maquiller->je me maquille avec de la poudre,* etc.

Pour chaque classe, on distingue aussi certaines sous-classes en fonction des propriétés syntaxico-sémantiques des prédicats de la classe. Les prédicats de la Classe1 sont sous-divisés en trois groupes : la Classe_1a, la Classe_1b et la Classe_1c. Le complément d'objet des prédicats de Classe_1a est le complément d'objet direct (par ex., *inventer une machine de*

cuisine, éteindre l'ordinateur, renouveler sa garde-robe, etc.), alors que celui des prédicats de la Classe_1b et de la Classe_1c est le complément d'objet indirect (par ex., jouer aux cartes/jeux vidéo……). Le complément d'objet indirect des prédicats de la Classe_1b est introduit par les prépositions comme *derrière, dessus, dessous, devant, sur*, etc., par exemple, *tirer sur le tiroir, appuyer sur le bouton*, …… Le complément d'objet indirect des prédicats de la Classe_1c est introduit par la préposition *à* ou *de*, par exemple, *jouer au foot*. La Classe2 a aussi trois sous-classes : la Classe_2a, la Classe_2b et la Classe_2c. Les compléments d'objet direct des prédicats de la Classe_2a sont toujours les noms d'artefacts, alors que ceux des prédicats de la Classe_2b et de la Classe_2c peuvent être à la fois les noms d'artefacts et les noms d'autres types. Les prépositions qui introduisent le complément de moyen des prédicats de la Classe_2b peuvent être *de, par* ou *avec* (par ex., *décrasser la casserole avec du vinaigre, bourrer un coussin des coupons*, etc.), alors que les prépositions qui introduisent le complément de moyen des prédicats de la Classe_2c ne peuvent être que *de,* par exemple, *équiper de matières modernes, orner une façade de drapeaux*, etc.

Pareillement pour la Classe3, il y a également trois sous-classes : la Classe_3a, la Class_3b et la Classe_3c. Le nom complément introduit par la préposition des prédicats de la Classe_3a est le complément de lieu (par ex., *déplacer les vêtements dans l'armoire, fixer l'annonce sur le mur, enfoncer un clou dans le panneau,* etc.), celui des prédicats de la Classe_3b et de la Classe_3c est le complément d'objet indirect (par ex., *connecter à l'internet, transformer en véhicule*, etc.). La Classe4 est divisé en quatre sous-classes : la Classe_4a, la Classe_4b, Classe_4c et Classe_4d. Les compléments d'objet directs de la Classe_4a et la Classe_4c peuvent être les noms d'artefacts ou les noms d'autres types, alors que ceux de la Classe_4b et de la Classe_4d sont toujours les noms d'autres types. Le nom complément introduit par la préposition des prédicats de la Classe_4a et de la Classe_4d est un complément circonstanciel de lieu (par ex., *placarder un avis sur le mur, transvaser du vin dans un verre*, etc.), mais celui de la Classe_4b et de la Classe_4c est un complément de moyen (par ex., *coiffer avec un lisseur, peigner avec un peigne, maquiller de rouge à lèvre,* etc.).

Dans le Tableau 8, nous établissons une liste de certains prédicats fournis en exemples pour chaque classe syntactico-sémantique. Les structures syntactico-sémantiques sont données sous la forme active. Dans l'analyse des distributions syntactico-sémantiques des prédicats appropriés, la liste de prédicats recensés et étudiés comprend certains prédicats qui ne sont pas les prédicats appropriés dans un sens rigoureux. On les accepte parmi les prédicats appropriés dans le but d'augmenter le rappel du résultat.

Information / **Classes**	**Prédicats appropriés**	**Distribution syntactico-sémantique**
Classe1		
Classe_1a	*éteindre*, *renouveler*, *inventer*, etc.	V+NAF
Classe_1b	*tirer*, *retirer*, *appuyer*, etc.	V+dessus/dessous/derrière/devant+NAF
Classe_1c	*jouer*	V+à/de (+Det)+NAF
Classe2		
Classe_2a	*récurer*, *réparer*, *tracter*, etc.	V+NAF+de/avec/par+NAF
Classe_2b	*découper*, *fouiller*, *décrasser*, etc.	V+NAF/Nc+de/avec/par+NAF
Classe_2c	*équiper*, *orner*, etc.	V+NAF/Nc+de+NAF
Classe3		
Classe_3a	*ranger*, *installer*, *contenir*, etc.	V+NAF/Nc+sous/sur/devant/derrière/au-dessus de/au-dessous de/ à la droite de…+NAF
Classe_3b	*transformer*	V+NAF/Nc+en+NAF
Classe_3c	*connecter*	V+NAF/Nc+à+NAF
Classe4		
Classe_4a	*transvaser*, *verser*, *enregistrer*, etc.	V+NAF/Nc+dans+NAF
Classe_4b	*peigner*, *maquiller*, *farder*, *coiffer*, etc.	V+Nc+avec/par/de+NAF
Classe_4c	*nourrir*, *alimenter*	V+Nc+à+NAF
Classe_4d	*afficher*, *placarder*, etc.	V+Nc+sur+NAF

Tableau 8 Classement des distributions syntactico-sémantiques des prédicats appropriés

2.1.4. Extraction automatique des structures prédicat-argument

L'extraction automatique des structures prédicat-argument est réalisée à l'aide des grammaires locales établies sur la plateforme Unitex. On tente de représenter respectivement les structures de syntagmes verbaux, les structures des groups nominaux et finalement les structures prédicat-argument par les patrons syntaxiques. La représentation des structures de syntagmes verbaux a été effectuée dans l'étape de désambiguïsation morphosynatxique des

prédicats (cf., Tableau 6, p. 139). Dans ce qui suit, on présente la construction des patrons syntaxiques adaptés aux groupes nominaux et aux structures prédicat-argument.

D'après Gross M. (1986), le groupe nominal (GN) est défini par la forme de base *Dét+N+Modifieur*. Les déterminants sont de deux sortes : définis et indéfinis. Les déterminants indéfinis sont sous-divisés en quatre classes en fonction des critères convenant aux trois propriétés syntaxiques : Dét peut se combiner directement avec N (cette propriété est notée comme Dét N, par ex, *j'ai lu chaque article*) ; Dét peut se combiner avec GN au moyen de la préposition *de* (notée comme Dét de GN, par ex, *beaucoup de mes amis*) ; Dét peut fonctionner comme adverbe (propriété notée comme N0 V Dét, par ex, N0 V Dét, *Luc dort beaucoup*). Les quatre classes sont les suivantes : Dadv qui comprend les Dét ayant les propriétés Dét de GN et N0 V Dét mais exclut la propriété Dét N ; Dadj qui est caractérisé par la propriété Dét N et l'absence de N0 V Dét ; Dnom qui comprend la propriété Dét de GN mais n'ayant pas la propriété Dét N ; N0 V Dét et Préd qui sont les prédéterminants ayant la propriété Dét GN mais ni la propriété Dét N, ni Dét de GN (cf., Tableau 9). En ce qui concerne le modifieur, il peut être un adjectif (par ex., *joli*), un adjectif modifié par l'adverbe (par ex., *très joli*), une série d'adjectifs (par ex., *un clou hexagonal rouge*) ou une proposition relative (par ex., *la poupée que je lui ai donnée*). Le modifieur peut se déplacer entre Dét et N (par ex., *une très jolie femme)* et un groupe nominal peut comprendre un autre GN comme composant (par ex., GN1 : ces petits insectes -> GN2 : *Beaucoup de ces petits insectes*). La structure de base de groupe nominal est représentée par les deux patrons syntaxiques suivants : Dét de (N/GN) et Dét (N/GN). La classification des déterminants de Gross M. (1986) se base complètement sur des critères formels sans tenir compte de leurs définitions traditionnelles. Néanmoins, cette description des déterminants et les groupes nominaux nous permet de construire une série de grammaires locales de la manière la plus exhaustive possible.

Classes de déterminants	Propriétés syntaxiques	Exemples
Déterminants définis		
		le, la, les, etc.
Déterminants indéfinis		
Dadj (Déterminants adjectivaux)	Dét N *N0 V Dét	*Il a lu chaque article.* **Luc dort chaque*
Dadv (Déterminants adverbiaux)	Dét de GN N0 V Dét *Dét N	*Il a acheté beaucoup de pommes.* *Il aime beaucoup ce roman.* **Il a mangé beaucoup gâteau.*
Dnom (Déterminants nominaux)	Dét de GN *Dét N *N0 V Dét	*Il y a plein de problèmes.* **L'industrie fabrique plein jouets.* **Il pleut plein.*
Préd (Prédéterminants)	Dét GN *Dét N *Dét de GN	*tous les jeunes amateurs de foot.* **tous amateurs de foot* **tous d'amateurs de foot*

Tableau 9 Classification des déterminants de Gross M. (1986)

Les structures prédicat-argument sont représentées par une série de patrons syntaxiques de base (par ex., V+NAF, V+à/de (+Det)+NAF, V+Nc+à+NAF, etc.) dans les analyses de distributions syntactico-sémantiques des prédicats (cf., Tableau 8, p. 144). On essaie de décrire les autres constructions syntaxiques observées par une série d'opérations. Par exemple,

35) *Le garagiste répare la voiture de Luc (Nhum+V+NAF)*

36) *La voiture de Luc est réparée par le garagiste* (NAF+être+Vpassif+par+Nhum)

37) *Luc a fait réparer sa voiture par le garagiste* (Nhum+faire+V+NAF+par+Nhum)

On utilise la règle Nhum+V+NAF <-> NAF+être+Vpassif+par+Nhum/ Nhum+faire+V+NAF+par+Nhum pour relier ces trois structures syntaxiques. Cette règle consiste plutôt à désigner une opération (i.e. la transformation de la forme passive en forme active ou la transformation de la forme active en forme passive) permettant de décrire la relation entre ces structures syntaxiques. Un autre exemple,

38) *Luc est en train de battre du beurre.* (Nhum+V+NAF)

39) *Luc, chef de cuisine de l'Élysée, est en train de battre du beurre.*
(Nhum+Apposition+V+NAF)

40) *Luc, chef de cuisine de l'Élysée, est en train de battre soigneusement du beurre.*
(Nhum+Apposition+V+ADV+NAF)

On représente la relation entre ces structures syntaxiques par les opérations suivantes : ← Apposition (Nhum+ V ← Apposition +NAF -> Nhum+Apposition+V+NAF) et ←ADV (Nhum+Apposition+ V ←ADV +NAF -> Nhum+Apposition+V+ADV+NAF). Dans le texte suivant :

41) *–Est-ce que tu as nettoyé la table ?* ($PRON_1$+V+NAF)
–Oui, je l'ai nettoyée. ($PRON_1$+$PRON_2$+V)

le complément d'objet direct de *nettoyer* est remplacé par le pronom *l'* dans la réponse, mais, en fonction du contexte, il est quand même assuré que le complément d'objet omis réfère à *la table*. On représente la relation entre ces deux constructions par l'opération ← Pronominalisation ($PRON_1$+V+NAF -> $PRON_1$+$PRON_2$+V). De plus, le complément d'objet de moyen (C.M.) est aussi souvent omis, par exemple,

42) Le professeur découpe des papiers avec une paire de ciseaux.
(N+V+NAF+avec+NAF)
43) Le professeur découpe des papiers. (N+V+NAF)

On lie ces deux structures par l'opération ←Omission (N+V+NAF+C.M. ←Omission -> N+V+NAF).

Une opération désigne la relation entre deux structures syntaxiques. Cette relation peut également exister entre deux autres structures bien différentes, par exemple, l'opération ← Transformation à la forme passive existe entre la structure V+NAF et NAF+être+Vpassif (V+NAF -> NAF+être +Vpassif) et aussi entre V+Nc+à+NAF et Nc+être+Vpassif+à+NAF (V+Nc+à+NAF->Nc+être+Vpassif+à+NAF). Si l'on constate une relation entre deux structures syntaxiques (V+NAF et NAF+être+Vpassif) et la représente par une opération, on peut appliquer cette opération à une autre structure (V+Nc+à+NAF) pour obtenir une nouvelle (Nc+être+Vpassif+à+NAF). Cette façon d'analyser permet de recenser plus de patrons syntaxiques (appelés **patrons syntaxiques dérivés**) d'une manière la plus exhaustive possible à partir d'un ensemble de patrons syntaxiques de base (cf. Annexe 1, section 1). Par exemple,

44) V+NAF -> V+ADV+NAF
-> PrédN+de+(Dét)+NAF
-> NAF+Vpp
… …

45) V+NAF+de/avec/par+NAF -> NAF+Vpp+de/avec/par+NAF
-> V+ADV+NAF+ de/avec/par+NAF
-> NAF, NAF+et/ou+NAF, qui+ pouvoir/devoir+ Vpassif +de/avec/par+NAF
… …

Néanmoins, la longueur (le nombre de tokens) d'une apposition ou d'une proposition relative est difficile à décider, parce que les appositions ou les propositions relatives peuvent comprendre des autres modifieurs. Les appositions, les modifieurs qui sont des propositions relatives et les modifieurs de plus de trois tokens ne sont pas pris en compte dans la construction des patrons syntaxiques pour les groupes nominaux et les structures prédicat-argument. Dans le prétraitement, les prédicats pré-donnés sont déjà étiquetés. Ensuite, on construit une série de graphes qui représentent les patrons syntaxiques établis en fonction des distributions syntactico-sémantiques afin d'identifier les structures prédicat-argument. Et puis, on enregistre le résultat d'étiquetage des structures prédicat-argument au format texte, transforme la case de toutes les lettres en minuscule et fait intervenir TreeTagger pour une lemmatisation. Finalement, on développe le script extrait_intersection.pl pour extraire toutes les structures prédicat-argument étiquetées.

2.1.5. Intersection pour obtenir une classe sémantique d'arguments

L'intersection effectuée sur les arguments a pour objet d'identifier un ensemble d'arguments communs de l'ensemble des prédicats appropriés définitionnels. Plus un argument est partagé par les prédicats, plus il est vraisemblable qu'il appartienne à la classe sémantique définie par ces prédicats. Le nombre de différents prédicats qui coexistent avec un argument est noté comme la fréquence d'intersection de cet argument. On sélectionne les

arguments les plus partagés par les prédicats donnés à l'avance comme les arguments de la classe sémantique définie par ces prédicats. On décide un seuil pour la fréquence d'intersection des arguments. Les arguments dont la fréquence d'intersection dépasse le seuil sont considérés comme les arguments les plus partagés par les prédicats. La Figure 21 est la capture d'écran d'un schéma qui représente le processus d'intersection des arguments. Dans ce schéma, on donne un exemple de quatre prédicats appropriés (Préd1, Préd2, Préd3 et Préd4). Chacun de ces prédicats a ses propres classes sémantiques. Les arguments qui coexistent avec ces prédicats sont différents. On voit bien que la classe sémantique partagée par ces quatre prédicats est la classe sémantique2 qui comprend un ensemble d'arguments. Ainsi, l'ensemble d'arguments obtenu par l'opération d'intersection constituera les arguments de la classe sémantique2 (à savoir Arg2, Arg3, Arg8, Arg17……). De plus, la classe sémantique1 est aussi une classe sémantique très partagée par les quatre prédicats. Les arguments de la classe sémantique1 seront également obtenus du fait de l'intersection des arguments. Dans le schéma, les parties en gris représentent respectivement la classe sémantique et les arguments de la classe sémantique communs de l'ensemble de prédicats appropriés définitionnels Préd1, Préd2, Préd3 et Préd4.

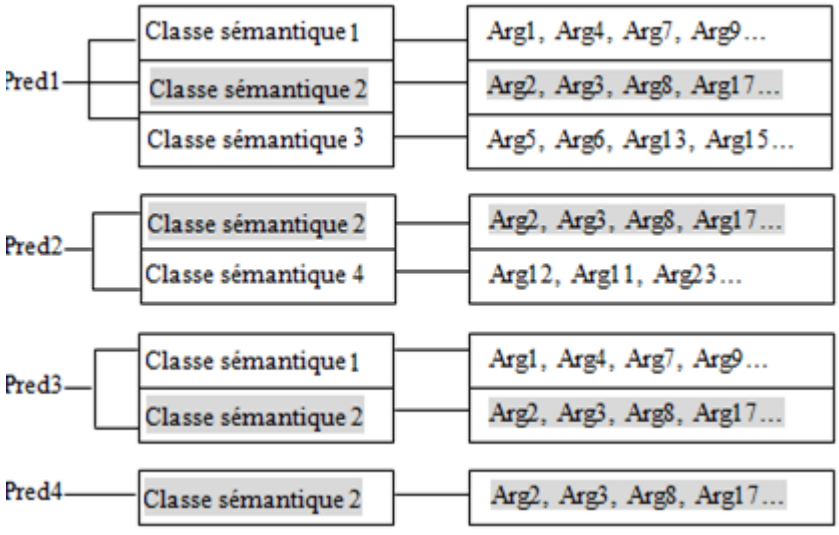

Figure 21 Intersection des arguments

À partir du résultat obtenu d'extraction des structures prédicat-argument, on trouve le prédicat, les arguments associés et le patron syntaxique du prédicat enregistré dans l'étiquette associée à chaque structure prédicat-argument extraite. Ces informations sont mises ensuite

dans un tableau dont chaque ligne commence par un prédicat qui est suivi de son patron syntaxique et de ses arguments associés. Ce processus est effectué par le script perl extrait_intersection.pl. Le résultat est montré dans la Figure 22.

```
V:nourri:i_p:de;adrénaline:1
V:remaniement:t_naf_p:;avis:1
V:affichage:i_p:sur;site:1;écran:1;tableau:1
V:eteint:t_naf:;navigation:1
V:pose:t_gn_p::sur;feuille:1;verrou:3;enregistreur:1;jante:2;autoradio:1;bande:1;parpaing:1;
V:ouverture:t_naf_p::de;intendo:1;trémie:18;réfrigérateur:1;cave:1;vitre:1;fil:1;capot:1;tra
V:plonger:t_gn_p::dans;jeu:1;thermo:1;ballon:1;verre:1;qualité:1
V:rangement:t_gn_p:dans::sous:sur;capacité:2;siège:2;endroit:1;appoint:1;siege:1;longueur:1;
V:tirer:t_naf:;fleur:1;il:1;photoshop:1;rivière:1;lecteur:1;string:1;image:2;sucre:1;étude:1
V:inventer:t_naf:;photo:1;ruban:1
V:cuit|cuite:t_naf_p:;réponse:1;génois:1
V:retirer:t_naf:;bouton:1;capuchon:2;electrovanne:2;point:1;battitures:1;avoir:1
V:essuyer:t_naf_p:de;flaque:1
V:essuiyage:t_naf_p:de;vitre:1
V:fabriquer:t_naf_p::de:avec:par;bouteille:1;pièce:2;matière:1;lcatel:1;samsung:1;reste:1;ch
V:cuisson:t_naf_p::de;crème:1;mixture:2;pièce:1;frite:1;sucre:1;fleureter:1;pâte:1;viande:1;
V:remplier:t_naf_p:de;humidité:1
V:ouvrer|ouvrir:t_naf_p::de:par;toit:15;hazard:1;voiture:1;capot:1;axe:1
V:produire:t_naf_p::de:par;toxine:1;w:1;luxe:1;page:2;vapeur:3;voiture:1;rinçage:1;bad:1;inj
V:renouveler:t_naf:;expérience:1
V:démarrage:t naf p:;calculateur|calculatrice:1;saxo:1;voiture:1;programme:1
```

Figure 22 Résultats d'intersection des arguments

Ensuite, on fait une intersection sur les arguments pour trouver les arguments de la classe sémantique définie par l'ensemble des prédicats donnés à l'avance. Pour cela, on procède comme suit : pour chaque argument, on assigne le score 0 au début et chaque fois qu'il apparaît dans une paire prédicat-argument différente, on incrémente son score par 1. Ce score permet de visualiser le nombre de prédicats partageant cet argument. On considère ce score final comme la fréquence d'intersection de cet argument. Les noms d'artefacts dont les fréquences d'intersection dépassent le seuil défini sont sélectionnés comme noms d'artefacts appartenant à l'ensemble des prédicats appropriés donnés à l'avance. La Figure 23 est la capture d'écran d'une partie du résultat d'intersection. Le chiffre dans la deuxième colonne désigne la fréquence d'intersection de l'argument et celui dans la troisième colonne réfère à la fréquence d'occurrence du nom d'artefact dans les structures prédicat-argument extraites.

```
voiture:28;104
four:22;77
eau:21;71
moteur:20;58
machine:19;48
carte:19;44
porte:19;103
fichier:19;31
pc:18;24
pièce:17;45
ordinateur:17;39
écran:15;25
véhicule:14;18
intérieur:14;17
moule:14;48
appareil:14;16
auto:13;21
tuyau:13;17
chose:13;26
pâte:13;47
plaque:13;39
vitre:12;13
crème:12;25
fenêtre:12;24
```

Figure 23 Résultat d'intersection

2.2. Méthode distributionnelle semi-supervisée

En se fondant sur la méthode distributionnelle supervisée, on développe une méthode distributionnelle semi-supervisée permettant l'apprentissage automatique des prédicats appropriés et de leurs patrons syntaxiques dans le corpus à partir d'un ensemble de noms d'artefacts donnés à l'avance. L'apprentissage automatique des patrons syntaxiques des prédicats appropriés est nécessaire en raison de l'impossibilité de prévoir les distributions syntactico-sémantiques des prédicats à partir des arguments. Nous essayons de prédire toutes les possibilités de patrons syntaxiques que les prédicats appropriés des noms d'artefacts peuvent avoir et nous effectuons ensuite un calcul probabiliste pour sélectionner le patron syntaxique adéquat de chaque prédicat. Avec les prédicats appropriés et leurs patrons syntaxiques appris, la méthode distributionnelle supervisée est appliquée encore une fois pour obtenir plus de noms d'artefacts. Nous pouvons répéter cet ensemble de processus itérativement jusqu'à ce qu'on ne trouve plus de nouveaux noms d'artefacts dans le corpus. La méthode distributionnelle semi-supervisée est composée de quatre étapes : l'extraction

automatique des structures prédicats-arguments, l'intersection des prédicats pour obtenir les prédicats de la classe d'arguments, le calcul probabiliste des patrons syntaxiques et l'application de la méthode distributionnelle supervisée. Dans ce qui suit, on présente en détail cette méthode.

2.2.1. Extraction automatique des structures prédicat-argument

Dans la méthode supervisée, on a présenté le classement de la distribution syntactico-sémantique des prédicats appropriés. Il s'agit d'une relation entre la couche syntaxique et la couche sémantique de la langue. Néanmoins, ce classement est effectué du point de vue de l'apprentissage supervisé, à savoir qu'il est fait à la condition que les prédicats appropriés soient déjà donnés à l'avance et les distributions syntactico-sémantiques soient prédictibles à partir de ces prédicats. Cependant, la méthode semi-supervisée consiste à identifier automatiquement les prédicats appropriés à partir d'un ensemble d'arguments d'une classe sémantique donnée à l'avance afin de récupérer plus d'arguments de cette classe sémantique. Dans le cadre de la méthode semi-supervisée, les distributions syntactico-sémantiques ne sont pas prédictibles à partir des arguments donnés à l'avance. Or, si l'on possède un nombre limité de possibilités de patrons syntaxiques des prédicats, on peut sélectionner le patron syntaxique adéquat de chaque prédicat parmi ces possibilités à l'aide d'un calcul probabiliste. Ces possibilités de patrons syntaxiques peuvent être prédites à partir des patrons syntaxiques établis dans la méthode supervisée en s'appuyant sur l'étude de la position de la distribution d'arguments.

Pour identifier automatiquement une structure prédicat-argument avec un argument donné à l'avance, le patron syntactico-sémantique établi dans l'apprentissage semi-supervisé doit respecter le critère suivant : le patron syntaxique doit comprendre au moins une position syntaxique où se trouve toujours un nom d'artefact. Nous listons de nouveau les treize classes de patrons syntaxiques établis dans la méthode supervisée : V+NAF, V+dessus/dessous/derrière/devant+NAF, V+à/de(+Det)+NAF, V+de/avec/par+NAF, V+NAF/Nc+de/avec/par+NAF, V+NAF/Nc+de+NAF, V+NAF/Nc+sous/sur/devant/derrière/au-dessus de/au-dessous de/à la droite de…+NAF,

V+NAF/Nc+en+NAF, V+NAF/Nc+à+NAF, V+NAF/Nc+dans+NAF, V+Nc+avec/par/de+NAF, V+Nc+à+NAF et V+Nc+sur+NAF. On se rend compte que la distribution d'arguments se situe souvent à la position du nom complément sans préposition (complément d'objet direct) ou à la position du nom complément introduite par une préposition (complément d'objet indirect, complément circonstanciel de lieu et complément circonstanciel de moyen). Tous les patrons syntaxiques établis dans la méthode supervisée concernent plutôt trois fonctions syntaxiques : verbe, nom complément sans préposition et nom complément introduit par une préposition. En fonction du critère pour l'établissement des patrons syntaxiques dans la méthode semi-supervisée, on peut avoir quatre possibilités de combinaisons avec les trois fonctions syntaxiques (verbe, nom complément sans préposition et nom complément introduit par une préposition) pour former les patrons syntaxiques désirés : V+NAF, V+NAF+prep+NAF, V+Nc+prep+NAF et V+prep+NAF (cf., Annexe 1, section 2).

De la même manière, on essaie de recenser d'autres patrons syntaxiques (patrons syntaxiques dérivés) à partir des quatre patrons syntaxiques de base prédits (par exemple, V+Nc+prep+NAF ->V+Nc, prep+NAF/être+Vpassif+prep+NAF/V+prep+NAF……, V+NAF+prep+NAF -> NAF+Vpp+prep+NAF/ NAF+être+Vpassif+prep+NAF V+ADV+NAF,avec+NAF……, ect.) dans le but de représenter les principales structures syntaxiques que les prédicats des noms d'artefacts peuvent avoir, Ensuite, on construit un ensemble de graphes en fonction des patrons syntaxiques établis pour étiqueter les structures prédicat-argument.

Dans le cadre de la méthode semi-supervisée, les arguments fournis à l'avance sont étiquetés par les étiquettes <NAF></NAF>. Les structures prédicat-argument ont les étiquettes <BLOC></BLOC> en ajoutant un attribut (tel que s=Vactif_GNAF, s=Vactif_GN_prep_GNAF, s=GNAF+Va+prep+GNAF, etc.) dans la balise ouvrante <BLOC>. Ces attributs indiquent les patrons syntaxiques des prédicats. GNAF désigne le groupe nominal de noms d'artefacts et GN indique le groupe nominal d'autres classes sémantiques. Les codes Vactif, Vpassif et Va représentent respectivement la forme active, la

forme passive et l'emploi du prédicat adjectival. Les prédicats sont étiquetés par les étiquettes <PVactif></PVactif>, <PVpassif></PVpassif> ou <PA></PA>. Les groupes nominaux de noms d'artefacts ont les étiquettes <O></O>. Le Tableau 10 liste toutes les étiquettes utilisées pour l'identification automatique des structures prédicat-argument dans la méthode distributionnelle semi-supervisée.

Patrons de base	**Séquences**	**Prédicats**	**GN de NAF**	**NAF**
V+NAF				
	<BLOC s=Vactif_GNAF></BLOC>	<PVactif></PVactif>	<O></O>	<NAF></NAF>
	<BLOC s=GNAF_Vpassif></BLOC>	<PVpassif></PVpassif>	<O></O>	<NAF></NAF>
	<BLOC s=GNAF_Va></BLOC>	<PA></PA>	<O></O>	<NAF></NAF>
V+NAF+prep+NAF				
	<BLOC s=Vactif_GNAF_prep_GNAF></BLOC>	<PVactif></PVactif>	<O></O>	<NAF></NAF>
	<BLOC s=GNAF_Vpassif_prep_GNAF></BLOC>	<PVpassif></PVpassif>	<O></O>	<NAF></NAF>
	<BLOC s=GNAF_Va_prep_GNAF></BLOC>	<PA></PA>	<O></O>	<NAF></NAF>
	<BLOC s=Vactif_prep_GNAF></BLOC>	<PVactif></PVactif>	<O></O>	<NAF></NAF>
	<BLOC s=Vactif_prep_GNAF_GNAF></BLOC>	<PVactif></PVactif>	<O></O>	<NAF></NAF>
	<BLOC s=GNAF_Vpassif></BLOC>	<PVpassif></PVpassif>	<O></O>	<NAF></NAF>
	<BLOC s=GNAF_Va></BLOC>	<PA></PA>	<O></O>	<NAF></NAF>
V+Nc+prep+NAF				
	<BLOC s=Vactif_GN_prep_GNAF></BLOC>	<PVactif></PVactif>	<O></O>	<NAF></NAF>
	<BLOC s=Vpassif_prep_GNAF></BLOC>	<PVpassif></PVpassif>	<O></O>	<NAF></NAF>
	<BLOC s=GN_Va_prep_GNAF></BLOC>	<PA></PA>	<O></O>	<NAF></NAF>
	<BLOC s=Vactif_prep_GNAF></BLOC>	<PVactif></PVactif>	<O></O>	<NAF></NAF>
V+prep+NAF				
	<BLOC s=Vactif_prep_GNAF></BLOC>	<PVactif></PVactif>	<O></O>	<NAF></NAF>

Tableau 10 Étiquettes utilisées pour étiqueter les structures prédicat-argument dans la méthode semi-supervisée

Le résultat d'étiquetage des structures prédicat-argument est enregistré au format texte brut. Pareillement, on transforme la case des lettres en minuscule et on fait intervenir TreeTagger pour une lemmatisation. La Figure 24 est la capture d'écran d'une partie du résultat d'extraction des structures prédicat-argument dans la méthode semi-supervisée.

```
bloc s=gnaf_va> ton;DET:POS;ton <o> <naf> compartiment;NOM;compartiment </naf> moteur;ADJ;moteur ,;PUN;, </o>
bloc s=vactif_gnaf> <pvactif> a;VER:pres;avoir </pvactif> <o> la;DET:ART;le <naf> valise;NOM;valise </naf> </o
bloc s=vactif_gnaf> <pvactif> ouvrir;VER:infi;ouvrir </pvactif> <o> le;DET:ART;le <naf> coffre;NOM;coffre </na
bloc s=vactif_gnaf> <pvactif> ouvrir;VER:infi;ouvrir </pvactif> <o> le;DET:ART;le <naf> coffre;NOM;coffre </na
bloc s=gnaf_va> pour;PRP;pour <o> <naf> coffre;NOM;coffre </naf> ouvert;VER:pper;ouvrir </o> (;PUN;( <pa> titr
bloc s=vactif_gnaf> <pvactif> maintenir;VER:infi;maintenir </pvactif> <o> le;DET:ART;le <naf> coffre;NOM;coffr
bloc s=vactif_gnaf> <pvactif> module;NOM;module </pvactif> <o> une;DET:ART;un petite;ADJ;petit <naf> plaque;VE
bloc s=gnaf_va> les;DET:ART;le <o> <naf> housses;NOM;housse </naf> </o> <pa> adaptée;VER:pper;adapter </pa> </
bloc s=vactif_gnaf> <pvactif> visser;VER:infi;visser </pvactif> <o> les;DET:ART;le <naf> plaques;NOM;plaque </
bloc s=vactif_gnaf> <pvactif> construire;VER:infi;construire </pvactif> <o> un;DET:ART;un petit;ADJ;petit <naf
bloc s=vactif_gnaf> <pvactif> fait;VER:pres;faire </pvactif> <o> les;DET:ART;le <naf> housses;NOM;housse </naf
bloc s=vactif_gnaf> <pvactif> pendre;VER:infi;pendre </pvactif> <o> une;DET:ART;un <naf> plaque;NOM;plaque </n
bloc s=vactif_gnaf> <pvactif> trouver;VER:infi;trouver </pvactif> <o> une;DET:ART;un <naf> plaque;NOM;plaque <
bloc s=vactif_gnaf> <pvactif> a;VER:pres;avoir </pvactif> <o> un;DET:ART;un <naf> placard;NOM;placard </naf> <
bloc s=vactif_gnaf> <pvactif> pendre;VER:infi;pendre </pvactif> <o> une;DET:ART;un <naf> plaque;NOM;plaque </n
bloc s=vactif_gnaf> <pvactif> trouver;VER:infi;trouver </pvactif> <o> une;DET:ART;un <naf> plaque;NOM;plaque <
bloc s=vactif_gnaf> <pvactif> a;VER:pres;avoir </pvactif> <o> un;DET:ART;un <naf> placard;NOM;placard </naf> <
bloc s=vactif_gnaf> <pvactif> fabriquer;VER:infi;fabriquer </pvactif> <o> une;DET:ART;un <naf> plaque;NOM;plaq
bloc s=vactif_avec_gn_gnaf> <pvactif> faire;VER:infi;faire </pvactif> avec;PRP;avec une;DET:ART;un lampe;NOM;l
bloc s=vactif_gn_sur_gnaf> <pvactif> est;VER:pres;être </pvactif> les;DET:ART;le données;NOM;donnée suivantes;
bloc s=vactif_avec_gnaf> a;VER:pres;avoir <pvactif> accouplé;VER:pper;accoupler </pvactif> avec;PRP;avec <o> u
```

Figure 24 Résultat d'extraction des structures prédicat-argument

2.2.2. Intersection des prédicats et élimination des prédicats basiques

L'intersection des prédicats consiste à trouver les prédicats les plus partagés de l'ensemble des noms d'artefacts donnés à l'avance. Dans un corpus, un nom d'artefact peut coexister avec les différents prédicats. Il est ainsi probable qu'une classe sémantique d'arguments partage certains prédicats communs et que ces prédicats communs sont les prédicats appropriés définitionnels des arguments. Les prédicats qui sont plus partagés par une classe sémantique de noms d'artefacts ont une plus grande possibilité d'appartenir à cette classe sémantique de noms d'artefacts. L'intersection permet de filtrer les prédicats qui ne font pas partie de la classe sémantique d'arguments donnée à l'avance. La Figure 25 est la capture d'écran du résultat d'intersection des prédicats. Le chiffre qui suit chaque prédicat indique le nombre de noms d'artefacts qui le partage. Le nombre de différents noms d'artefacts qui coexistent avec un prédicat est défini comme la fréquence d'intersection de ce prédicat.

```
ouvrir:11;s=vactif_gnaf:7;s=vactif_gn_de_gnaf:2;s=gnaf_va:2;s=vpassif_dans_gnaf:1
acheter:11;s=vactif_gnaf:11;s=gnaf_va:1
utiliser:9;s=vactif_gnaf:8;s=vactif_gn_de_gnaf:2;s=gnaf_va:1
ouvrer|ouvrir:9;s=vactif_gnaf:9;s=vactif_gn_de_gnaf:1
remplir:8;s=vactif_gnaf:8;s=gnaf_vpassif:1;s=gnaf_va:1;s=vactif_à_gnaf:1
poser:8;s=vactif_gnaf:8
enlever:8;s=vactif_gnaf:4;s=vactif_gn_de_gnaf:3;s=vactif_gn_dans_gnaf:1
voir:8;s=vactif_gnaf:5;s=gnaf_va:1;s=vactif_gn_de_gnaf:1;s=vactif_dans_gnaf:1
disposer:8;s=vactif_de_gnaf:4;s=vactif_gnaf:3;s=vactif_gn_de_gnaf:1;s=vactif_gn_sur_gnaf:1
vider:7;s=vactif_gnaf:7
coller:6;s=vactif_gnaf:3;s=gnaf_va_à_gnaf:2;s=gnaf_va:2;s=vactif_sur_gnaf:1;s=gn_va_à_gnaf:1;
commander:6;s=vactif_gnaf:6;s=vactif_de_gnaf:1;s=vactif_gn_de_gnaf:1
proposer:6;s=vactif_gnaf:6
changer:6;s=vactif_gnaf:5;s=vactif_gn_de_gnaf:2;s=vactif_gn_sur_gnaf:1
placer:6;s=vactif_dans_gnaf:3;s=vactif_gn_dans_gnaf:1;s=gn_va_sur_gnaf:1;s=vactif_gnaf:1
vendre:6;s=vactif_gnaf:4;s=gnaf_va:1;s=vactif_dans_gnaf:1
recevoir:5;s=vactif_gnaf:3;s=vactif_gn_dans_gnaf:1;s=vactif_gn_de_gnaf:1;s=vactif_dans_gnaf:1
retirer:5;s=vactif_gnaf:4;s=vactif_gn_de_gnaf:1
garder:5;s=vactif_gnaf:3;s=vactif_gn_dans_gnaf:1;s=vactif_gn_de_gnaf:1
tenir:5;s=vactif_dans_gnaf:3;s=vactif_gnaf:1;s=vactif_à_gnaf:1
ajouter:5;s=vactif_gn_de_gnaf:3;s=vactif_gnaf:3;s=vactif_gn_à_gnaf:2;s=vactif_à_gnaf:1;s=vact
transporter:5;s=vactif_gnaf:4;s=vpassif_dans_gnaf:1
```

Figure 25 Résultat d'intersection des prédicats & Résultat d'organisation des informations extraites

Le résultat d'intersection permet d'observer que de nombreux prédicats basiques occupent le haut de la liste, tels que *acheter, disposer, voir,* etc. Ces prédicats ne sont pas les prédicats appropriés des noms d'artefacts, mais ils ont une classe sémantique de noms d'artefacts. Pour les prédicats appropriés qui appartiennent à la classe d'arguments donnée à l'avance, leurs occurrences dans les séquences prédicat-argument extraites et leurs occurrences dans le corpus total sont plus ou moins similaires. Cependant, pour les prédicats basiques, il existe un grand écart entre leurs occurrences dans les séquences prédicat-argument extraites et leurs occurrences dans le corpus total, parce que les prédicats basiques ont un champ sémantique plus large. Par exemple, *acheter*, un prédicat basique, peut avoir des arguments de la classe sémantique de moyen de transport (*acheter une voiture, acheter une camionette, acheter un tracteur……*) et aussi des arguments d'autres classes sémantiques (*acheter un poële, acheter un manteau, acheter un ordinateur……*). Si les structures prédicat-argument extraites à partir des arguments de la classe sémantique de moyen de transport est noté comme sous-corpus1, il aura un grand écart entre la fréquence d'occurrence du prédicat *acheter* dans le sous-corpus1 et celle dans le corpus total qui comprend également les arguments d'autres classes sémantiques. En revanche, pour le prédicat approprié de la classe sémantique de moyen de transport *garer* qui ne coexiste presque qu'avec les arguments de la classe sémantique de moyen de transport, l'écart entre sa fréquence d'occurrence dans le

sous-corpus1 et celle dans le corpus total est beaucoup moins grand. Dans cette section, toutes les structures prédicat-argument extraites sont appelées sous-corpus1. En se basant sur l'écart de fréquence d'occurrence, on peut faire un filtrage sur les prédicats pour éliminer les prédicats basiques.

Cependant, il est possible qu'un prédicat approprié ne soit pas reconnu à cause de l'omission de son complément d'objet, car l'identification d'un prédicat ou d'un argument est réalisée par l'extraction des structures prédicat-argument dans lesquelles les arguments se trouvent à la position du sujet (plutôt pour les phrases passives) ou du complément d'objet. Cela conduit à un résultat inexact pour le calcul d'occurrences des prédicats dans les structures prédicat-argument extraites. Pour obtenir un résultat le plus précis possible, on procède comme suit : toutes les structures prédicat-argument qui concernent ou non les noms d'artefacts donnés à l'avance sont extraites. Ces structures extraites sont appelées sous-corpus2. De plus, la fréquence d'occurrence de chaque candidat prédicat récupéré dans ce sous-corpus est aussi calculé. Les candidats prédicats sont ceux obtenus à partir des arguments donnés à l'avance par les patrons syntaxiques. Pour le sous-corpus2, les structures avec l'omission du complément d'objet ne peuvent pas être récupérées non plus. Le schéma dans la Figure 26 représente la relation entre le corpus total, le sous-corpus2 et le sous-corpus1.

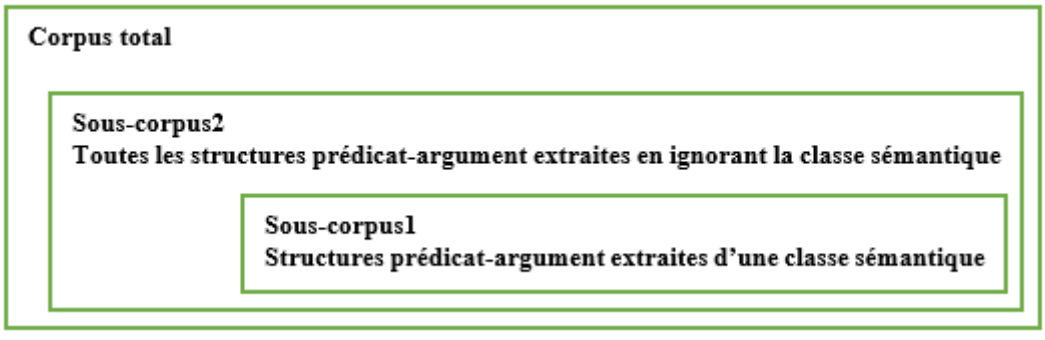

Figure 26 Relation entre le corpus total, le sous-corpus2 et le sous-corpus1

On calcule respectivement la fréquence d'occurrence du prédicat dans le sous-corpus1 (noté comme *FC1*), celle dans le corpus total (noté comme *FT*) et celle dans le sous-corpus2 (noté comme *FC2*). Ensuite, on calcule l'écart entre la fréquence d'occurrence du prédicat dans le sous-corpus1 et celle dans le corpus total (noté comme *Ecart1*). Et puis, on calcule

l'écart entre la fréquence d'occurrence du prédicat dans le sous-corpus1 et celle dans le sous-corpus2 (noté comme *Ecart2*). Ecart1 et Ecart2 sont respectivement calculés par les équations suivantes :

$$Ecart1 = \frac{FT - FC1}{FT} \tag{29}$$

$$Ecart2 = \frac{FC2 - FC1}{FT} \tag{30}$$

Le résultat de ce calcul est montré dans la Figure 27.

```
remplir:73:719:0.898470097357441:394:0.446453407510431;s=vactif_gnaf:69;s=gnaf_va:2;s=gnaf_vp
ouvrir:21:1125:0.981333333333333:423:0.357333333333333;s=vactif_gnaf:16;s=vactif_gn_de_gnaf:2
ouvrer|ouvrir:21:688:0.969476744186046:357:0.488372093023256;s=vactif_gnaf:20;s=vactif_gn_de_
vider:13:228:0.942982456140351:88:0.328947368421053;s=vactif_gnaf:13
coller:11:786:0.986005089058524:313:0.384223918575064;s=vactif_gnaf:5;s=gnaf_va_à_gnaf:2;s=gn
entremettre:9:117:0.923076923076923:31:0.188034188034188;s=vactif_gnaf:9
accrocher:9:169:0.946745562130177:65:0.331360946745562;s=vactif_gn_à_gnaf:3;s=vactif_gn_de_gn
transporter:9:90:0.9:49:0.444444444444444;s=vactif_gnaf:8;s=vpassif_dans_gnaf:1
réparer:8:423:0.981087470449173:168:0.378250591016548;s=vactif_gnaf:8
enregistrer:7:291:0.975945017182131:126:0.40893470790378;s=vactif_sur_gnaf:3;s=vactif_gn_sur_
afficher:7:675:0.98962962962963:304:0.44;s=vactif_gnaf:4;s=vactif_dans_gnaf:2;s=vactif_sur_gn
déplacer:6:271:0.977859778597786:120:0.420664206642066;s=vactif_gnaf:4;s=vactif_gn_dans_gnaf:
ranger:6:240:0.975:95:0.370833333333333;s=vactif_dans_gnaf:4;s=vactif_gn_dans_gnaf:1;s=vactif
copier:5:172:0.970930232558139:74:0.401162790697674;s=vactif_gnaf:4;s=vactif_gn_de_gnaf:1
organiser:5:194:0.974226804123711:99:0.484536082474227;s=vactif_gnaf:4;s=vactif_gn_en_gnaf:1
stocker:5:162:0.969135802469136:86:0.5;s=vactif_gnaf:3;s=vactif_gn_dans_gnaf:1;s=vactif_dans_
```

Figure 27 Résultat de calcul de l'écart de fréquence

Dans le résultat, le premier chiffre après le prédicat de chaque ligne indique la fréquence d'occurrence du prédicat dans le sous-corpus1; le deuxième chiffre réfère à la fréquence d'occurrence du prédicat dans le corpus total ; le troisième chiffre indique l'écart entre la fréquence d'occurrence du prédicat dans le sous-corpus1 et celle dans le corpus total ; le quatrième chiffre réfère à la fréquence d'occurrence du prédicat dans le sous-corpus2, et le cinquième chiffre indique l'écart entre la fréquence d'occurrence du prédicat dans le sous-corpus1 et celle dans le sous-corpus2. Le filtrage est effectué en référence à l'*Ecart1* et à l'*Ecart2* en même temps. On décide respectivement un seuil pour les deux types d'informations numériques (*Ecart1* et *Ecart2*) après plusieurs tests et on établit la condition suivante pour le filtrage :

Condition : si *Ecart1* d'un prédicat dépasse 0.978 et son *Ecart2* dépasse 0.450, on considère ce prédicat comme prédicat basique et on l'enlève de la liste.

Néanmoins, les graphes établis pour reconnaître les structures prédicat-argument ne sont pas exhaustifs. En conséquence, les structures prédicat-argument extraites ne sont pas exhaustives non plus. Ainsi, l'écart calculé en fonction des occurrences des prédicats dans le sous-corpus (structures prédicat-argument) et dans le corpus total n'a pas une précision de 100%. De plus, certains arguments ne se trouvent pas fréquemment dans les structures prédicat-argument. Pour cela, ce filtrage ne permet pas d'éliminer tous les prédicats basiques. Il reste potentiellement certains prédicats basiques et ces prédicats basiques amènent probablement de nombreux noms des autres classes sémantiques.

2.2.3. Calcul des patrons syntaxiques

Les quatre patrons syntaxiques de base sont les quatre possibilités prédites des distributions syntactico-sémantiques des prédicats appropriés des noms d'artefacts. Il est possible que le patron syntaxique (de base ou dérivé) avec lequel un prédicat est identifié ne soit pas le patron syntaxique de ce prédicat. Par exemple, la séquence *utiliser le klaxon d'un autre frigo* est mal reconnue comme structure prédicat-argument avec le patron syntaxique Vactif+GN+de+GNAF (*s=vactif_gn_de_gnaf*). La préposition *de* introduit en réalité un modifieur du substantif-tête *klaxon* au lieu d'un complément circonstanciel de moyen. Un autre exemple, si l'on a une séquence étiquetée comme *déchirer la poche à <NAF>douilles</NAF>.*, un patron syntaxique Vactif+GN+à+GNAF (*s=vactif_gn_à_gnaf*) pour le prédicat *déchirer* est obtenu à partir de cette séquence ; or *la poche à douilles* est un nom composé et la préposition *à* est un élément constitutif au sein de l'unité lexicale au lieu d'une préposition introduisant un complément d'objet indirect. L'objectif du calcul des patrons syntaxiques est de sélectionner le patron syntaxique adéquat du prédicat parmi les quatre patrons syntaxiques prédits.

Dans les structures prédicat-argument extraites, on récupère les prédicats, les arguments correspondants et les patrons syntaxiques enregistrés dans les étiquettes (telles que

s=Vactif_GNAF, Vpassif_prep_GNAF, s=GN_Va_prep_GNAF, etc.). On met ces informations dans un tableau où chaque ligne commence par un prédicat, suivi de sa fréquence d'intersection et les patrons syntaxiques associés. Un prédicat peut avoir plusieurs relations syntaxiques avec un même argument et un prédicat peut avoir encore plus de relations syntaxiques avec de différents arguments. Par exemple, pour le prédicat *coller*, il a apparu ensemble avec six différents arguments dans le corpus : *pochette* (*s=vactif_gnaf*), *trousse* (*s=gnaf_va, s=gnaf_va_à_gnaf*), *serviette* (*s=vactif_gnaf*), *classeur* (*s=gnaf_va, s=gnaf_va_à_gnaf*), *carton* (*s=vactif_gnaf, s=gn_va_à_gnaf*) et *frigo* (*s=vactif_gn_sur_gnaf, s=vactif_sur_gnaf*). La fréquence d'occurrence de chaque patron syntaxique associé au prédicat *coller* varie. Par exemple, *coller* est repéré par le patron syntaxique *s=vactif_gnaf* quatre fois, par le patrons syntaxique *s=gnaf_va* trois fois et par le patron syntaxique *s=gnaf_va_à_gnaf* deux fois. Les patrons syntaxiques de chaque prédicat sont mis en ordre par la fréquence d'occurrence décroissante. Le résultat d'organisation des informations extraites est montré dans la Figure 25.

Afin de sélectionner le patron syntaxique adéquat du prédicat parmi les quatre patrons syntaxiques de base (V+NAF, V+NAF+prep+NAF, V+Nc+prep+NAF, V+prep+NAF), nous calculons respectivement la probabilité d'avoir un complément d'objet direct, la probabilité que cet objet direct soit toujours un nom d'artefact et la probabilité d'avoir un complément introduit par une préposition à partir des informations syntaxiques extraites (telles que, *s=vactif_gnaf, s=vactif_gn_de_gnaf*, ……) pour chaque prédicat. La probabilité d'avoir un objet direct *P(cod)* est calculée en fonction de la formule :

$$P(cod) = \frac{C(gnaf)+c(gn)}{c(s)} \qquad (31)$$

c(gnaf) réfère à la fréquence d'occurrence des patrons syntaxiques contenant *gnaf* à la position d'objet direct, par exemple, *s=vactif_gnaf, s=gnaf_va* et *s=gnaf_vpassif* sont tous les patrons syntaxiques comprenant *gnaf* à la position d'objet direct. *c(gn)* désigne la fréquence d'occurrence des patrons syntaxiques contenant *gn* à la position d'objet direct. *c(s)* indique la fréquence d'occurrence totale de tous les patrons syntaxiques. Par exemple, dans la ligne suivante :

46) remplir : 77 ;s=vactif_gnaf :70 ;s=gnaf_vpassif :4 ;s=gnaf_va :2 ;s=vactif_à_gnaf :1,

s=vactif_gnaf (70), s=gnaf_vpassif (4) et *vactif_à_gnaf (2) sont* tous les patrons syntaxiques comprenant *gnaf* à la position d'objet direct. Ainsi, c(gnaf) de *remplir* est calculée comme suit :

c(gnaf)=70(s=vactif_gnaf)+4(s=gnaf_vpassif)+2(vactif_à_gnaf)=76

Pour *remplir*, il n'y a pas de patrons syntaxiques contenant *gn* à la position d'objet direct et *c(gn)* de *remplir* égale à 0 :

c(gn)=0

La fréquence d'occurrence totale de tous les patrons syntaxiques de *remplir* c*(s)* est obtenue par le calcul suivant :

c(s)=70(s=vactif_gnaf)+4(s=gnaf_vpassif)+2(vactif_à_gnaf)+1(s=vactif_à_gnaf)=77

Ainsi, *P(cod)* du prédicat *remplir* est calculée comme suit :

$$P(cod)= \frac{C(gnaf)+c(gn)}{c(s)} = (76+0)/77 \approx 0.9870$$

Ensuite, la probabilité d'avoir un objet direct qui est toujours un nom d'artefact *P(codnaf)* est calculée en fonction de l'équation suivante :

$$P(codnaf) = \frac{C(gnaf)}{c(s)} \quad (32)$$

La probabilité d'avoir un complément introduit par une préposition P(codi) est calculé au moyen de la formule suivante :

$$P(codi) = \frac{C(prep)}{c(s)} \quad (33)$$

Pour chaque probabilité (*P(cod), P(codnaf)* et *P(codi)*), nous décidons un seuil. Si *P(cod), P(codnaf)* ou *P(codi)* d'un prédicat dépasse le seuil défini, l'information indiquée par la probabilité est considérée positive. Par exemple, si *P(codnaf)* d'un prédicat dépasse le seuil

décidé pour *P(codnaf),* on croit que ce prédicat a toujours ou souvent un nom d'artefact à la position du complément d'objet direct. En fonction de ces trois valeures de probabilité, le patron syntaxique adéquat peut être sélectionné pour chaque prédicat parmi les quatre candidats patrons (V+NAF, V+NAF+prep+NAF, V+Nc+prep+NAF, V+prep+NAF). Finalement, les prédicats extraits sont classés dans quatre groupes en fonction de leurs patrons syntaxiques : le groupe de V+NAF, le groupe de V+NAF+prep+NAF, le groupe de V+Nc+prep+NAF et le groupe de V+prep+NAF. Le Tableau 11 donne un exemple sur le processus de sélection des patrons syntaxiques.

Prédicats	**P(cod)**	**P(codnaf)**	**P(codi)**	**Patron syntaxique sélectionné**
Préd1	positif	positif	positif	V+NAF+prep+NAF
Préd2	négatif	négatif	positif	V+prep+NAF
Préd3	positif	négatif	positif	V+GN+prep+NAF
Préd4	positif	positif	négatif	V+NAF

Tableau 11 Sélection des patrons syntaxiques

Dans le calcul des patrons syntaxiques, on prédit qu'il existe quatre possibilités de distributions syntactico-sémantiques (V+NAF, V+NAF+prep+NAF, V+Nc+prep+NAF etV+prep+NAF) pour les prédicats appropriés des noms d'artefacts. Ensuite, on fait l'hypothèse que tous les prédicats appropriés de noms d'artefacts ont ces quatre distributions syntactico-sémantiques et on étiquette toutes les structures prédicat-argument de chaque prédicat en fonction de ces quatre distributions syntactico-sémantiques. Finalement, on intègre le calcul probabiliste pour sélectionner le patron syntaxique adéquat de chaque prédicat.

2.2.4. Application de la méthode distributionnelle supervisée

À l'aide du calcul des patrons syntaxiques, les prédicats appropriés récupérés à partir d'un ensemble d'arguments sont classés en quatre groupes en fonction de leurs patrons syntaxiques (V+NAF, V+NAF+prep+NAF, V+Nc+prep+NAF et V+prep+NAF). Ces prédicats sont ensuite étiquetés dans le corpus (c'est le corpus de noms d'artefacts non étiqueté) avec les étiquettes <t_naf></t_naf>, <t_naf_p_naf></t_naf_p_naf>, <t_gn_p_naf></t_gn_p_naf> et <i_p></i_p> qui indiquent leurs distributions syntactico-

sémantiques par un script perl. L'ensemble des noms d'artefacts donnés à l'avance sont aussi étiquetés avec les balises <NAF><NAF>. Un graphe (appelé v_cod_coi_boucle.grf) est établi pour reconnaître les structures prédicat-argument à partir des prédicats étiquetés dans le corpus en fonction des informations syntactico-sémantiques enregistrées dans les étiquettes. Ensuite, on retourne vers la méthode distributionnelle supervisée et on répète tous les autres processus (l'extraction des structures prédicat-argument, l'organisation des informations étiquetées et l'intersection des arguments). Les nouveaux noms d'artefacts reconnus sont étiquetés par les étiquettes <NAFN></NAFN>.

La construction du graphe pour reconnaître automatiquement les structures prédicat-argument est réalisée en quatre étapes : la construction des grammaires locales pour les prédicats étiquetés par <t_naf></t_naf>, la construction des grammaires locales pour les prédicats étiquetés par <t_naf_p_naf></t_naf_p_naf>, la construction des grammaires locales pour les prédicats étiquetés par <t_gn_p_naf></t_gn_p_naf> et la construction des grammaires locales pour les prédicats étiquetés par <i_p></i_p>. Pour chaque groupe de prédicats, les autres patrons syntaxiques dérivés du patron syntaxique de base sont aussi établis. Par exemple, pour les prédicats étiquetés par <t_naf></t_naf>, en plus du patron syntaxique de base du groupe V+NAF, on a également les patrons syntaxiques dérivés : être+Vpassif, V+ADV+Vpassif, V+Vpp, etc.

On projette les graphes établis pour le corpus et les structures prédicat-argument sont étiquetées. Le résultat est enregistré au format texte brut et la case des lettres est transformée en minuscules. Ensuite, on lemmatise les textes avec TreeTagger. Finalement, on lance le script extrait_intersection_BversA développé pour l'intersection des arguments reconnus dans la méthode semi-supervisée.

3. Évaluation

Les différents types d'évaluation ont été développées pour évaluer les différents systèmes d'applications. Le choix du type d'évaluation dépend du type de systèmes (les systèmes d'analyse, les systèmes qui produisent les sorties ou les systèmes interactifs) ainsi

que des entrées et des sorties du système. Dans le cadre du projet de la thèse, on se concentre seulement sur les types d'évaluation pour les systèmes d'acquisition terminologique. Dans cette partie, on présente d'abord l'état de l'art concernant l'évaluation des systèmes de terminologie. Ensuite, on expose certaines mesures d'évaluation souvent utilisées pour évaluer les systèmes de terminologie. Finalement, on présente la méthode d'évaluation pour la méthode distributionnelle et les résultats d'évaluation.

3.1. État de l'art concernant l'évaluation

L'évaluation des systèmes de terminologie s'effectue souvent sous deux angles : celle des aspects terminologiques et celle de la convivialité de l'interface d'utilisateur. La sélection du type d'évaluation des aspects terminologiques détermine directement la pertinence de l'évaluation pour le résultat du système. Dans ce qui suit, on présente quatre protocoles d'évaluation proposés respectivement par Hamon & Hû (2002), Christophe Jouis & ARC A3 (2004), Timimi (2006) et Nazarenko et al. (2009) pour les systèmes de terminologie.

A. How to evaluate necessary cooperative systems of terminology building? (Hamon & Hû, 2002)

Hamon et Hû (2002) ont défini une série de critères pour l'évaluation des systèmes d'acquisition terminologique. Ils évaluent les systèmes d'acquisition terminologique sous deux angles : les aspects terminologiques des systèmes et les caractéristiques ergonomiques de l'interface d'utilisateur parce que la plupart des systèmes d'acquisition terminologique sont développés pour identifier, classifier et organiser les termes à l'aide des terminologues et sont en général composés de deux étapes : l'extraction des termes et la structuration supervisée par les experts. Certaines limites de précision et de rappel pour l'évaluation des systèmes d'acquisition terminologique sont indiquées par Hamon et Hû (2002), telles que l'impossibilité de refléter la spécificité du système terminologique, l'ignorance du type de validation stricte ou pas (par ex., ceux qui sont refusés par une validation stricte peuvent être, en revanche, utiles pour les terminologues.), l'incompatibilité du standard construit avec

d'autres systèmes pour l'évaluation d'exhaustivité des termes et l'impossibilité d'évaluer la convivialité ainsi que les autres paramètres techniques du système.

Dans la méthode d'évaluation des systèmes terminologiques de Hamon et Hû (2002), il y a une série de critères pour l'évaluation des aspects terminologiques des systèmes et une série de critères pour l'évaluation ergonomique des systèmes. Les critères établis pour l'évaluation des aspects terminologiques sont classés en fonction de six méta-critères : le but du système, la stratégie, les paramètres, les fonctionnalités, les qualités des résultats et les formats d'échange. Les critères de chaque classe sont résumés dans le Tableau 12. Certains critères listés sont évalués selon la valeur booléenne et les autres sont évalués selon la valeur numérique ou textuelle.

Méta-critères	**Critères**	**Descriptions**
But du système	Extraction des termes	unités identifiées (telles que les groupes nominaux, les collocations, etc.)
	Structuration des termes	acquisition des relations (type de relation), similarité entre les termes et la classification
	Autres	taille du corpus, prétraitement demandé, etc.
Stratégie	Approche linguistique	règles de transformation, patrons lexico-sémantiques, frontières des mots
	Approche statistique	endogène, fondé sur les ressources
	Apprentissage artificiel	
	Ressources	dictionnaire de langue générale, thésaurus, ressources spécialisées, données construites automatiquement
Paramètres	Réglage	avant ou au cours du traitement
	Définition des connaissances acquises	application, demandes des terminologues
Fonctionnalités	Connaissances des terminologues	correction, addition
	Traitement transparent	méthode de calcul, entrée des connaissances
Qualités	Type de validation	strict, large
	Mesures	précision, rappel, F-mesure, minoring-error précision
	Type d'erreur	ambiguïté, données partielles ou erronées, stemming erreurs
Formats d'échange	Formats standard	TEI (Ideand Veronis, 1995), GENETER (Le Meur, 1998)
	Formats spécifiques	

Tableau 12 Critères d'évaluation

L'évaluation ergonomique comprend trois méta-critères : la convivialité, la qualité technique et le sentiment général. L'évaluation de la convivialité du système consiste à évaluer l'utilisation de l'interface et la qualité d'interaction. L'évaluation de la qualité technique a pour objet d'évaluer la robustesse et la portabilité du système et l'évaluation du

sentiment général est d'évaluer la différence entre la qualité du système et la satisfaction de l'utilisateur.

B. ARC A3: A Method for evaluation term extracting tools and/or semantic relations between terms from corpora (Christophe Jouis & ARC A3, 2004)

Dans la méthode d'évaluation présentée par Christophe Jouis et ARC A3 (2004), on introduit un ensemble de références de terminologie établie manuellement. Ces références sont associées au domaine du corpus et les sorties variées sont comparées avec ces références de terminologie. Ces références sont considérées comme des références relatives. On a proposé de faire l'évaluation qualitative et quantitative par une présentation normalisée. D'abord, les résultats obtenus par le système doivent être mis dans un tableau comme le suivant :

T	A	B	C	D	E	T1	T2	R

Tableau 13 Tableau d'évaluation

T représente les classes de termes, A signifie très bien, B signifie acceptable, C indique acceptable après modification, D désigne que ce n'est pas absurde mais ce n'est pas très caractéristique, E signifie artefact, T1 réfère au terme modifié par le spécialiste, T2 réfère au nombre des systèmes ayant trouvé le même terme et R permet de noter les autres avis. L'expert balise une des colonnes (A, B, C, D, et E) pour chaque classe de termes et remplit T1 et R si nécessaire. Ensuite, chaque résultat doit être présenté dans le tableau suivant :

T	A	B	C	D	E	F

Tableau 14 Tableau de représentation

Dans ce tableau, T représente également les classes de termes, A signifie que la classe de termes appartient à la référence de thésaurus relative (RTHR), B désigne que la classe de

termes n'appartient pas à la RTHR mais elle est envoyée ici pour la synonymie, C signifie que la classe de termes n'appartient pas à la RTHR mais doit apparaître dans la RTHR, D indique que la classe de termes n'appartient pas à la RTHR mais doit apparaitre dans RTHR sous un autre descripteur, E réfère aux mesures de bruit et F désigne les mesures de silence.

C. Évaluation des systèmes d'acquisition de terminologie : nouvelles pratiques, nouvelles métriques (Timimi, 2006)

Timimi (2006) propose une réflexion sur la pratique d'évaluation et tente de déterminer quels sont les critères d'évaluation des systèmes d'acquisition terminologique et comment ils sont déterminants pour les besoins informationnels en présentant le protocole d'évaluation dans le projet CESART (Mustafa el Hadi, Timimi et Dabbadie, 2004). Trois types de démarches d'évaluation sont indiqués : l'évaluation diachronique qui compare le système avec ses versions antérieures en vue d'une étude diachronique de ses performances ; l'évaluation d'appariement (transversale) qui a pour objet de comparer les performances du système avec d'autres systèmes ou avec des résultats de référence établis à l'avance ; l'évaluation de diagnostic dans laquelle l'évaluateur cherche les sources de performance d'un système conçu pour une tâche précise à partir d'une série de tests. On a aussi résumé les facettes de l'évaluation : l'évaluation avec interface statique (juger les performances sans interventions ou enrichissement terminologique) par opposition à l'interface dynamique (juger les performances suite à une intégration de ressources extérieures) et l'évaluation de type boîte noire (juger les performances globales à partir seulement des entrées et des sorties) par opposition à la boîte transparente (juger le fonctionnement interne du système à travers ses différents modules et prétraitements).

Le projet CESART consiste à élaborer un protocole pour l'évaluation des systèmes d'acquisition terminologique. Le protocole de CESART est une évaluation boîte noire et orientée applications. « L'appréciation (ou non) des performances d'un système ne peut être indépendante de l'application industrielle ou langagière pour laquelle le système a été conçu. » (Timimi, 2006). Trois applications sont fixées dans le protocole de CESART :

l'extraction de termes pour l'enrichissement de ressources terminologiques, l'indexation contrôlée et enrichissement ainsi que l'extraction de relations. D'après le protocole d'évaluation de CESART, le corpus utilisé pour l'évaluation doit obligatoirement être représentatif du domaine, homogène et volumineux (pour les systèmes basés sur les approches statistiques). Des ressources complémentaires sont aussi mises à la disposition en fonction des besoins spécifiques des participants (un corpus d'apprentissage similaire au corpus de l'évaluation du point de vue de la taille, du format et du thème et un référentiel validé par les experts pour le calcul de rappel et de précision). Les ressources de référence doivent être complétées par les experts et doivent être liées au domaine du corpus. L'expertise humaine est nécessaire en l'absence des références ou des méthodes automatisées. Un expert humain doit être à la fois un spécialiste du domaine du corpus et un familier des pratiques documentaires et langagières vu l'aspect interdisciplinaire du projet CESART. Les espaces métriques utilisés sont les espaces métriques de performance de l'extraction d'information rappel et précision. De plus, pour les systèmes de constitution terminologique, trois niveaux dégressifs A, B et C de pertinence des termes et une classification qualitative de l'ensemble des systèmes expertisés en leur attribuant une note globale choisie entre 1 et 5 assortie d'une appréciation sont recommandés aux experts pour l'évaluation manuelle. Pour les systèmes d'indexation, on effectue une évaluation automatique avec le référentiel humain et une évaluation manuelle par expert pour les termes d'indexation supplémentaires proposés par les systèmes mais qui ne figurent pas dans le référentiel. Pour les systèmes d'extraction de relations, on restreint l'évaluation aux trois types de relations : synonymie, hyperonymie ainsi que méronymie et l'expert est demandé de donner une note entre 1 à 3 pour mesurer la pertinence de la relation.

D. Évaluation des outils terminologiques : enjeux, difficultés et propositions (Nazarenko et al., 2009)

Nazarenko et al. (2009) ont proposé une méthode d'évaluation basée sur une décomposition des outils d'analyse terminologique en fonctionnalités élémentaires et sur la définition de mesures de précision et de rappel. Il s'agit d'un nouveau protocole d'évaluation adapté à la complexité des produits terminologiques, la dépendance aux applications, le rôle

de l'interaction avec l'utilisateur et la variabilité des terminologies de référence. D'après Nazarenko et al. (2009), l'analyse terminologique doit être décomposée en fonctionnalités élémentaires afin de les évaluer séparément, car la qualité de la terminologie doit être mesurée selon plusieurs facteurs (la qualité des termes extraits et celle des relations qu'ils entretiennent). De plus, en raison de la relativité de référence (c'est-à-dire que la référence construite par les terminologues n'est jamais absolument exhaustive ou correcte), on propose également de multiplier les évaluations pour tester les systèmes en comparaison des différentes références. Enfin, l'application pour laquelle les systèmes sont développés doit également être prise en compte pour produire les critères d'évaluation et l'évaluation sur l'interaction est aussi nécessaire pour la plupart des systèmes d'acquisition terminologique.

Nazarenko et al. (2009) indiquent que les mesures de rappel et de précision permettent de prendre en compte le bruit et le silence des résultats mais reposent sur l'hypothèse à l'effet que la pertinence est une notion binaire (oui/non), alors qu'un candidat-terme peut être seulement proche d'un terme de référence. Ainsi, on propose une évaluation qui permet de tenir compte de la pertinence graduée. Les mesures de précision et de rappel terminologiques proposées reposent sur une distance terminologique pour respecter le caractère gradué de la pertinence des termes et sur un ajustement de la sortie à la référence en raison de la relativité de la référence. Le comportement des espaces métriques proposés est le suivant : un système parfait doit avoir une valeur de qualité maximale ; un système qui renvoie en plus des termes proches des termes de la référence ne doit pas être pénalisé ou ne l'être que faiblement par rapport au système précédent ; un système qui renvoie à une liste de non-termes doit avoir la valeur minimale. La distance terminologique consiste à mesurer la distance entre deux termes en calculant le coût de transformation de l'une à l'autre en se basant sur trois opérations élémentaires : ajout, substitution et suppression (Sant, 2004). Le coût de transformation est la somme des coûts des trois opérations élémentaires et la somme du coût est ensuite divisée par le nombre de lettres du terme le plus long des deux termes comparés pour obtenir une distance normalisée. On distingue deux distances : une distance calculée en s'appuyant sur les chaînes de caractères et une distance calculée en s'appuyant sur les termes de multi-mots. La distance terminologique est définie comme la moyenne des distances sur les chaînes de caractères et

sur les termes. L'évaluation graduée est effectuée en calculant les distances terminologiques entre les éléments de la sortie et les éléments de la référence. En ce qui concerne l'ajustement de la sortie à la référence, on considère les termes de la sortie qui correspondent au même terme de la référence en bloc et on calcule les mesures de précision et de rappel sur la partition de S qui est soit un ensemble de termes de la sortie qui se rapprochent du même terme de R avec une distance inférieure à un seuil soit composé d'un terme seul.

3.2. Mesures

Les mesures les plus utilisées pour l'évaluation des systèmes d'acquisition terminologique sont les mesures de rappel et de précision. Cependant, les désavantages de rappel et de précision ont été beaucoup présentés dans la section 3.1. Ainsi, de nombreuses autres mesures sont développées pour l'évaluation des systèmes d'acquisition terminologique, telles que la métrique LA (leaf-ancestor) (Smpson et Babarczy, 2003) qui permet d'évaluer la précision de parsing, la mesure UAS (Unlabeled Attachment Score) (Tsarfaty, Nivre et Ndersson, 2011) qui consiste à évaluer la qualité des liens de dépendance, la précision pondérée de Hamon et Nazarenko (2001) qui a pour objet d'évaluer les relations sémantiques, UAP (un-interpolated Average Precision) de Schone et Jurafsky (2001) qui permet de calculer une précision moyenne, etc. Dans ce qui suit, on présente trois types de mesures souvent utilisées dans l'évaluation des systèmes d'acquisition terminologique.

3.2.1. Précision, rappel et F-mesure

Les sorties d'un système peuvent être divisées en quatre cas : vrai positif, faux positif, vrai négatif et faux négatif. Le vrai positif réfère au nombre de termes (messages, documents, ou séquences) trouvés avec justesse comme ceux qui doivent être reconnus. Le faux positif indique le nombre de termes trouvés par erreur comme ceux qui doivent être reconnus. Le vrai négatif représente le nombre de termes qui sont trouvés à raison comme ceux qui ne doivent pas être reconnus et le faux négatif est le nombre de termes qui sont trouvés par erreur comme ceux qui ne doivent pas être reconnus. Ainsi, la précision est définie comme la proportion de solutions trouvées qui sont pertinentes et elle permet de mesurer la capacité du

système à refuser les solutions non-pertinentes. La précision (*P*) est calculée par la formule suivante :

$$P = \frac{vrai\ positif}{vrai\ positif + faux\ positif} \quad (34)$$

Le rappel désigne la proportion des solutions pertinentes qui sont trouvées et consiste à mesurer la capacité du système à donner toutes les solutions pertinentes. La formule pour calculer le rappel (*R*) est la suivante :

$$R = \frac{vrai\ positif}{vrai\ positif + faux\ négatif} \quad (35)$$

La F-mesure est la moyenne harmonique du rappel et de la précision ($\beta > 0$) (Cornuéjols et Miclet, 2010) :

$$F\text{-}mesure = \frac{(1+\beta^2)\cdot rappel\cdot précision}{\beta^2\cdot rappel + précision} \quad (36)$$

β est un coefficient pour pondérer la moyenne harmonique selon les besoins. Par exemple, la plupart des systèmes d'acquisition terminologique semi-automatique exigent un rappel assez élevé même si en sacrifiant la précision afin de garantir la quantité de candidat-termes proposés aux terminologues pour la construction de la ressource terminologique. Cela exige de sanctionner plus fortement les faux négatifs que les faux positifs en donnant une valeur de plus de 1 à β. Au contraire, pour les systèmes exigeant une précision élevée même si en sacrifiant le rappel, on est obligé de donner une valeur de moins de 1 à β pour être plus exigeant sur la précision. On confond souvent la F-mesure avec la F1-mesure (dans lequel $\beta=1$) :

$$F1\text{-}mesure = \frac{2\cdot rappel\cdot précision}{rappel + précision} \quad (37)$$

D'après Popescu-Belis (1999), l'un des scores est plus proche de zéro (le score de rappel ou le score de précision), le score de F-mesure est plus proche de zéro. Ainsi, entre le score de rappel et le score de précision, le score F-mesure est proche de celui qui est plus bas. Si l'un

des scores (le score de rappel ou le score de précision) est égal à zéro, le score F-mesure égale à zéro aussi.

3.2.2. Précision UAP

UAP (un-interpolated Average Precision) (Schone et Jurafsky, 2001) a pour objet de faire la moyenne de la précision à la position du i^{ème} terme pertinent trouvé dans la sortie (une liste de termes). UAP est calculé par la formule suivante :

$$UAP = \frac{1}{k} \sum_{i-1}^{k} P_i \quad (38)$$

k indique le nombre total des termes corrects dans la sortie, P_i est la précision à la position *i* qui est défini par la formule suivante :

$$P_i = \frac{i}{H_i} \quad (39)$$

H_i est le nombre de termes qu'on suppose qu'on doit parcourir pour trouver le i^{ème} terme correct dans la sortie. La somme de P_i est divisée par le nombre total de termes corrects dans la sortie *k* et le résultat obtenu est l'UAP. Par rapport à la mesure de précision et de rappel qui est souvent calculée sur une partie des résultats, UAP permet de prendre en compte tout l'ensemble de la sortie.

3.2.3. Précision pondérée

La précision pondérée de Hamon et Nazarenko (2001) est développée pour évaluer l'utilité et le caractère suggestif des systèmes d'extraction des relations terminologiques (relation de synonymie, relation d'hypomimie, etc.). Elle a pour objet de minimiser le poids de chaque erreur dans les classes de relations extraites (la classe de relation d'hyponymie, la classe de relation de synonymie, etc.) dans lesquelles la plupart des relations ne sont pas validées. Cette précision permet de mieux refléter les avis des terminologues et l'utilité des relations inférées pour la structuration terminologique (qui est effectuée en fonction des relations terminologiques). Les relations inférées sont les relations reconnues à l'aide des

règles d'inférence établies à l'avance à partir des règles de synonymie dans le projet présenté par Hamon et Nazarenko (2001). La précision pondérée (*WeightedPrecision*) est calculée en fonction de la formule suivante :

$$WeightedPrecision = \sum_{i=1}^{N_f} NdErrors_i \times \frac{W_i}{NbNbLinks_i} \quad (40)$$

dans laquelle N_f est le nombre de classes de relations terminologiques et W_i définit le poids de chaque erreur dans une classe de relations terminologiques et il est calculé par la formule suivante :

$$W_i = \min (R_i, 1) \quad (41)$$

R_i est le ratio d'erreur de la classe de relations terminologiques i et il est défini comme ce qui suit :

$$R_i = \max (NbLinks/2, 1)/NbErrors_i, \quad if\ NbErrors_i > 0 \quad (42)$$

dans laquelle $NbLinks_i$ indique le nombre de relations dans la classe de relations terminologiques i, et $NbErrors_i$ est le nombre d'erreurs dans la classe de relations terminologiques i.

3.3. Expérimentation et évaluation de la méthode distributionnelle

L'évaluation de la méthode distributionnelle est effectuée sous deux aspects : quantitative et qualitative. Dans l'évaluation quantitative, on fait appel au dictionnaire en ajoutant une annotation manuelle pour établir un standard. Le résultat obtenu par la méthode distributionnelle est comparé avec le standard pour calculer le taux de précision, le taux de rappel et le taux de F-mesure. Dans l'évaluation qualitative, on analyse respectivement les types d'erreurs et les types de silences de la méthode distributionnelle supervisée et la méthode distributionnelle semi-supervisée. Dans ce qui suit, on détaille la méthode d'évaluation et les résultats d'évaluation.

3.3.1. Évaluation quantitative

Pour l'expérimentation de la méthode supervisée, on a choisi une centaine de prédicats appropriés pour la classe sémantique générale des noms d'artefacts. Les prédicats appropriés sont classés en fonction de leurs particularités syntactico-sémantiques. Les patrons syntaxiques sont établis en se basant sur la distribution syntactico-sémantique des prédicats donnés à l'avance. Pour la méthode semi-supervisée, elle est testée sur trois classes sémantiques : contenants d'arrangement, moyens de transport et appareils de cuisson. Pour chaque classe sémantique, on a construit une liste d'arguments manuellement. Chaque liste d'arguments comporte environ une vingtaine de noms d'artefacts de la classe sémantique.

Dans l'évaluation quantitative, le résultat est évalué en faisant appel à une ressource lexicale de noms d'artefacts [3]. Elle comprend 13,400 noms d'artefacts monolexicaux. Les noms d'artefacts sont regroupés par classes sémantiques (Buvet, 2009b). On étiquette respectivement les noms d'artefacts de la classe sémantique des contenants d'arrangement, de la classe sémantique des appareils de cuisson et de la classe sémantique des moyens de transport dans le corpus avec ce dictionnaire. Et puis, on ajoute une annotation manuelle pour compléter l'étiquetage des noms d'artefacts dans le corpus du fait que ce dictionnaire n'est pas complet. Le résultat obtenu est considéré comme le standard. Le standard et le résultat obtenu par la méthode distributionnelle sont d'abord lemmatisés par TreeTagger. Et puis, on extrait respectivement les termes étiquetés dans les deux résultats en éliminant la répétition. Finalement, on compare la liste de termes obtenue par la méthode distributionnelle avec le standard en calculant la précision, le rappel et la F-mesure.

Dans le résultat de la méthode supervisée, on découvre qu'il existe de plus en plus de bruits parmi les arguments dont les fréquences d'intersection sont inférieures à 5. Pour décider si 5 est le meilleur seuil d'intersection d'arguments, on fait un test comme suit : pour la méthode supervisée, on définit respectivement le seuil d'intersection d'arguments à 4, 5, 6, 7 et 8. Ensuite, on calcule respectivement le taux de précision, le taux de rappel et le taux de

[3] Cette ressource lexicale nous a été fournie par P.-A. Buvet.

F-mesure. Les résultats d'évaluation obtenus avec les différents seuils sont listés dans le Tableau 15.

Seuils	Précision	Rappel	F-mesure
4	68.31%	76.20%	72.03%
5	70.08%	74.16%	72.06%
6	89.40%	71.78%	79.63%
7	90.27%	67.45%	77.21%
8	90.59%	65.33%	75.91%

Tableau 15 Résultats d'évaluation avec les différents seuils

La Figure 28 montre une comparaison entre les résultats d'évaluation obtenus selon les différents seuils. On voit bien que le taux de rappel baisse et le taux de précision augmente au fur et à mesure de l'augmentation du seuil. Quand le seuil est supérieur à 7, le taux de précision commence à devenir relativement stable. Le résultat est le meilleur quand le seuil égale à 6.

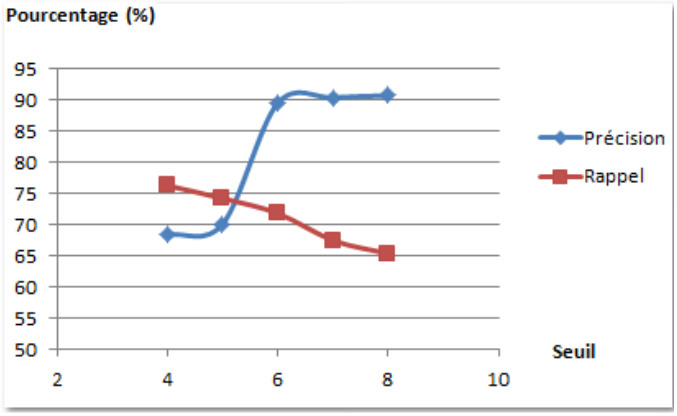

Figure 28 Comparaison des résultats obtenus avec les différents seuils

Pour la méthode semi-supervisée, l'expérimentation est d'abord exécutée pour la classe sémantique des contenants d'arrangement. La méthode supervisée est répétée itérativement cinq fois. Les résultats obtenus après chaque itération sont respectivement évalués. On découvre que la proportion de bruits commence à augmenter largement parmi les arguments dont les fréquences d'intersection sont inférieures à 3. Ainsi, le seuil d'intersection d'arguments est d'abord défini à 3. Le résultat obtenu par la méthode semi-supervisée comprend les termes donnés à l'avance. Dans le Tableau 16, on liste les résultats d'évaluation obtenus avec les différents nombres d'itération.

Nombres d'itération	Précision	Rappel	F-mesure
1	86.12%	29.41%	43.85%
2	84.07%	58.82%	69.21%
3	81.34%	81.02%	81.20%
4	76.10%	81.02%	78.48%
5	57.79%	79.87%	67.06 %

Tableau 16 Résultats d'évaluation avec les différents nombres d'itération

La Figure 29 montre la comparaison entre ces résultats d'évaluation. On voit que la F-mesure du résultat obtenu après trois fois d'itérations est la plus élevée. Après quatre fois d'itérations, le taux de précision commence à baisser très vite et le taux de rappel atteint une valeur relativement stable. Le taux de précision est abaissé à cause des bruits amenés potentiellement par les prédicats basiques non éliminés et les arguments d'autres classes sémantiques obtenus après chaque itération.

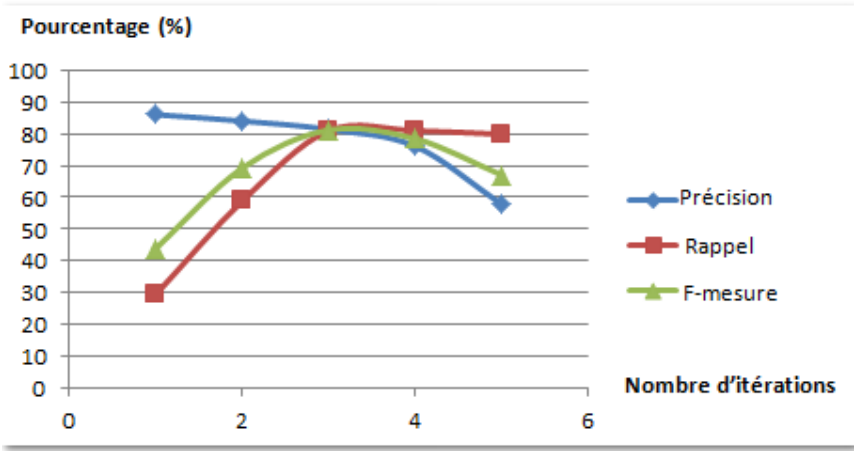

Figure 29 Comparaison des résultats obtenus avec les différents nombres d'itération

Ensuite, pour décider si le seuil 3 est le seuil d'intersection d'arguments permettant d'obtenir le meilleur résultat, on définit respectivement le seuil d'intersection d'arguments à 2, 4, 5 et 6 pour tester. Et puis, les mêmes processus d'expérimentation présentés ci-dessus sont répétés pour chaque seuil. Les résultats obtenus après chaque itération avec chaque seuil sont respectivement évalués. De même, pour chaque seuil, on peut obtenir un résultat qui est meilleur que les autres. On a fait une comparaison entre les meilleurs résultats obtenus avec chaque seuil. Dans la Figure 30, on peut voir la comparaison des F-mesures les plus élevés qui peuvent être obtenues avec les différents seuils.

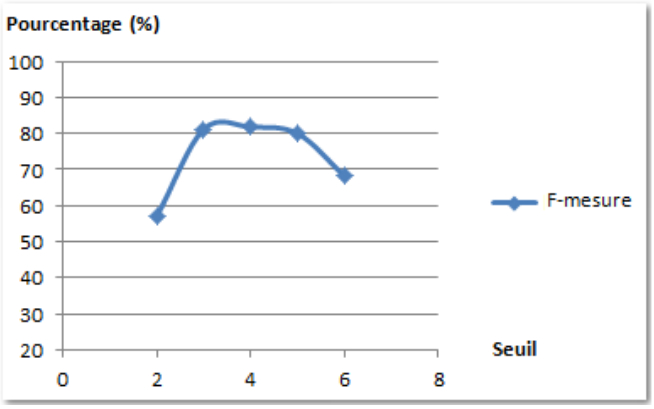

Figure 30 Comparaison des meilleurs résultats obtenus avec les différents seuils

Pour les deux autres classes sémantiques : la classe sémantique des moyens de transport et la classe sémantique des appareils de cuisson, les mêmes expérimentations et évaluations sont effectuées. Après le test, on choisit respectivement 3, 2 et 2 comme le seuil d'intersection d'arguments de la classe sémantique des contenants d'arrangement, de la classe sémantique des appareils de cuisson et de la classe sémantique des moyens de transport. 3, 3 et 2 sont respectivement sélectionnés comme le nombre d'itérations de la classe sémantique des contenants d'arrangement, de la classe sémantique des appareils de cuisson et de la classe sémantique des moyens de transport. Ces seuils permettent d'obtenir un meilleur résultat pour chaque classe sémantique. Dans le Tableau 17, on liste le résultat d'évaluation final sur chaque classe sémantique d'arguments.

Classes sémantiques	**Précision**	**Rappel**	**F-mesure**
Moyens de transport	62.46%	58.53%	60.43%
Appareils de cuisson	70.14%	76.87%	73.35%
Contenants d'arrangement	81.34%	81.02%	81.20%

Tableau 17 Résultat d'évaluation final de la méthode distributionnelle semi-supervisée

3.3.2. Évaluation qualitative

Pour analyser les types d'erreurs dans le résultat de la méthode distributionnelle supervisée, on a choisi au hasard 150 termes dans le résultat final ; on a trouvé 16 erreurs. Ces erreurs peuvent être divisées en cinq types. Le type A regroupe les termes dont les référents peuvent varier selon les contextes, tels que *type, chose, morceau, version*, etc. Le type B comporte les erreurs produites par certains prédicats qui ne sont pas les prédicats appropriés mais qui sont néanmoins utilisés pour repérer les noms d'artefacts afin d'augmenter le rappel

du résultat, par exemple, dans les séquences *verser de l'eau*, *verser du vin* et *verser du jus de fruit*, le prédicat *verser* est donné dans la liste pour repérer plus de noms d'artefacts comme *vin* et *jus de fruit*, mais *eau* est aussi reconnu à tort en même temps. Le type C sont les erreurs produites en raison des constituants syntaxiques mal reconnus, par exemple, dans *ouverture de session*, *session* est reconnu à tort comme un nom d'artefact en considérant *de* comme une préposition introduisant le complément de moyen. Le type D comprend les termes qui sont en fait les constituants des termes composés, tels que *terre* dans *tuile terre*, *air* dans *air bag*, *liquide* dans *liquide vaisselle*, etc. Enfin, on calcule respectivement la quantité d'erreurs de chaque type et la proportion de chaque type d'erreur. Dans le Tableau 18, on liste les cinq types d'erreurs avec les exemples correspondants et leur proportion respective parmi tous les types d'erreurs. On peut voir que les termes d'autres classes sémantiques récupérés par les prédicats non appropriés (type B) occupent la proportion la plus importante de tous les types d'erreurs. De plus, les erreurs produites en raison des constituants syntaxiques mal reconnus (type C) occupent également une grande proportion. Finalement, le manque de patrons morphosyntaxiques sur les termes composés (type D) est aussi une raison importante pour expliquer les erreurs.

Types d'erreurs	**Nombre**	**Proportion**	**Exemples**
A	4	25.00%	*chose* (*chose collée*/*chose bloquée*), *type* (*assemble de type*), *version* (*version fermée*), *morceau* (*couper plusieurs morceaux*)
B	5	31.25%	*pomme* (*laver les pommes*), *eau* (*rempli d'eau*/*rassembler l'eau*/*changer l'eau*/*verser de l'eau*/*cuire le sucre avec l'eau*/*retirer l'eau des tomates farcies*), *résistance* (*bouger des résistances mécaniques*), *discussion* (*poser la question dans une nouvelle discussion*), *préparation* (*verser la préparation dans le bol*)
C	4	25.00%	*vitesse* (*changement de vitesse*), *session* (*ouverture de session*), *contact* (*un film alimentaire au contact*), *fruit* (*cuisson de la purée de fruit*)
D	3	18.75%	*terre* (*tuile terre cuite*), *air* (*équipé d'un double air bag*), *liquide* (*liquide vaisselle*)
Total	16	100%	

Tableau 18 Types d'erreurs dans le résultat de la méthode distributionnelle supervisée

Ensuite, on étiquette le corpus avec le résultat obtenu par la méthode distributionnelle supervisée et on analyse une soixantaine de textes dans lesquels il y a environ 35 silences. Ces silences s'expliquent de la façon suivante : ils sont provoqués par l'intersection des arguments

(type de silence A) ; ils sont produits faute de patrons morphosyntaxiques pour les identifier (type B) ; ils ne se trouve pas fréquemment dans les structures prédicat-argument (type C); ils sont qui coexistent souvent avec les prédicats basiques (comme *avoir, mettre, trouver*, etc.) filtrés ou les prédicats appropriés qui ne sont pas listés dans la liste de prédicats donnée à l'avance (type D). Pour les silences du type A, il existe d'autres raisons. Par exemple, *saxo* est un terme supprimé par l'intersection et sa fréquence d'intersection est relativement basse parce qu'il se trouve souvent dans les titres (qui ne sont pas les structures prédicat-argument) ou après les prédicats basiques (tels que *acquérir une saxo, touche les saxos, parler de ton saxo*, etc.). De plus, la liste incomplète de prédicats appropriés donnée à l'avance et le manque de patrons syntaxiques (ou morphosyntaxiques) pour repérer les structures prédicat-argument (ou les termes composés) sont aussi deux raisons importantes pour les silences du type A. Dans le Tableau 19, on liste tous les types de silences et les exemples associés. On calcule aussi la proportion respective de chaque type de silence. La plupart des silences sont provoqués par leur coexistence fréquente avec les prédicats basiques et non avec des prédicats appropriés. De plus, les patrons syntaxiques (ou les patrons morphosyntaxiques) incomplets baisse également le rappel du résultat. Enfin, la liste incomplète des prédicats appropriés donnée à l'avance est aussi un facteur important pour la production des silences.

Types de silences	**Nombre**	**Proportion**	**Exemples**
A	25	71.43%	*coque, sphère, capteur, catalyseur, amortisseur, courroie, bougie, afficheur, étrier, autoradio, ecran, Berlingo, cartouche, coffre, eprom, vérin, ambulance, corbillard, rétro, saxo, flaque, capuchon, chiffon, rotule, Xantia*
Raison 1 : présence fréquente dans les structures non prédicat-argument			*saxo* (*achat saxo 1.5l d _ citroën _ forum marques*/*problème direction assisté saxo 1.5d 2001 _ saxo _ citroën _ forum marques*), *flaque* (*Pas énormément, mais une flaque*/*Je démarre et la un grosse flaque d'huile K*), *capuchon* (*un système de ventilation sur le capuchon*/*cerise sur le gâteau*, ou *capuchon sur le stylo*), etc.
Raison 2 : coexistence fréquente avec les prédicats basiques			*saxo* (*acquérir un saxo/touche les saxos/parler de ta saxo*), *flaque* (*se trouve la flaque d'eau/ça fait une jolie flaque*/*il y a fréquemment une petite flaque d'eau*), *rotule* (*tu mets de graisse sur les rotules*/*commander une rotule de suspension*), *capuchon* (*remettre le capuchon/la présence de capuchon*), *chiffon* (*mettre un chiffon dessous*), etc.
Raison 3 : absence de prédicats appropriés dans la liste donnée à l'avance			*rotule* (*déboiter des rotules*), *capuchon* (*dévisser le capuchon*), etc.
Raison 4 : absence de patrons syntaxiques (ou morphosyntaxiques) pour repérer les structures prédicat-argument (ou les termes composés)			*rotule* (*tu as une référence de rotule*/*la coquille s'est fondue à l'enclenchement sur la rotule*/*j'ai vu le prix de la rotule*), *chiffon* (*étaient parfaitement récollées, d'un chiffon humide*/*essuyer la carrosserie à l'aide du chiffon/un petit coup d'un chiffon sec enlève toute bidons ou des chiffons*), etc.
B	4	11.43%	*Xsara Picasso, Xantia HDI 90 CV, auto car, voiture à vocation sportive*
C	1	2.86%	*Citroën*
D	5	14.28%	*plip, filtre, boulon, chauffe, Ulysse*
Total	35	100%	

Tableau 19 Types de silences dans le résultat de la méthode distributionnelle supervisée

Pour analyser les types d'erreurs dans le résultat de la méthode distributionnelle semi-supervisée, on analyse respectivement 60 termes dans le résultat de la classe sémantique des contenants d'arrangement, de la classe sémantique des appareils de cuisson et de la classe sémantique des moyens de transport. Il y a respectivement 12, 14 et 24 erreurs dans les 60 termes de la classe sémantique des contenants d'arrangement, de la classe sémantique des appareils de cuisson et de la classe sémantique des moyens de transport. Les erreurs de chaque classe sémantique ont plusieurs explications. Bref, la plupart des erreurs sont dépendent des prédicats basiques récupérés à partir des arguments donnés à l'avance (type d'erreur A), par exemple, à partir de l'argument de contenants d'arrangement *caisse* (*torpiller la caisse*), on obtient le bruit *travail* (*torpiller le travail*). De plus, certaines erreurs sont les

termes reconnus à tort en raison de leur fonction référentielle (par ex., *reste ->remplir le reste*) (type B). Finalement, d' autres phénomènes langagiers peuvent également causer les erreurs (type C), par exemple, *croquer des voitures* (il est possible que *voiture* renvoie à un gâteau sous forme de voiture). Dans le Tableau 20, on liste respectivement les types d'erreurs, leurs proportions et les exemples correspondants à chaque classe sémantique. Cela permet d'observer que la plupart des erreurs sont des termes d'autres classes sémantiques qui dépendent des prédicats basiques récupérés à partir des arguments donnés à l'avance. De plus, le résultat de la classe des moyens de transport est moins pertinent que celui des deux autres classes sémantiques, parce que les arguments de la classe sémantique des moyens de transport coexistent plus fréquemment avec les prédicats basiques par rapport à ceux des autres classes sémantiques. Cela est dû au fait que la classe de moyens de transport possède moins de prédicats appropriés par rapport à ceux des deux autres classes sémantiques. Le nombre de prédicats appropriés qu'une classe sémantique peut posséder importe pour la performance de la méthode semi-supervisée sur cette classe sémantique.

Classes sémantiques	**Types d'erreurs**	**Nombre**	**Proportion**	**Exemples**
Contenants d'arrangement				
	A	10	83.33%	*stylo* (*trimbaler la trousse -> se trimbaler avec plein de stylos à la poche*), *travail* (*torpiller la caisse -> torpiller le travail*), *pâte/mousse* (*le classeur est rempli-> remplir de pâte/mousse*), *chose* (*transporter un frigo -> transporter de choses memes*), etc.
	B	2	16.66%	*reste* (*remplir le reste*), *contenu* (*vider le contenu*)
	Total	12	100%	
Appareils de cuisson				
	A	13	92.86%	*eau, batterie, morceau* (*réchauffer le poêle/four -> réchauffer de l'eau/la batterie/le morceau*), etc.
	C	1	7.14%	*voiture* (*croquer des voitures*)
	Total	14	100%	
Moyens de transport				
	A	23	95.83%	*stylo* (*manier la voiture -> manier le stylo*), *maison* (*louer une voiture -> louer une maison*), *cuisine* (*vibrer la voiture -> vibrer la cuisine*), etc.
	B	1	4.17%	*objet* (*concevoir son objet)*
	Total	24	100%	

Tableau 20 Types d'erreurs dans le résultat de la méthode distributionnelle semi-supervisée

En ce qui concerne les types de silences de la méthode semi-supervisée, on analyse respectivement une cinquantaine de textes et on trouve respectivement 8, 13 et 17 silences pour les classes sémantiques : contenants d'arrangement, appareils de cuisson et moyens de transport. Les silences dépendent de trois raisons principales : les termes qui ne se trouvent pas dans les structures prédicat-argument, mais dans les structures prédicat-argument non reconnues faute de patrons syntaxiques ou qui coexistent souvent avec les prédicats basiques filtrés ne peuvent être étiquetés (type de silence A) ; l'intersection des arguments élimine certains termes qui doivent être reconnus (type B) ; il manque les patrons morphosyntaxiques pour repérer les termes composés (par ex., *cuve acier*, etc.) (type C). Dans le Tableau 21, on liste les types de silences de chaque classe sémantique avec les exemples correspondants.

Pour la classe sémantique des contenants d'arrangement et la classe sémantique des moyens de transport, la plupart des silences sont provoqués par l'intersection des arguments (cf. Tableau 21). Les termes qui doivent être reconnus mais qui sont supprimés par l'intersection se trouvent fréquemment dans les structures non prédicat-argument ou les structures prédicat-argument qui n'ont pas pu être reconnues par les patrons syntaxiques incomplets, ou coexistent souvent avec les prédicats basiques filtrés. La cooccurrence insuffisante d'un argument avec les différents prédicats ne permet pas de lui assigner une fréquence d'intersection assez haute. Bien que les silences provoqués par l'intersection n'occupent pas la proportion la plus importante parmi tous les types de silences, la pertinence et la complétude de l'identification des structures prédicat-argument est toujours un facteur important pour augmenter le taux de rappel du résultat.

Classes sémantiques	**Types de silences**	**Nombre**	**Proportion**	**Exemples**
Contenants d'arrangement				
	A	1	12.50%	*cuve* (*cuve acier*)
	B	4	50.00%	*tonneau, cocotte, collecteur, sceau*
	C	3	37.50%	*cellier* (*j'ai bien une prise femelle de ce type mais dans mon cellier*), *solive* (*solives qui reposent sur deux bastaings*), *cloison* (*as _tu prévu une cloison*)
	Total	8	100%	
Appareils de cuisson				
	A	10	76.92%	*friteuse/cocotte* (*mettre dans la friteuse/la cocotte*), *sautoir* (*mettre les dolmades dans un sautoir*), *mixeur/batteur* (*acheter un mixeur/batteur*), etc.
	B	3	23.07%	*lacanche* (*le lacanche fabriqué dans le village*), *Westhal* (*chercher des retours d'expérience sur les modèles Westhal*), *casserole* (*pour les casseroles, utilisez de préférence de l'inox*)
	Total	13	100%	
Moyens de transport				
	A	7	41.17%	*Xsara Picasso, Xantia HDI 90 cv, Citroën ZX*, etc.
	B	10	58.82%	*camion* (*mettre sur le camion/j'ai un camion branché*), *camionnette* (*si la jolie camionnette bleue passe dans votre ville*), *Fiat* (*j'ai déja eu une FIAT*), *citroën* (*titre:temps d` attente livraison _ citroën _ forum marques*), etc.
	total	17	100%	

Tableau 21 Types de silences dans le résultat de la méthode distributionnelle semi-supervisée

Chapitre 2 Méthode morphosémantique

La méthode morphosémantique a pour objet d'exploiter les analyses morphologiques afin d'étiqueter le vocabulaire. Les analyses morphologiques s'effectuent en tenant compte de la sémantique des formes. Pour les unités monolexicales, la segmentation morphématique des bases permet d'obtenir les nouvelles combinaisons entre les morphèmes et cela permet de former les nouvelles bases potentielles à partir desquelles on obtient de nouvelles unités par dérivation morphologique. Pour les unités polylexicales, on les étiquette à partir des unités monolexicales à l'aide des patrons morphosyntaxiques construits en se fondant sur les analyses des structures internes et des relations sémantiques internes entre les constituants. La méthode morphosémantique est plutôt une méthode d'extension du vocabulaire. Nous choisissons un vocabulaire spécifique (les noms de métiers) et appliquons la méthode à ce vocabulaire comme objet de l'expérimentation. Dans ce chapitre, on présente d'abord le corpus adopté pour la méthode morphosémantique. Ensuite, on détaille la méthode morphosémantique (y compris la méthodologie d'analyses morphologiques et les analyses morphologiques appliquées aux noms de métiers). Finalement, on présente les résultats et leur évaluation.

1. Corpus

Le corpus constitué pour la méthode morphosémantique doit comprendre suffisamment d'informations morphosémantiques concernant le vocabulaire étudié. Les contextes ne sont pas nécessaires pour le corpus adopté dans le cadre de la méthode morphosémantique. Un corpus spécialisé riche en vocabulaire est exploitable. La méthode morphosémantique est appliquée au vocabulaire des noms de métiers. Dans cette section, on présente d'abord certains genres de textes qui sont riches en noms de métiers. Ensuite, on présente le corpus constitué pour l'expérimentation de la méthode morphosémantique.

1.1. Genres de textes dans le cadre de la méthode morphosémantique

Les informations publiées dans les blogs d'emploi ou dans les communautés d'emploi diffusent parfois des messages pour documenter un appel d'entretien, un appel de communication ou un appel de concours professionnel. Elles peuvent aussi décrire des expériences professionnelles, des astuces ou des solutions concernant les affaires professionnelles (telles qu'une solution pour le ciblage publicitaire, une astuce favorisant la communication avec les collègues, etc.). Cependant, bien que les infos sur les blogs d'emploi ou sur les communautés d'emploi soient sur la thématique d'emploi, elles ne comprennent pas beaucoup de noms de métiers.

Les discussions sur les forums d'emploi concernent souvent les expériences professionnelles, les arguments à l'appui d'une idée sur la thématique, des conseils ou des informations associées. Les discussions sur les forums comprennent de nombreuses fautes d'orthographes, des fautes grammaticales, des langages SMS et des langages Internet mais elle est riche en vocabulaire des noms de métiers.

À partir des sites de communautés ou des blogs d'emploi, on trouve souvent un espace permettant d'afficher les annonces d'emploi ou un moteur de recherche qui permet d'accéder aux bases de données concernant les offres d'emploi. Un moteur de recherche permet d'accéder aux annonces d'emploi archivées sur le site interne ou sur les sites externes.

Une annonce d'emploi comprend en général les informations suivantes : le nom du poste offert, le nom de l'organisme qui embauche, une brève description de ce dernier ou du contexte de travail, les missions du poste, le profil du candidat souhaité, les conditions du contrat (y compris le remboursement éventuel, la durée, le type de contrat, etc.) et la manière de poser la candidature. Beaucoup de descriptions comprennent des syntagmes verbaux qui correspondent à la structure Vinf + C.O.D. La structure de la langue des annonces d'emploi est simplifiée et correspond bien aussi à la structure prédicat-argument, mais les arguments ne sont pas souvent les noms de métiers. Les noms de métiers se trouvent souvent dans les titres

des annonces. La description dans une annonce d'emploi se déroule à propos de l'emploi, mais elle comprend peu d'occurrences de noms de métiers associés. Les termes décrivant les activités d'une profession comme *développer, organiser, optimiser, assurer,* …, les termes à propos de la qualification comme *être compétent de, être capable de, maîtriser, diplômé, ...* et les termes comme *recruter, embaucher, chercher* … montrent une grande présence dans les annonces d'emploi.

1.2. Corpus constitué pour la méthode morphosémantique

La méthode morphosémantique est appliquée au vocabulaire des noms de métiers. On a choisi ainsi les discussions sur les forums d'emploi, les actualités et les annonces d'emploi sur les blogs ainsi que sur les sites de communauté. Le volume de corpus sur les noms de métiers atteint 23,571 Ko, à savoir 3,754,334 mots, 18,871,715 caractères (espaces non compris) ou 22,562,359 caractères (espaces compris). Les textes de différents genres dans le corpus de noms de métiers occupent à peu près la même proportion. Le Tableau 22 liste les sites web qu'on traite pour la construction du corpus de noms de métiers :

Informations / Sites	**Genre de textes**	**Thème**	**URL**
Forum emploi	Forum	Emploi	http://www.le-forum-emploi.com
Rejoins Job	Actualités	Emploi	http://www.regionsjob.com/actualites/archives
Blog Sudouest	Blog	Emploi	http://emploi.blogs.sudouest.fr/archives
Id-carrières	Blog	Emploi	http://www.id-carrieres.com/blog/category/emploi_et_rh/
Blog emploi	Blog	Emploi	http://www.blog-emploi.com
Blog emploi	Annonces d'emploi	Emploi	http://emploi.blog-emploi.com/offre_emploi/?v=___0_0_0_0_0_0

Tableau 22 Liste de sites à aspirer pour constituer le corpus de la méthode morphosémantique

L'outil qu'on utilise pour construire le corpus de la méthode morphosémantique est encore celui qu'on développe dans le cadre de la méthode distributionnelle : RENE. Le nombre de niveaux d'aspiration, le tourne-pages, les hyperliens à aspirer pour chaque site web et les textes à extraire sur chaque page web téléchargée sont tous indiqués dans les deux fichiers de paramètres qui alimentent respectivement le module d'aspiration et le module d'extraction.

2. Présentation de la méthode

Dans la méthode morphosémantique, on fournit à l'avance une liste de noms de métiers simples et deux listes de bases (une liste de verbes et une liste de noms) permettant de produire les noms de métiers dérivés. À partir des bases données à l'avance, la segmentation et la recomposition morphématiques sont effectuées pour enrichir la liste de bases. Ensuite, les candidats noms de métiers dérivés sont produits par la dérivation morphologique. Ils sont validés à condition d'être trouvés dans le corpus. Finalement, on étiquette les noms de métiers composés à partir des noms de métiers monolexicaux (simples et dérivés) à l'aide de patrons morphosyntaxiques. Un algorithme pour mesurer l'information mutuelle de chaque candidat-terme composé est développé pour détecter les collocations. L'objectif de la méthode morphosémantique est d'enrichir le vocabulaire à partir d'un ensemble de termes donnés à l'avance en se basant sur le corpus. Dans cette section, on présente premièrement les études effectuées sur la morphologie. Deuxièmement, on présente les analyses morphologiques respectivement appliquées aux noms de métiers monolexicaux et aux noms de métiers polylexicaux. Finalement, on détaille chaque étape de la méthode morphosémantique.

2.1. État de l'art sur les études de morphologie

« La morphologie, quant à elle, se préoccupe de la forme des mots, dans leurs différents emplois et constructions, et de la part d'interprétation liée à cette forme même. » (Huot, 2001, p.9). Le champ de la morphologie peut être divisé en deux parties : la morphologie flexionnelle qui a pour objet d'étudier les variations formelles que subissent les unités lexicales en rapport avec leur fonction dans la phrase et la morphologie lexicale qui s'intéresse à la structure interne des unités lexicales. D'après Fradin (2003), la morphologie morphématique combinatoire qui a été longtemps dominante en morphologie présente des faibles pour décrire la formation des mots. Il défend les raisons conduisant à l'abandon de la notion de morphème et expose un nouveau modèle descriptif : la morphologie lexématique classique. Dans cette section, nous présentons d'abord les deux modèles descriptifs en morphologie (la morphologie morphématique combinatoire et la morphologie lexématique

classique) et exposons respectivement leurs avancées et limites. Ensuite, nous mettons l'accent sur la présentation des analyses des mots composés (y compris la discussion sur la notion des noms composés et les analyses morphologiques des noms composés).

2.1.1. Morphologie morphématique combinatoire

La morphologie morphématique combinatoire se caractérise par les hypothèses portant sur les unités minimales, la construction du complexe et la réalisation matérielle des unités linguistiques. Dans la morphologie morphématique combinatoire, on considère que les morphèmes sont les unités minimales des langues et que les expressions lexicales complexes sont construites de manière exhaustive à partir des morphèmes. La réalisation matérielle des unités linguistiques s'explique par les règles morphophonologiques qui postulent les variations phonologiques et prosodiques des morphèmes selon les différents contextes. Dans ce qui suit, on présente d'abord certaines notions basiques dans le modèle descriptif de la morphologie morphématique combinatoire. Ensuite, on expose les procédés de la formation des mots. Finalement, on présente les analyses flexionnelles et dérivationnelles.

« Un **morphème** est la plus petite unité formelle dotée d'un sens ; on dit aussi que c'est une unité significative minimale. » (Apothéloz, 2002, p.3). Les **morphèmes lexicaux** (qui constituent le lexique de la langue) et les **morphèmes grammaticaux** (qui incluent les pronoms, les articles, les prépositions, les conjonctions ainsi que les affixes de flexion et les affixes de dérivation) sont distingués. On dégage également les **morphèmes libres** s'ils ont l'existence autonome et les **morphèmes liés**. Les **affixes de flexion** (appelés aussi **flexifs**) sont des morphèmes liés et ils ne créent pas de nouveaux lexèmes mais une autre forme d'un même lexème. Les **affixes de dérivation** permettent de former les nouveaux lexèmes et de structurer une partie du lexique français en maintenant un rapport perceptible formel et sémantique. D'après Apothéloz (2002, p.14), le rapport entre une série de mots obtenus par la même dérivation (telle que *vitrerie, horlogerie, marbrerie, lingerie*, etc.) est le **rapport paradigmatique** et une telle série de mots est appelé un **paradigme**. Le rapport paradigmatique implique la possibilité d'une substitution à une position donnée. Le rapport entre *vitre* et *vitrerie* est le **rapport dérivationnel**. Le lexème construit par une opération

d'affixation est appelé **dérivé**. Les affixes comprennent les **préfixes** qui se trouvent à gauche de la base, les **infixes** qui se placent à l'intérieur de la base et les **suffixes** qui se situent à droite de la base. La **base** est l'élément sur lequel opère un affixe. Une base peut être composée d'un ou plusieurs morphèmes, par exemple, le nom *nationalisation* comporte quatre morphèmes, à savoir un morphème lexical et trois suffixes : *nation-al-is-ation*. Le morphème lexical qui subsiste après l'enlèvement de tous les affixes (les affixes de flexion et les affixes de dérivation) est appelé **radical**. Dans des commentaires étymologiques, le terme de **racine** est parfois utilisé pour désigner un élément qui a été un morphème dans un état antérieur de la langue et qu'on retrouve dans une famille de mots (Apothéloz, 2002, p.16). « Certaines racines peuvent être accompagnées d'un « allongement » dont les formes diverses répondent néanmoins à un schéma structural précis. Cet allongement constitue avec la racine un ensemble morphologique spécifique, appelé **thème**, et qui a une fonction tout à fait particulière dans le processus de construction des mots. » (Huot, 2001, p.43). Par exemple, dans *form-at-ion*, *-at* est la suite d'allongement de la racine *form-* et l'assemblage de la racine et de l'allongement *format-* constitue un thème, sur lequel les mots comme *format-ion, format-eur, format-if* …… peuvent être construits.

Apothéloz (2002, p.16) a présenté six mécanismes de formation des mots : l'**emprunt**, la **dérivation affixale**, la **dérivation non affixale**, la **composition**, la **troncation** et le **procédé du mot-valise**. Le vocabulaire d'une langue peut être enrichi par l'emprunt à une autre langue contemporaine ou ancienne. Le **calque** est une variante de l'emprunt qui consiste à emprunter un signifié sans que le signifiant soit emprunté, par exemple, *souris* provient d'un calque d'un emploi de l'anglais *mouse* ; *réaliser* vient du mot anglais *realize* ; *gratte-ciel* a été obtenu par traduction littérale de *skyscraper*. La dérivation affixale permet de créer les nouveaux lexèmes avec l'intervention d'affixes, alors que la dérivation non affixale est effectuée sans intervention d'un affixe. Il existe deux types de dérivations non affixales : le mécanisme purement sémantique (le plus souvent la métaphore ou la métonymie) et la **conversion** qui consiste à modifier la catégorie grammaticale de la base (un nom en verbe, un adjectif en nom, etc.). La composition est définie par Apothéloz (2002, p.18) comme la construction d'une unité lexicale complexe au moyen d'un morphème grammatical non

affixal et d'un morphème lexical, ou d'au moins deux morphèmes lexicaux libres ou liés. La **composition populaire** (à partir des mots français) et la **composition savante** (à partir des éléments latins ou grecs) se distinguent du point de vue de l'évolution du français. La troncation est un procédé de réduction du signifiant du lexème, tel que *périph (périphérique)*, *appart (appartement)*, *sax (saxophone)*, etc. Le procédé mot-valise a pour objet de construire un nouveau lexème à partir de plusieurs lexèmes par la composition et la troncation, tel que *autobus* (*auto*mobile+omni*bus*), *motel* (*mo*tocar+hô*tel*), *caméscope* (*cam*éra+magnéto*scope*), etc. Du point de vue de l'analyse morphologique, le mot est **simple** s'il est constitué d'un seul morphème ou **construit** s'il est constitué de plusieurs morphèmes. Les mots construits par dérivation sont appelés les **mots dérivés** et les mots construits par composition sont appelés les **mots composés**. Il existe également de nombreux **faux dérivés,** tels que *maquette, omelette* ou *ombrette* se terminant *–ette* mais qui n'ont aucun rapport avec le paradigme des dérivés en *–ette* qui concerne les dénominaux (comme *hache->hachette*) ou les déverbaux (comme *allumer->allumette*). Un mot est un **mot complexe non construit** s'il présente les propriétés suivantes : un de ses segments a la forme d'un affixe ; le segment restant n'est pas un morphème connu ; la fonction de l'affixe existe dans le sens du mot (Apothéloz, 2002, p.69-70). La forme d'affixe est nommée **pseudo-affixe** et le segment restant est appelé **pseudo-base** (Corbin, 1987).

Dans la morphologie morphématique combinatoire, la réalisation matérielle des unités linguistiques s'explique par les règles morphophonologiques. La morphophonologie est l'étude de la structure phonologique et prosodique des morphèmes, les variations phonologiques et prosodiques que subissent les morphèmes selon le contexte et les variations phonologiques et prosodiques qui remplissent une fonction morphologique. Le concept **morphe** est introduit pour désigner la manifestation concrète du morphème. Par exemple, le préfixe *dé-* (*défait*) se présente sous la forme *dés-* (*désaccord*) quand la base est commencée par une voyelle ; *dé-* et *dés-* sont ainsi deux morphes représentant le même morphème. Deux morphes sont en distribution complémentaire quand un morphe apparaît dans des contextes phonologiques où n'apparaît jamais l'autre, et réciproquement. Les morphes en distribution complémentaire sont appelés **allomorphes**. Ce type de variation est appelé **allomorphie**. Le

supplétisme (ou **supplétion**) est un type particulier d'allomorphie. Il s'agit des cas où les allomorphes ont des formes complètement différentes, par exemple, *cuisine ->culin-aire, cheveu->capill-aire, pierre->lith-ique*, etc. L'allomorphie morphophonologique et l'allomorphie morphologique sont distinguées. L'allomorphie morphophonologique illustre le cas où le conditionnement est phonologique (par exemple, *dé-* se présente sous la forme *dés-* devant la voyelle). L'allomorphie morphologique est conditionnée par la morphologie, par exemple, pour le verbe *aller*, il présente trois allomorphes aux formes de l'indicatif : *v-, all-* et *i-*. (Corbin, 1987 :283-385)

Pour certains dérivés, la segmentation en morphèmes est relativement aisée, pour certains, elle devient délicate, même indécidable. Apothéloz (2002, p.49) définit la **diagrammaticité** comme un paramètre permettant d'évaluer la lisibilité de la structure interne d'un mot construit. La diagrammaticité est analysée comme la résultante de plusieurs facteurs : la compositionnalité, la productivité et la transparence formelle. La baisse de la diagrammaticité peut avoir plusieurs causes : l'écart entre le sens prédictible et le sens **effectif**, l'existence de plusieurs parcours dérivationnels, la **coalescence** d'affixes, la **mécoupure**, l'affixation non sémantique, l'**orphelinisation** et les troncations.

Le sens d'un dérivé est compositionnel s'il est prédictible à partir des morphèmes qui le constituent, sinon, le sens d'un dérivé est effectif. Par exemple, les verbes *stérilise, banalise, stabilise* et *brutalise* sont tous des dérivés de structure $[[x]_{adj}$-*ise*$]_{v}$. Le sens de *stérilise, banalise* et *stabilise* peut être paraphrasé par « agit de façon à rendre X », alors que le sens de *brutalise* n'est pas le cas. Le sens de *brutalise* présente un écart entre le sens prédictible et le sens effectif. Certains dérivés peuvent avoir plusieurs interprétations dérivationnelles qui produisent des sens compositionnels différents, par exemple, *invalidable* peut être représenté par [[*in*-[*valide*]$_{adj}$]$_{adj}$-*able*] ou par [*in*-[[*valide*]$_{adj}$-*able*]]. La coalescence d'affixes est un phénomène de fusion d'affixes. Par exemple, les parcours dérivationnels des noms *aulnaie, bananeraie, cerisaie* et *aspergeraie* sont respectivement $[[X]_N$ -aie$]_N$, $[[[X]_N$ –(i)er$]_N$ -aie$]_N$, $[[X]_N$ -aie$]_N$ et $[[X]_N$ -eraie$]_N$. L'existence du suffixe *–eraie* s'explique par la fusion des suffixes *–er* et *–aie* en raison de leur fréquente composition. La mécoupure est une segmentation morphologique erronée de la chaîne parlée et elle aboutit à une recomposition

des morphèmes. L'affixation non sémantique contribue à faire baisser la compositionalité du dérivé. Par exemple, l'ajout du suffixe *-ier* au nom d'un arbre en ancien français *peuple* consiste à aligner le mot sur le paradigme des dérivés en *–ier* au lieu de construire le nom de l'arbre. L'orphelinisation du dérivé diminue la productivité de la base lexicale. Par exemple, dans le verbe *déceler*, *celer* est pratiquement tombé en désuétude et il est rare que les locuteurs francophones traitent encore le verbe *déceler* comme une forme construite. Les troncations comprennent les troncations liées à une opération d'affixation, les troncations simples et l'haplologie. Par exemple, quand certains adjectifs (tels que *électrique, mathématique, synthétique,* etc.) se terminant en *–ique* sont convertis en verbes par le suffixe *–ise* (tels que *électriser, mathématiser, synthétiser,* etc.), il s'agit de troncations liées à une opération d'affixation ; quand une séquence de phonèmes qui apparaît deux fois successivement est supprimée (comme *gratuit->*gratuitité->gratuité*), il s'agit de l'haplologie ; les formes comme *appart* (*appartement*), *dissert* (*dissertation*), *médic* (*médicament*)… sont les troncations simples.

Il n'est pas évident de délimiter la frontière entre mots préfixés et mots composés quand le premier élément constitutif n'est ni un verbe ni un nom ni un adjectif. Par exemple, si *polylexical, contre-emploi* et *micro-onde* sont considérés comme des composés, l'avis est-il le même pour *polycéphale, contrepartie* et *microchirurgie* ? *bi- (bigame), anti-(antiride)* et *pluri-(plurilingue)* devraient-ils être considérés comme préfixes ou comme parties des mots composés ? Huot (2001, p.95) a donné une définition de mots composés pour les distinguer des préfixes : « quels qu'en soient l'origine et le degré d'autonomie, les prépositions pourvues d'un sens plein, les éléments ayant une interprétation locative, temporelle ou quantitative devraient être considérés comme des parties de mots composés dès lors que le terme dans lequel ils figurent (avec ou sans trait d'union) est pourvu d'une interprétation unique, mais dans laquelle ils restent repérables. ». D'après Huot (2001), *bi- (bicentenaire), circum-(circumpolaire), demi-(demi-frère), infra- (infrarouge), non-(non-fumeur), mono-(monogame), sub-(subtropical)*…… sont tous des éléments de noms composés. Apothéloz (2002, p.73) distingue deux groupes d'affixes dérivationnels : **les affixes transcatégoriels** qui ont pour fonction de changer la catégorie grammaticale de la base et **les affixes**

intracatégoriels qui ont pour fonction d'opérer sur le signifié de la base sans modifier sa catégorie grammaticale. Il existe deux types d'affixes intracatégoriels : les suffixes qui ajoutent une spécification sémantique à la base (tels que *fille->fillette, célèbre->célèbrissime, éléphant->éléphanteau*, etc.) et les suffixes qui créent une nouvelle notion par exemple en transformant un mot désignant un objet en un mot désignant un être humain (*milliard->milliardaire, rosier->roseraie, bureau->bureautique,* etc.). Certains suffixes sélectionnent leur base selon des critères sémantiques (par ex., le suffixe adverbial *–(e)ment* sélectionne des bases adjectivales dont le sens qualifie un procès) et certains nécessitent la troncation des bases en raison des contraintes morphophonologiques, par exemple, pour le suffixe *–if*, il nécessite la troncation en transformant un nom en un adjectif comme *désignation->désignatif, révulsion->révulsif, défense->défensif*, etc. Certains verbes pourraient être analysés comme dérivés des locutions, tels que *atterrit (à terre), entasse (en tas), enterre (en terre), attable (à table),* etc. Ces verbes sont analysés par la **dérivation délocutive**.

La conversion consiste à modifier la catégorie grammaticale de la base sans aucune modification formelle. Elle est aussi appelée l'affixation zéro. La conversion aboutit à un double modèle dérivationnel. Par exemple, pour le nom *offense* et le verbe *offense* de la 3^{e} personne du singulier au présent, la dérivation peut être interprétée comme un modèle pour dériver du verbe au nom ou comme un modèle pour dériver du nom au verbe.

Les morphèmes flexionnels se distinguent des morphèmes dérivationnels. Ils ne peuvent apparaître qu'en fin de mot et ne permettent pas de former les nouveaux mots construits. Ils sélectionnent la base ou mot construit pour s'adjoindre mais ne déterminent pas la catégorie des mots issus de leur adjonction. Ils n'ont que la valeur grammaticale. La plupart d'entre eux ne sont pas syllabiques (mais un seul morphème consonantique) et sont soumis à la règle de troncation. Dans les langues de la famille indo-européenne, les principales indications grammaticales fournies par les morphèmes flexionnels sont : les cas issus du latin et qui ont disparu du français, le genre, le nombre, la personne, le temps, le mode et l'aspect. Le français moderne ne connaît que les morphèmes flexionnels de genre, de nombre, de personne, de temps et de mode.

2.1.2. Morphologie lexématique classique

La morphologie lexématique classique est distinguée de la morphologie morphématique combinatoire par deux traits : « l'unité de base est le lexème et non le morphème (d'où son nom), les mécanismes de construction des unités complexes sont processuels : les unités morphologiques complexes ne résultent plus de la combinaison d'unités atomiques mais de l'application de fonctions à un lexème. » (Fradin, 2003 : 79).

La morphologie lexématique classique pose le **lexème** comme le signe linguistique minimal hors emploi. Il est une unité abstraite sous-spécifiée pour la flexion correspondant aux unités linguistiques lexicales. Le **grammème** est l'unité abstraite correspondant aux unités linguistiques grammaticales. L'infinitif est aussi considéré comme une forme flexionnelle qui se rapporte à un lexème, par exemple, *manger, mangeait* et *mangerai* sont tous des instances du lexème MANGER. La morphologie lexématique classique (MLC) envisage la construction des unités complexes de la manière suivante : les unités construites complexes sont le résultat de l'application d'une série de règles morphologiques (les règles flexionnelles, certaines règles constructionnelles qui prennent un argument, ainsi que les règles de composition et d'incorporation qui en prennent deux) et ces règles sont des fonctions qui régissent d'autres fonctions ; l'affixation est un procédé morphologique qui présenterait un caractère diagrammatique plus fort que d'autres procédés ; les procédés comme la réduplication, la soustraction, le changement tonal ou accentuel sont décrits comme le résultat de fonctions appliquées pour les changement requis ; la fonction sémantique associée au **trait-valué** (qui est un trait auquel est associé une valeur, par ex., [ger :mas]) est uniforme quel que soit la réalisation du trait en question (par exemple, le pluriel des mots *oiseau (oiseaux), œuf (œufs), émail (émaux), œil (yeux) et ail (aulx)* se manifeste par de diverses marques mais le résultat est interprété par la fonction unique «rendre au pluriel».) ; l'ordre des affixes reflète l'ordre d'application des fonctions morphologiques ; les mécanismes qui construisent les lexèmes complexes ne produisent pas les structures internes. Les fonctions s'appliquent à toutes les dimensions du lexème. Elles se divisent en deux types : les règles morphologiques et les opérations morphologiques. « Les changements qu'une règle

morphologique lexicale entraîne dans le substrat sont effectués par les opérations morphologiques. Ces opérations se regroupent en familles : affixation, apophonie, réduplication, méthathèse, soustraction » (Fradin, 2003, p.117). Chaque règle est associée au moyen d'un ensemble d'opérations. Par exemple, *dépersonnaliser* s'analyse par les opérations : $FCT_{A<N}(X)$=Xal, $FCT_{V<N}(X)$=Xiz et $FCT_{inv\text{-}V}(X)$=déX dans lesquelles $FCT_{A<N}$ est la fonction de transformation de l'adjectif en nom et X indique l'argument (*personne*) pris par la fonction. Les règles morphologiques mettent en rapport la face matérielle des signes linguistiques (le **substrat**) et leur contenu (l'**abstrat**). « Cependant, une grande partie des phénomènes affectant le substrat phonique ne peut pas être traitée au moyen d'opérations morphologiques telles que les envisage la MLC. En particulier, tous les phénomènes où intervient la prosodie demeurent hors de sa portée. » (Fradin, 2003, p.132).

Le modèle MLC présente des avancées pour la description des phénomènes qui soulèvent des problèmes dans la morphologie morphématique combinatoire, tels que les lexèmes à radicaux multiples, les lexèmes formellement complexes non construits, la conversion, etc. Le fait que certains lexèmes peuvent avoir plusieurs radicaux s'explique par la sélection du radical en fonction de **l'ensemble-valué** (qui est un ensemble cohérent de traits munis de leur valeur, par ex., [ger :fem, nb :sg, per :3]) associé au verbe. Par exemple, le verbe *acquérir* a deux radicaux : /aker/ (qui sert pour les personnes 4 et 5 du présent, pour l'imparfait, le prétérit et le participe passé, l'infinitif et le future/conditionnel) et /akjɛr/ (qui sert pour la personne 6 et les autres personnes du présent ainsi que pour le subjonctif présent) ; *démontrer* a aussi deux radicaux : l'un qui sert pour le verbe et une partie des dérivés (*démontrable*, *redémontrer*) et l'autre qui sert pour les dérivés dont l'origine est dite savante (*démonstration, démonstratif, démonstrateur*). Les lexèmes formellement complexes mais non construits sont les unités dont les constituants ne sont ni catégorisables ni dotés d'une signification ; leur interprétation n'est pas prédictible et leur forme ne peut être assignée à une distorsion identifiée par la grammaire (Fradin, 2003 : p.140). Les lexèmes formellement complexes non construits où c'est la base lexicale qui est identifiable se distinguent de ceux où c'est l'**exposant** d'une règle morphologique qui l'est. L'exposant est le moyen de réalisation formelle du procédé de construction. Pour les premiers, la description se fait par

l'introduction du trait "base" dans la règle morphologique. Par exemple, pour *royaume*, le trait "base" *roi* est introduit pour pointer sur le lexème qui constitue la base du lexème complexe non construit (cf., Figure 31). Les lexèmes complexes non construits où seulement l'exposant de la règle morphologique est identifiable sont décrits comme des sorties appauvries d'une règle morphologique. Par exemple, le lexème complexe non construit *peuplier* partage les informations rassemblées sous (I) avec *poirier (pommier, mirabellier…)* à l'exception des prédicats concernant le fruit (cf., Figure 32).

(G)	royaume#
(F)	(rwa.jom)
(SX)	cat:n
(M)	base:roi#
(S)	royaume'

Figure 31 Exemple "royaume"

	I	II
(G)	peuplier#	poirier#
(F)	(pØ.pli.je)	(pwa.rje)
(SX)	cat:n, ger:mas	cat:n, ger:mas
(S)	arbre'•x	arbre'•x ∧ pousser'•y•(sur•x) ∧ fruit'•y ∧ poire'•y ∧...

Figure 32 Exemple de "poirier"

De la même manière, la conversion est décrite par la règle morphologique qui indique le changement des traits valués du mot-forme (tels que, la catégorie grammaticale, la sémantique, la phonologie, etc.). Par exemple, la règle de conversion pour les déverbaux féminins (comme *nager->nage, récolter->récolte, épater->épate*) est :

	I	II
(G)	$X^{\#}$	$X^{\#}$
(F)	(…)	(…)
(SX)	cat:v	cat:n ∩ ger:fem
(S)	$(V'•ev•x_1...x_n)$	$(\cap V'•ev•X_2...X_n \wedge \neg agent•X_2)$

Figure 33 Exemple de conversion

La MLC est élaborée pour traiter les phénomènes de flexion. Elle connaît des limites dans le domaine de la construction des lexèmes. En conséquence, un nouveau mode de traitement est inventé pour la morphologie constructionnelle. D'après Fradin (2003, p.192), une unité est complexe si elle analysable en formants identifiables et une unité est définie comme construite si elle peut être obtenue par les procédés formels disponibles en langue.

Une unité complexe peut être construite ou non. Elle peut être construite par la morphologie, la syntaxe, ou des procédés de grammaticalisation.

Les **procédés de la morphologie constructionnelle non savante** du français comprennent l'affixation, la conversion, la réduplication, la composition **VN** et la composition **NN**. Les unités construites par la composition sont polylexématiques (appelées noms composés). La composition VN (un verbe transitif et un nom) construit les noms composés dénotant l'objet ou l'humain qui 'Ver N' (par ex., *tire-bouchon, crève-cœur,* etc.). La composition NN construit les noms composés N_1N_2 dans lesquels N_1 indique le type du référent du nom composé ou une propriété (constitutive ou fonctionnelle) de N_2 caractérisant N_1. Les unités dans lesquels figure un élément phonologiquement non distinct d'une préposition (*contre-épreuve, sous-sol, avant-trou…*) sont considérées comme formées par préfixation, parce que la composition présente une proximité avec la syntaxe alors que les prépositions qui se construisent avec un verbe n'apparaissent jamais au début de verbes (Fradin, 2003, p.197). Les **procédés savants** de construction de lexèmes mettent en jeu des unités appartenant à d'autres langues, par exemple, *mètre* dans *hydromètre* qui signifie *mesurer* provient d'un mot grec.

Les noms composés provenant de la lexicalisation des unités syntaxiques sont des **unités syntaxiquement construites**. « La composition nominale est une micro-syntaxe » (Benveniste, 1967 :15-31). Il existe plusieurs types de procédés construisant ces unités : le type de synapsies-DE (*heure de pointe, fil de fer…*), le type de synapsies-A (*avion à réaction, moulin à vent, brosse à dents…*), le type de syntagmes SN (*sauce à l'ail, os à la moelle…*), le type de syntagmes NA (*poids lourd, chambre froide…*), etc. « Les synapsies sont des syntagmes nominaux formant une unité lexicale stabilisée » (Benveniste, 1974 : p.172). La **morphologie grammaticale** se distingue de la **morphologie extragrammaticale**. Dans la morphologie grammaticale, on discerne aussi la morphologie prototypique (centrale) et la morphologie marginale (périphérique). Les **unités construites par des moyens extragrammaticaux** sont des mots crées sans relever à proprement dit de la grammaire. Les procédés extragrammaticaux mettent en jeu des processus linguistiques qui n'appartiennent pas à la grammaire de la langue étudiée mais qui correspondent quand même aux exigences

de la grammaire universelle, telles que les mots-échos (*glouglou, flip flop, pif paf…*), les mots-valises (*ridicoculiser, délyrer, ennuiversel…*), la siglaison (*SMIC, CUC, ONU…*), etc. « La morphologie prototypique est celle qui met en œuvre les procédés les moins marqués, c'est-à-dire les plus diagrammatiques (affixation) et qui sert à des fins descriptives. » (Fradin, 2003, p.207). La morphologie périphérique se situe sur les marges de la morphologie centrale. Les **procédés périphériques** comprennent l'accourcissement, la réduplication et la **suffixation poly-lexématique** (tel que, *fleurdelisé < fleur de lys, charcutier < chair cuite, mainmortable < main morte, quarderonner < quart de rond…*).

2.1.3. Analyse des noms composés

D'après Saussure (1982, p.242), les noms composés sont formés par la composition qui est le fait de « deux ou de plusieurs termes originairement distincts, mais qui se rencontrant fréquemment en syntaxe, au sein d'une phrase, se soudent en une unité absolue et difficilement analysable ». Grevisse (1986) distingue le nom composé du syntagme en fonction des critères syntaxiques : le nom composé est une unité lexicale permanente alors que le syntagme est une forme libre occasionnelle. Gross G. et al. (1986) décrit les noms composés en introduisant la notion de figement : « La notion d' "idée unique" qui correspond à celle de signifié unique ne concerne que les mots composés "figés" ceux dont aucun élément ne peut faire l'objet d'un choix : […] le figement peut affecter l'ensemble ou un seul des éléments constituants, avec toutes les combinaisons intermédiaires possibles. ». L'introduction de la notion de figement permet de distinguer les noms composés des formes libres : « […] le critère de discrimination est le degré de figement. Il se mesure en appliquant à la structure les transformations normalement admises par un syntagme libre de même nature et en notant l'acceptabilité des formes obtenues » (Jacquemin, 1991, p.18).

Cependant, ces définitions ne permettent pas de répondre à la question suivante : comment délimiter les noms composés des noms préfixés ? Si *contre-épreuve* est considéré comme composé, comment doit-on considérer *contrepartie* ? Si *anthropophage (anthropo-)* est considéré comme composé savant dans la tradition grammaticale, où faut-il ranger *téléphone (télé-), antitabac (anti-) ou automobile (auto-)* ? Sur la distinction entre les

composés et les préfixés, une vive discussion existe depuis longtemps. D'après Apothéloz (2002, p.18-19), les noms composés sont les mots formés par la composition qui est définie comme la construction d'une unité lexicale complexe au moyen d'un morphème grammatical non affixal et d'un morphème lexical, ou d'au moins deux morphèmes lexicaux libres ou liés ; le problème de distinguer les composés des préfixés provient de la difficulté de discerner les morphèmes affixaux des morphèmes lexicaux. D'après Huot (2001, p.95), « quels qu'en soient l'origine et le degré d'autonomie, les prépositions pourvues d'un sens plein, les éléments ayant une interprétation locative, temporelle ou quantitative devraient être considérés comme des parties de mots composés dès lors que le terme dans lequel ils figurent (avec ou sans trait d'union) est pourvu d'une interprétation unique, mais dans laquelle ils restent repérables.». Ainsi, *bi- (bicentenaire), circum- (circumpolaire), demi-(demi-frère), infra- (infrarouge), non-(non-fumeur), mono-(monogame)…*sont tous considérés comme parties des composés. Fradin (2003, p.197) a tenté de résoudre le problème du point de vue de la lexicalisation : puisque la composition présente une proximité avec la syntaxe et que les prépositions qui se construisent avec un verbe n'apparaissent jamais au début de ce verbe, les unités dans lesquels figure un élément phonologiquement non distinct d'une préposition (*contre-épreuve, sous-sol, avant-trou…*) doivent être considérées comme formées par préfixation. Guilbert (1963) classe les éléments selon qu'ils ont une nature essentiellement grammaticale ou lexicale : les unités qui ne sont formées que d'éléments de nature lexicale sont composées et les unités précédées d'éléments de valeur prépositionnelle ou adverbiale sont préfixées.

En linguistique informatique, le mot est défini comme une chaîne de caractères séparés par deux blancs. L'identification des catégories polylexicales devient un travail qui a pour objet de chercher les associations privilégiées entre les mots voisins (Benson et al., 1986). Smadja (1993) a défini les noms composés comme « une combinaison récurrente de mots qui se trouvent ensemble plus souvent que par le simple fait du hasard et qui correspondent à une utilisation arbitraire ».

Mejri (1997, p.133-136) a présenté un ensemble de différents critères de classement des noms composés selon leurs structures internes. D'après le constituant qui joue le rôle de

base, les noms à base verbale (*cure-oreille, trompe-la-mort*) et les noms à base nominale (*bateau-feu, aide-soignante*) peuvent être distingués. Dans le cas où il s'agit du figement, il distingue deux types de composés selon qu'ils ont une structure conforme ou non aux règles de la langue. Par exemple, *cercle vicieux* et *grande surface* sont conformes aux règles de la langue, alors que *café-filtre* et *cheval-vapeur* ne le sont pas. En se référant à Darmesteter (1875) qui oppose la composition (reposant sur l'ellipse) et la juxtaposition (reposant sur la soudure sans ellipse), on a un critère purement syntaxique (absence ou présence d'ellipse) permettant de distinguer les composés (*cure-oreille* et *bateau-feu*) et les juxtaposés (*trompe-la-mort* et *aide-soignant*), mais selon lequel on hésite devant *pomme de terre* en raison de l'absence du déterminant. En fonction du critère qui tient compte de la relation entre les réalités dénommées et les constituants du nom composé, *bateau-feu* et *aide-soignant* renvoient à des réalités dénommées par l'un de leurs constituants (soit *bateau* et *aide*), alors que *cure-oreille* et *trompe-la-mort* ne comportent aucun constituant permettant d'indiquer leurs réalités dénommées.

Fotopoulou (1996) a présenté une série de transformations de structure constatées lors de l'étude des structures composées : la surcomposition, l'ellipse, les modifieurs et les variantes. Fotopoulou (1996, p.93) distingue la surcomposition et les modifieurs selon la possibilité de transformation du constituant surajouté. Par exemple, dans *convertisseur élévateur de fréquence*, *convertisseur de fréquence* est considéré comme la forme de base et *élévateur* ne peut être remplacé que par *abaisseur* (c'est-à-dire que *élévateur* n'est pas un modifieur libre et il est ajouté par la surcomposition), alors que dans *réseaux mondiaux de télécommunication*, la forme de base est *réseaux de télécommunication* et *mondiaux* peut être remplacé par *internationaux* sans changer le sens du terme de base (c'est-à-dire que *mondiaux* est un modifieur libre). Sur l'ellipse, « un nom composé peut être évoqué par un nom composé elliptique où un ou plusieurs de ses éléments ont disparu » (Fotopoulou, 1996, p.93), par exemple, *gain (isotrope) partiel, les (téléphones) sans fil*, etc. Deux principales variantes sont aussi présentées : les variantes orthographiques (caractère optionnel du trait d'union, par ex., *émetteur-récepteur/émetteur récepteur*) et les variantes morphosyntaxiques (simplification de la structure du nom composé en éliminant la préposition ou le déterminant,

par ex., *tension d'hélice/tension hélice*).

Mathieu-Colas (2009) a proposé une typologie de noms composés en fonction des analyses de leurs structures internes. La description des noms composés se fonde sur le classement morphologique. On ne retient que les composés graphiques (ceux qui sont articulés par séparateurs, tels que trait d'union, apostrophe, espace, etc.) en tenant compte du traitement informatique. 17 classes élémentaires (EMPRUNTS, ONOMATOPEES et mots assimilés, Composés sur PARTICULES, Composés NOM+de+X, Composés NOM+en+x, etc.) et 8 classes complémentaires réservées aux composés complexes (Expansions de A+N, Expansions de N+N, etc.) sont établies. Les composés complexes mettent en jeu plus de deux termes nominaux et/ou adjectivaux (Mathieu-Colas, 2009, p.3). Deux types de formations des composés complexes sont distingués : la surcomposition (*Moyen Age->haut Moyen Age, libre échange->zone de libre-échange, épingle à cheveux->virage en épingle à cheveux*) et les syntagmes figés (*beau brin de fille ->*beau brin/*brin de fille, bon vieux temps->*vieux temps, cas de force majeure->*cas de force/*force majeure*) (Mathieu-Colas, 2009, p.44). De plus, chaque classe groupe encore les sous-classes éventuelles et 700 types de noms composés sont recensés en total. Ce classement a pour objet de rendre compte de la structure morphologique des composés qui permet de faciliter l'analyse des procédés compositionnels ainsi que la saisie et le traitement informatique de classes formellement homogènes (Mathieu-Colas, 2009, p.4).

2.2. Analyses morphosémantiques des noms de métiers

L'analyse morphosémantique des noms de métiers se décline en deux parties : l'analyse morphologique et l'analyse sémantique. L'analyse morphologique a pour objet d'étudier la structure formelle des noms de métiers[4] dans le but de reconnaître les candidats-termes en fonctions des structures correspondantes. L'analyse sémantique consiste à étudier la relation sémantique entre les noms de métiers et leurs constituants dans le but de délimiter les

[4] Toutes les observations sur les noms de métiers sont faites à partir de la ressource lexicale (comprenant 5185 noms de métiers) fournie par P.-A. Buvet.

noms de métiers. L'analyse morphologique dans notre méthode se fonde sur la morphologie morphématique combinatoire. En tenant compte du traitement automatique, on distingue les noms de métiers monolexicaux et les noms de métiers polylexicaux. On ne considère que les composés graphiques (ceux qui sont articulés par des séparateurs, tels que le trait d'union, l'apostrophe ou l'espace) comme unités polylexicales. Les composés sans séparateurs (*psychanalyste, psychiatre, musicothérapeute, cartomancien*), les dérivés (*chauffeur, chanteur, comédien*) et les simples (*syndic, choreute, chef, garde, expert, maître, maçon, porion, maquignon, mannequin, marin, médecin, camelot, matelot*) sont tous considérés comme des unités monolexicales. L'analyse morphologique des unités monolexicales est effectuée à partir des morphèmes et celle des unités polylexicales est réalisée à partir du lexique.

2.2.1. Analyses morphosémantiques des noms de métiers monolexicaux

Les noms de métiers dérivés peuvent avoir les suffixes comme *–eur* (*chamoiseur, chanfreineur, chasseur, chauffeur, chroniqueur, cisailleur*), *-ien* (*clinicien, comédien, cosméticien, électricien, mécanicien, esthéticien*), *-graphe* (par ex., *géographe, hydrographe, iconographe, infographe, typographe, aérographe, cartographe, chorégraphe*), *-iste* (*journaliste, buraliste, camiste, caricaturiste, cariste, cellophaniste*), *-logue* (*cosmétologue, étheirologue, gemmologue, géologue, graphologue, gynécologue, hydrogéologue, iridologue*, etc.), *-ier* (par ex., *ouvrier, ivoirier, jardinier, joaillier, journalier, laitier, layetier, licier*, etc.), *-iatre* (par ex., *pédiatre, psychiatre,* etc.), ou *-er* (par ex., *boulanger, horloger, vacher, boucher,* etc.), etc. Il existe également de nombreux noms de métiers empruntés, tels que *speaker, doker, designer, manager…* qui ont un suffixe d'une langue étrangère. (cf., Annexe 1, section 3)

D'après Apothéloz (2002, p.18), les mots construits au moyen d'un morphème grammatical non affixal et d'un morphème lexical, ou d'au moins deux morphèmes lexicaux libres ou liés sont les mots composés. Les noms comme *hydrogéologue, hydrographe, pédiatre, ostéopathe* sont ainsi considérés comme composés du fait que *–graphe, -iatre, -pathe, hydro-, pédi-*, ainsi que *ostéo-* sont les morphèmes lexicaux. Nous appelons ces noms

de métiers les composés sans séparateurs pour les distinguer des composés graphiques dans le cadre de la linguistique informatique (Mathieu-Colas, 2009). De plus, en enlevant les suffixes des noms de métiers dérivés, on trouve que la base de certains d'entre eux est elle-même un construit (un dérivé ou un composé). Par exemple, la base de *cartomancien* est *cartomancie* qui est construite au moyen de deux morphèmes lexicaux : *carto-* et *mancie* ; la base *magnétothécaire* est *magnétothèque* qui est composé de deux morphèmes lexicaux : *magnéto-* et *thèque* ; la base *anesthésiste* est *anesthésie* qui est dérivé d'*esthésie*. (cf., Annexe 1, section 4)

Cependant, toutes les unités lexicales se terminant par un suffixe appartenant au vocabulaire des noms de métiers ne sont pas des noms de métiers. Par exemple, les noms se terminant en *–iste* : *liste, triste, touriste, sexiste, optimiste, pessimiste, piste,* … ne sont pas des noms de métiers ; les noms se terminant en *–eur* : *antérieur, couleur, erreur, extérieur, intérieur, faveur, hauteur, honneur, humeur, vigueur, valeur, accélérateur* … ne sont pas des noms de métiers ; pour ceux qui se terminent en *–ien, acinétien, acœlien, ancien*, … ils ne sont pas des noms de métiers non plus. Parmi les unités lexicales se terminant en *–iste, -eur, -ien*…mais qui ne sont pas des noms de métiers, certaines sont des faux dérivés comme *erreur, couleur, ancien, frein, chocolat, résultat, ficaire, triste…* et certaines sont des dérivés dont les bases sont sémantiquement étrangères à une activité professionnelle, telles que *accélérateur* (*accélérer*), *acinétien* (*acinète*), *diluvien* (*diluvium*), *sexiste* (*sexe*), *abolitionniste* (*abolition*), *filaire* (*fil*), etc. Si l'on remonte au radical des noms de métiers, on trouve que certains noms de métiers sont dérivés du même radical, mais ils appartiennent à différentes classes sémantiques en raison des différents parcours de dérivation. Par exemple, *aspirateur* (*[[aspirer]$_v$-eur]$_n$->aspirateur*) est un nom d'artefact, alors que *aspirant* (*[[[aspirer]$_v$-ant]$_{adj}$]$_n$->aspirant*) est un nom de métier ; *compteur* (*[[compter]$_v$-eur]$_n$->compteur*) est un nom d'artefact, alors que *comptable* (*[[[compter]$_v$-able]$_{adj}$Ø]$_n$->comptable*) est un nom de métier ; *informaticien* (*[[[[informer]$_v$-ique]$_n$-ien]$_n$->informaticien*) est un nom de métier, alors que *informateur* (*[[informer]$_v$-eur]$_n$->informateur*) ne l'est pas. Le fait que certaines unités lexicales dérivées partagent le même suffixe que les noms de métiers mais ne le sont pas s'explique par la sémantique de la base de ces unités lexicales dérivées. Dans ce qui suit,

on présente une analyse de la relation sémantique entre les noms de métiers et leurs bases dans le but de délimiter les noms de métiers.

Certains noms de métiers proviennent des noms qui sont sémantiquement associés à l'objet du travail d'une profession, tels qu'*accessoiriste* (*accessoire*), *aciériste* (*acier*), *ingénieur* (*engin*), etc. La relation sémantique entre ces noms de métiers et leurs bases peut être paraphrasée par « X est la personne qui V+Y», par exemple, *aciériste* (*acier*) : « aciériste est la personne qui fabrique l'acier ». Il existe également de nombreux noms de métiers dérivés des noms de domaines ou de disciplines, tels que *graphologue* (*graphologie*), *chirurgien* (*chirurgie*), *informaticien* (*informatique*), etc. La relation sémantique entre ces noms de métiers et leurs bases peut être paraphrasée par « X est la personne qui travaille en Y» ou « X est spécialiste en Y », par exemple, *psychologue* (*psychologie*) : « psychologue est spécialiste en psychologie ». Un nom de métier peut aussi être dérivé d'un nom décrivant l'activité professionnelle, tels que *cartomancien* (*cartomancie*), *manutentionnaire* (*manutention*), etc. La relation sémantique entre X et Y est paraphrasée par « X est la personne qui fait Y», par exemple, *cartomancien* (*cartomancie*) : « cartomancien est la personne qui fait la cartomancie». *mandataire* est aussi un nom de métier dérivé d'un nom (*mandat*) décrivant l'activité professionnelle, mais dans la relation sémantique entre *mandataire* et *mandat*, *mandataire* est plutôt un récipient au lieu d'un acteur. La relation sémantique entre ce type de noms de métiers et leurs bases est paraphrasée par « X est la personne qui reçoit Y », par exemple, *mandataire* (*mandat*) : « mandataire est la personne qui reçoit le mandat ». Il y a aussi les noms de métiers dérivés des noms décrivant les missions exercées, par exemple, *ambassadeur* (*ambassade*) : « ambassadeur est la personne qui gère l'ambassade ». Un nom de métier peut aussi être formé à partir d'un nom de lieu, par exemple, *laborantin* (*laboratoire*) : « laborantin est la personne qui travaille dans un laboratoire ». Pour certains noms de métiers, la relation entre eux et leurs bases est la métonymie, tels que *bâtonnier* (*bâton*), *motard* (*moto*), etc.

Un nom de métier peut aussi être dérivé d'un verbe qui décrit l'activité professionnelle, tel que *directeur* (*diriger*), *assistant* (*assister*), *conseiller* (*conseiller*), *développeur* (*développer*), etc. La relation sémantique entre ces noms de métiers (X) et leurs bases (Y)

peut être paraphrasée par « X est la personne qui Y », par exemple, *conseiller* (*conseiller*) : « conseiller est la personne qui conseille » ; *assistant* (*assister*) : « assistant est la personne qui assiste » ; *directeur* (*diriger*) : « directeur est la personne qui dirige ». Il existe également certains noms de métiers qui sont dérivés des adjectifs par la conversion, tels que *comptable* (*comptable*), *responsable* (*responsable*), *commercial* (*commercial*), etc. Dans le Tableau 23, on liste la typologie de relations sémantiques recensée entre les noms de métiers monolexicaux et leurs bases.

Types sémantiques	**Exemples**
X est la personne qui V+Y(objet)	*appareilleur* (*appareil*), *aciériste* (*acier*), *accessoiriste* (*accessoire*), *batelier* (*bateau*), *bijoutier* (*bijou*), *mercier* (*mercerie*), *métallier* (*métal*), *archiviste* (*archive*), *volailler* (*volaille*), etc.
X est la personne qui travaille /est spécialisé en Y (domaine/discipline)	*artiste* (*art*), *analyste* (*analyse*), *banquier* (*banque*), *agronome* (*agronomie*), *comédien* (*comédie*), *électricien* (*électrique*), *géologue* (*géologie*), *musicothérapeute* (*musicothérapie*), etc.
X est la personne qui fait Y(activité)	*cartomancien* (*cartomancie*), *manutentionnaire* (*manutention*), *anesthésiste* (*anesthésie*), *gestionnaire* (*gestion*), etc.
X est la personne qui reçoit Y (activité)	mandataire (mandat), etc.
X est la personne qui a Y(mission)	*ambassadeur* (*ambassade*), etc.
X est la personne qui travaille dans Y(lieu)	*laborantin* (*laboratoire*), etc.
Y est métonyme de X	*bâtonnier* (*bâton*), *motard* (*moto*), etc.
X est la personne qui Y(verbe)	*conseiller* (*conseiller*), *vendeur* (*vendre*), *abatteur* (*abattre*), *bobinier* (*bobiner*), *assistant* (*assister*), *commerçant* (*commercer*), *tisserand* (*tisser*), *peintre* (*peindre*), *écrivain* (*écrire*), *magistrat* (Empr. du lat. magistratus dér. du verbe magister), etc.
X <- Y(adjectif)	*comptable* (*comptable*), *responsable* (*responsable*), *commercial* (*commercial*), *externe*, *interne*, etc.
Composés non dérivés	*manucure* (*manu/cure*), *pédicure* (*pédi/cure*), etc.

Tableau 23 Analyses des relations sémantiques internes

2.2.2. Analyses morphosémantiques des noms de métiers polylexicaux

Un nom de métier polylexical (composé) est souvent formé à partir d'un nom de métier monolexical en ajoutant une expansion, tel que, *adjoint → adjoint de sécurité*, *accordeur → accordeur de piano*, *directeur → directeur de service, assistant → assistant administration, compagnon → compagnon routier, biologiste → biologiste médical*, *boucher → boucher chevalin*, *accompagnateur → accompagnateur à la mobilité, accompagnateur → accompagnateur en moyenne montagne*, etc. L'expansion ajoutée sous-spécifie en général le genre du travail que le nouveau nom de métier implique.

L'expansion ajoutée peut être une séquence introduite par une préposition, telle que de+N/GN, en+N/GN, sur+Det+N/GN ou à+Det+GN/N dans lesquelles le nom (N) introduit par la préposition peut être non seulement un nom monolexical (par ex., *sécurité* dans *adjoint de sécurité*) mais aussi un nom polylexical (par ex., *machine à imprimer* dans *conducteur d'une machine à imprimer*). L'expansion ajoutée peut également être un adjectif (par ex., *technique* dans *éducateur technique*) ou même plusieurs adjectifs (par ex., *administratif* et *territorial* dans *adjoint administratif territorial*), un nom monolexical (par ex., *administration* dans *assistant administration*), un nom polylexical (par ex., *bilan carbone* dans *expert bilan carbone*), ou un groupe nominal (par ex., *conseiller vie privé*). Parfois, l'expansion est une construction d'un verbe au participe passé (par ex., dans *journaliste spécialisé en environnement*) qui correspond à la structure Vpp+en+N. La structure des groupes nominaux dans les noms de métiers composés est flexible et variable. Il peut y avoir les groupes nominaux de type N+de+N (par ex., le GN *ouvrages d'art* dans *constructeur d'ouvrages d'art*), NN (par ex., *produits écoconception* dans *ingénieur produits écoconception*), NA (par ex., *la conduite automobile* dans *enseignant de la conduite automobile*), ou même les groupes nominaux qui contiennent des autres groupes nominaux, par exemple, N+A+de+Det+N (par ex., *la protection judiciaire de la jeunesse* dans *éducateur de la protection judiciaire de la jeunesse*).

Un nom de métier composé peut aussi être formé à partir d'un nom de métier monolexical en ajoutant un morphème lexical lié (technico-, bi-, demi-, électro-, éco-, podo-, médico-, télé-, oto-) ou un mot grammatical (contre-, sous-) à gauche, tels que *technico-commercial, télé-commercial, podo-orthésiste, électro-technicien, éco-concepteur, médico-psychologue, sous-chef,* etc. Dans cette thèse, on n'évalue pas la délimitation des préfixes des éléments composés. On appelle tous les éléments ajoutés à gauche des préfixes. Pour certains noms de métiers empruntés (par ex., les emprunts anglais), l'expansion en N/GN est ajoutée en avant, par exemple, *game designer, campus manager, web planner, data protection manager,* etc.

Il existe également les noms de métiers composés qui sont formés par la combinaison de deux noms de métiers (NMP_1+NMP_2). La combinaison peut être réalisée en juxtaposant deux

noms de métiers par un espace, par une préposition ou même par une conjonction *et*. Chacun des deux noms de métiers peut être un nom de métier monolexical (tels que *coffreur bancheur, linguiste informaticien, maître assistant, ébéniste menuisier, ingénieur numéricien, journaliste reporter, médecin légiste, enseignant chercheur, inspecteur comptable, guide accompagnateur*, etc.) ou un nom de métier polylexical (tels que *infirmier sapeur-pompier, médecin pharmacien de sapeur-pompier, capitaine de sapeur-pompier, chef d'équipe paysagiste, magasinier en chef des bibliothèques,* etc.).

De plus, il y a aussi beaucoup de noms de métiers composés qui sont formés à partir d'un autre nom de métier composé. Par exemple, pour *directeur financier du cabinet*, il est formé à partir d'un autre nom de métier composé *directeur financier.* La structure interne de *directeur financier du cabinet* peut être représentée par le patron morphosyntaxique (NMP1+NMP2)+de(DET)+N2/GN2 qui contient un autre patron morphosyntaxique du nom de métier composé NMP1+NMP2 (NMP réfère à un nom de métier monolexical). Le patron morphosyntaxique dans lequel se trouve un autre patron morphosyntaxique du nom composé est appelé le **patron morphosyntaxique complexe**. Les noms composés correspondant aux patrons morphosyntaxiques complexes sont appelés les **noms composés complexes**. En revanche, le patron morphosyntaxique qui ne comprend pas un autre patron morphosyntaxique du nom composé est appelé le **patron morphosyntaxique simple** et les noms composés correspondants sont appelés les **noms composés de base**. Il existe également beaucoup d'autres types de noms de métiers composés complexes, tels que (NMP1+NMP2)+prep(DET)+N3/GN3 (par ex., *concepteur réalisateur de loisirs durables, conseiller vendeur en agence de voyages,* etc.*),* (NMP1+NMP2)+NMP3 (par ex., *maître-nageur sauveteur, médecin-pharmacien biologiste,* etc.), (Préfixe+NMP1)+prep(Det)+N2 (par ex., *technico-commercial en agrofournitures, sous-officier de gendarmerie,* etc.), …… (cf. Annexe 1, section 5)

La formation des noms de métiers composés (NMPC) peut être en fait représentée par une série d'opérateurs et un ensemble d'éléments constitutifs. Les éléments constitutifs sont les constituants élémentaires pour former un nom de métier composé et les opérateurs indiquent les façons de former les noms de métiers avec ces éléments constitutifs. Les

éléments constitutifs présentés ici le sont au niveau de la formation de noms de métiers composés. Ce sont le préfixe, le nom de métier monolexical (NMP), le nom (N) (simple, dérivé ou composé) et le groupe nominal (GN). De plus, un nom de métier composé peut également être utilisé comme un constituant pour former un autre nom de métier composé. Les opérateurs établis sont : ←N/GN, N/GNEmprunt→, ←NMP, ←NMPC, Préfixe→, ←A, ←prep(Det)+N/GN(prep={de, à, sur, en, dans}), ←prep(Det)+NMPC, ←Vpp+en+N/GN, ←et/ou+A, ←et+NMP, ←et/ou+NMPC, ←et/ou+prep(Det)+N/GN(prep={de, à, sur, en, dans}) ainsi que ←et/ou+N/GN. La flèche désigne la direction d'opération : "←" signifie l'ajout à la fin et "→" signifie l'ajout au début. Dans la Figure 34, on donne quelques exemples de la représentation des structures internes des noms de métiers composés de base. Les noms de métiers composés complexes sont construits à partir des noms de métiers composés de base. Une autre ou plusieurs autres opérations sont surajoutées aux noms de métiers composés de base pour former les noms de métiers composés complexes. La plupart des noms de métiers complexes sont formés à partir d'un nom de métier composé en surajoutant uniquement une opération. Dans la Figure 35, on présente une analyse de la structure de certains noms de métiers composés complexes.

Cependant, un élément constitutif d'un groupe nominal au sein d'un nom de métier composé est distingué d'un élément constitutif ajouté à travers une opération indiquée par les opérateurs qu'on a relevés (à savoir un élément constitutif au niveau de la formation de noms de métiers composés). Les éléments constitutifs ajoutés par une opération peuvent être un préfixe, un nom de métier monolexical (NMP), un nom (N) (simple, dérivé ou composé), ou un groupe nominal (GN), mais ils sont ajoutés comme une expansion au nom de métier initial. Néanmoins, un élément constitutif d'un groupe nominal au sein d'un nom de métier composé n'assure un rôle de modifieur qu'au substantif-tête du groupe nominal. Par exemple, dans le nom de métier *directeur du bâtiment en énergies renouvelables*, *du bâtiment en énergies renouvelables* est l'expansion ajoutée par l'opération ←prep(Det)+N/GN(prep={de, à, sur, en, dans}), alors que *renouvelables* est un constituant du groupe nominal *énergies renouvelables.* Un autre exemple, dans le nom de métier composé *responsable d'atelier de production*, *de production* est un élément constitutif du groupe nominal *atelier de production* et *atelier de*

production est l'élément constitutif introduit par l'opération ←prep(Det)+N/GN(prep={de, à, sur, en, dans}).

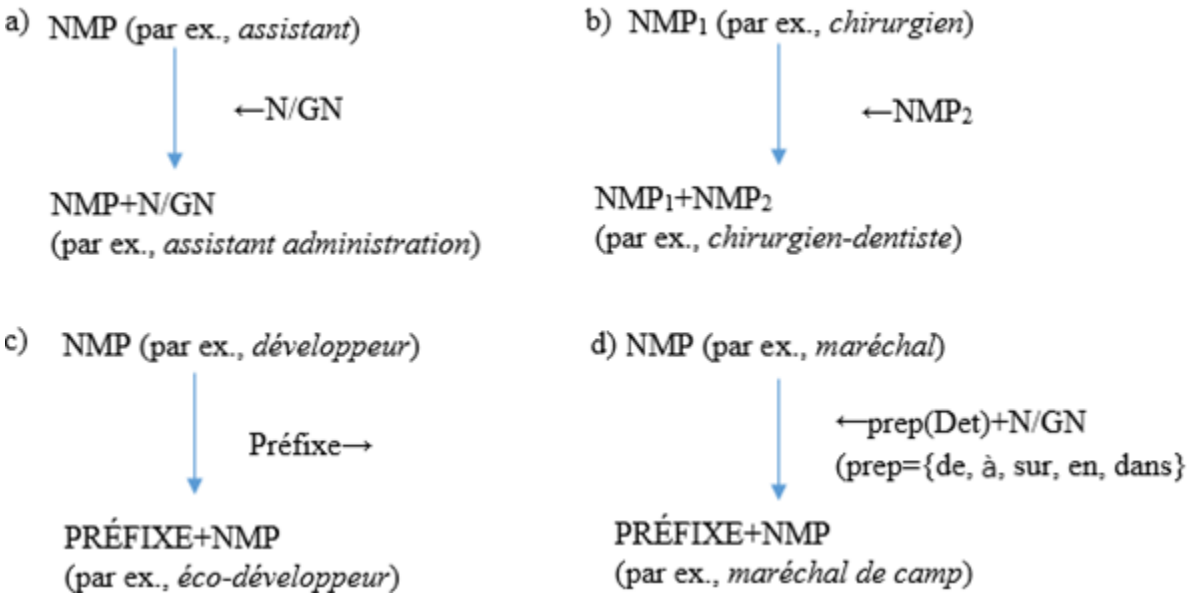

Figure 34 Exemple d'opérations pour former les noms de métiers composés de base

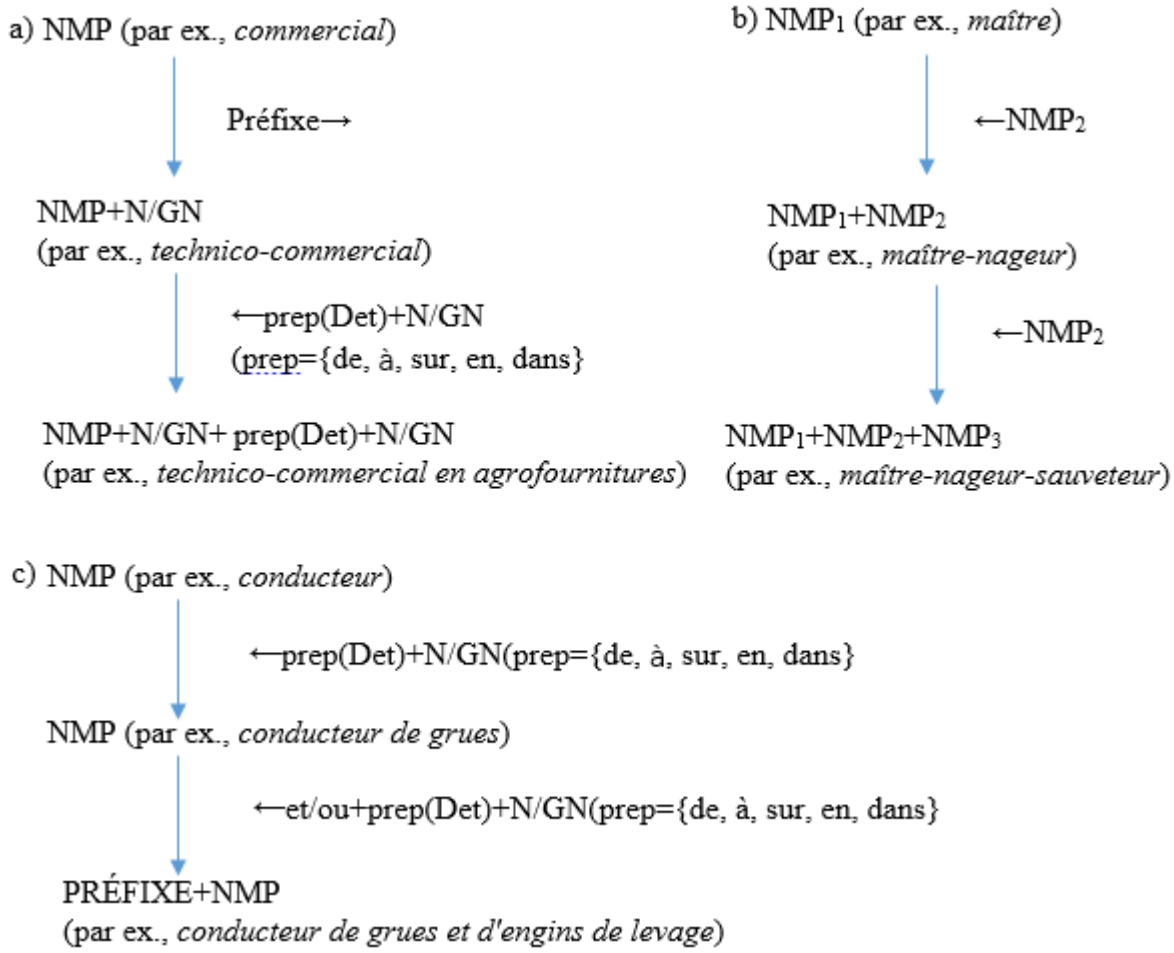

Figure 35 Exemple d'opérations pour former les noms de métiers composés complexes

Dans le schéma de la formation des noms composés complexes (cf. la Figure 35), on voit bien qu'il existe au moins deux opérateurs pour la formation de chaque nom de métier composé complexe. Néanmoins, tous les deux opérateurs ne sont pas combinables pour former un nom de métier composé. Par exemple, les opérateurs comme ←et/ou+A, ←et/ou+prep(Det)+N/GN(prep={de, à, sur, en, dans}) et ←et/ou+N/GN qui exigent une structure symétrique sont souvent combinés avec les opérateurs qui leur permettent de former une structure symétrique, tels que ←A avec ←et/ou+A, ←N/GN avec ←et/ou+N/GN et ←prep(Det)+N/GN(prep={de, à, sur, en, dans}) avec ←et/ou+prep(Det)+N/GN(prep={de, à, sur, en, dans}).

Or, certains opérateurs présentent une grande puissance de combinaison et ils peuvent être combinés avec presque tous les autres opérateurs, par exemple, les opérateurs ←A et ←prep(Det)+N/GN(prep={de, à, sur, en, dans}) semblent assez puissants de sorte qu'ils peuvent être combinés avec presque tous les autres opérateurs. La possibilité de combinaison entre deux opérateurs dépend aussi de l'ordre établi entre eux. Par exemple, l'opérateur ←A est combinable avec ←et/ou+A à condition que ←A soit mise en œuvre avant ←et/ou+A, alors que l'opération ←A est rarement surajoutée si l'opération ←et/ou+A a déjà été exécutée (par ex., *conseiller conjugal et familial* → ? (*conseiller conjugal et familial)+A*). Certains opérateurs ne peuvent être appliqués qu'aux noms de métiers composés, tels que ←et/ou+A (par ex., **technicien et électronique,* mais *technicien mécanique et électronique*), ←et/ou+N/GN (**consultant et réseaux,* mais *consultant communication et réseau*) et ←et/ou+prep(Det)+N/GN(prep={de,à,sur,en,dans}) (par ex., **directeur et d'information,* mais *directeur de service et d'information*).

Certains opérateurs ne peuvent pas se combiner directement mais peuvent l'être à travers un autre opérateur, par exemple, ←NMP et ←et/ou+N/GN ne sont pas combinables mais ←NMP, ←N/GN et ←et/ou+N/GN peuvent être mis en place l'un après l'autre pour former un nouveau nom de métier composé (par ex., **ingénieur technico-commercial et sols pollués,* mais *ingénieur technico-commercial sites et sols pollués*). Pour bien analyser les combinaisons entre les différents opérateurs, on établit une matrice carrée de taille 14 $A=(a_{i,j})$ $(1\leq i,j\leq 14)$ dans laquelle chaque ligne (i) et chaque colonne (j) représente respectivement un

opérateur. Un coefficient $a_{i,j}$ de la matrice représente une combinaison de deux opérateurs (un opérateur représenté par la ligne i (opérateur$_i$) et un opérateur représenté par la colonne j (opérateur$_j$)). On considère que l'opérateur$_i$ est supérieur à l'opérateur$_j$. Dans le Tableau 24, on liste tous les opérateurs en leur associant un identifiant dans la première colonne et chaque opérateur est retenu par son identifiant correspondant dans la première ligne. Si la combinaison d'un opérateur$_i$ avec un opérateur$_j$ est acceptable, on note O (qui signifie OUI). Si la combinaison se produit rarement, on note "-".

Identifiants / Opérateurs	**(1)**	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**	**(7)**	**(8)**	**(9)**	**(10)**	**(11)**	**(12)**	**(13)**
(1) ←N/GN	O	O	O	O	O	-	O	O	O	O	O	O	O
(2) ←A	O	O	O	O	O	-	O	O	O	O	O	O	O
(3) ←prep(Det)+N/GN (prep={de,à,sur,en,dans})	O	O	O	O	O	-	O	O	O	O	O	O	O
(4) ←Vpp+en+ N/GN	-	-	-	-	O	-	O	O	-	O	O	O	O
(5) ←et/ou+A	-	-	O	O	-	-	-	-	-	O	-	O	O
(6) ←et+NMP	O	O	O	O	-	-	-	-	O	O	-	O	O
(7) ←et/ou+prep(Det)+N/GN (prep={de,à,sur,en,dans})	-	-	-	-	-	-	-	-	-	O	-	O	O
(8) ←et/ou+N/GN	-	-	O	O	-	-	-	-	-	O	-	O	O
(9) ←NMP	O	O	O	O	-	O	-	-	O	O	O	O	O
(10) N/GNEmprunt→	O	O	O	O	O	O	O	O	O	O	O	O	O
(11) ←NMPC	O	O	O	O	O	O	O	O	O	O	O	O	O
(13) ←prep(Det)+NMPC	O	O	O	O	O	O	O	O	O	O	O	O	O
(14) Préfixe→	O	O	O	O	O	O	O	O	O	-	O	O	O

Tableau 24 Possibilité de combinaison entre les opérateurs

Malgré une certaine contrainte de combinaison, la plupart des opérateurs peuvent se combiner entre eux. En exécutant un ensemble d'opérateurs combinables, les nouveaux éléments constitutifs peuvent toujours être ajoutés successivement. De plus, dans la langue française, la plupart des éléments constitutifs (sauf Préfixe→) pour former un nouveau nom de métier composé sont en général ajoutés à la fin du nom de métier initial (par ex., *chef → chef de produit → chef de produit en informatique, adjoint→ adjoint technique→ adjoint technique de recherche→adjoint technique de recherche et de formation,* etc.). Cependant, les emprunts changent cette formation de noms composés par l'ajout de constituants. Une expansion (un nom ou un groupe nominal) est ajoutée au début dans les emprunts à l'anglais, tel que *designer->game designer, manager->campus manager, manager->data protection manager,*

etc. L'ajout des éléments constitutifs devient complètement à double sens avec l'entrée des emprunts.

2.3. Méthode

2.3.1. Extension des termes monolexicaux

L'extension des termes monolexicaux a pour objet d'enrichir la liste de noms de métiers à partir d'un ensemble de bases données à l'avance par une série d'opérations morphologiques. On commence par établir manuellement deux listes de mots : une liste de verbes décrivant l'activité professionnelle et une liste de noms décrivant l'activité professionnelle, l'objet, la mission ou le lieu de travail, la discipline et le domaine. Ensuite, on segmente automatiquement les noms en morphèmes à l'aide de Morfetik transformé en DELA. Morfetik transformé est complété par 1,000 morphèmes non autonomes (tels que *ethno-*, *-iatrie*, *musico-*, *info-*, etc.) recensés en se basant sur le DELAS et le TLFI (Trésor de la langue française informatisé). Ces morphèmes sont enregistrés comme entrées PFX (préfixe) ou SFX (suffixe) dans Morfetik. L'objectif de la segmentation des noms est de refaire une combinaison entre les morphèmes pour obtenir les nouvelles bases potentielles afin d'enrichir la liste de bases. Finalement, on établit une série de règles d'allomorphie permettant de produire les noms de métiers dérivés. L'écriture des règles d'allomorphie s'appuie sur le statut et la nature linguistique des règles d'allomorphie définis par Corbin (1987 : 283-314) : « une allomorphie est une variation de nature phonologique, non explicable phonologiquement, qui affecte un morphème appartenant à une catégorie lexicale majeure ou affixale lors d'une opération dérivationnelle ou dans un contexte phonologique ».

La segmentation automatique est effectuée selon l'algorithme suivant :

- pour les mots de la liste, on parcourt chacun de ces mots caractère par caractère du début jusqu'à la fin et on concatène un caractère chaque fois pour obtenir une sous-chaîne de caractères ;
- on cherche la sous-chaîne de caractères obtenue après chaque concaténation dans

Morfetik transformé en DELA pour vérifier si cette sous-chaîne de caractères existe comme une entrée autonome ;

- si cette sous-chaîne de caractères concaténée est une entrée dans Morfetik, on vérifie si la partie qui reste dans le mot existe aussi comme une entrée dans Morfetik ; si oui, on enregistre d'un part cette sous-chaîne de caractères comme morphème dans la liste A et d'autre part la partie qui reste dans la liste B, et on recommence la concaténation à partir de la partie qui reste ; si non, on continue la concaténation ;
- on répète la deuxième et la troisième étape jusqu'à la fin du mot ;
- chaque mot lui-même est aussi enregistré dans la liste B.

Dans la Figure 36, on donne un exemple pour montrer le processus de segmentation. Finalement, on obtient deux listes : la liste A comprenant les morphèmes qui se trouvent normalement au début ou au milieu d'un mot et la liste B comprenant les mots et les morphèmes qui se trouvent en général à la fin du mot.

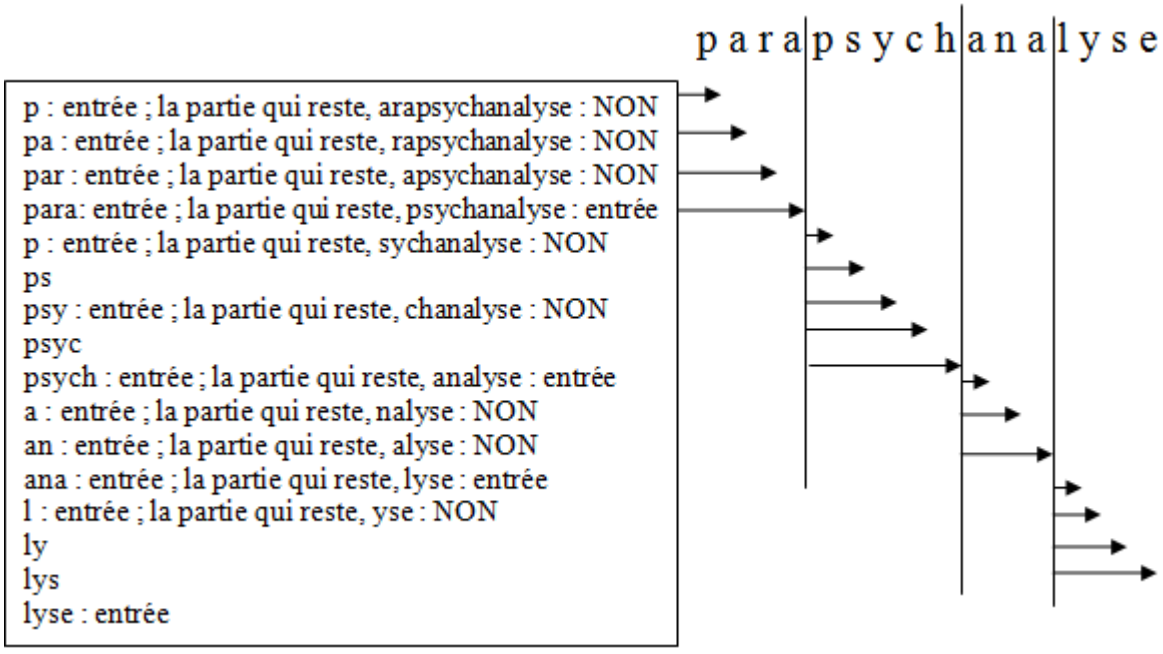

Figure 36 Processus de segmentation

Avec la liste A et B, on exécute la recomposition automatique des morphèmes pour enrichir la liste de bases. La recomposition se fait entre un morphème de la liste A et un morphème de la liste B. On prend chaque morphème dans la liste A pour combiner avec tous les morphèmes dans la liste B. Les candidats obtenus seront validés s'ils sont présents dans le

corpus. On vérifie la présence de chaque candidat dans le corpus en les étiquetant à l'aide d'Unitex. Le processus de recomposition est montré dans la Figure 37. Dans la recomposition automatique des morphèmes, l'allomorphie pour certains morphèmes est également considérée, par exemple, *psych(o)*- est sous la forme *psych*- devant la voyelle (*psychanalyse*) et est sous la forme *psycho*- (*psychogéologie*) devant la consonne.

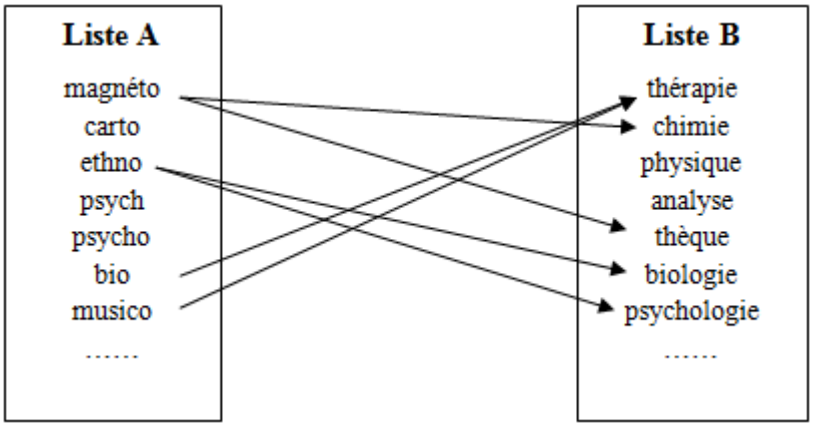

Figure 37 Processus de recomposition

Les règles d'allomorphie dans la dérivation morphologique se divisent en deux groupes : les règles d'allomorphie pour les verbes et les règles d'allomorphie pour les noms. Pour chaque verbe, on cherche d'abord dans la table de lemmes de Morfetik son code flexionnel et on trouve ensuite la façon de récupérer le radical de ce type de verbes dans la table de codes de Morfetik selon le code flexionnel trouvé. Dans la table de codes de Morfetik, la colonne « Rad » enregistre le nombre de caractères à enlever à partir de la fin pour obtenir le radical du verbe. On concatène le radical du verbe obtenu selon ce qui est indiqué dans la colonne « Rad » et la chaîne de caractères obtenue après l'enlèvement du suffixe *–ent* de la forme enregistrée dans la colonne « Ind-pr:3:S » afin d'obtenir la base du verbe à utiliser pour la dérivation. Et puis, on établit respectivement une série de règles d'allomorphie pour les verbes et les noms (cf. Annexe 1, section6) permettant de produire les formes dérivées de la variation de la base verbale (par ex., *recevoir->recev- ->récepteur*) et de prendre en compte certains attachements entre suffixes et bases (par ex, *commercer -> commerc- -> commerçant*). Finalement, on ajoute les suffixes (*-eur, -ier, -iste, -ien*…) pour former les noms de métiers. On a recensé une liste de suffixes appartenant au vocabulaire des

noms de métiers (cf. Annexe1, section 3). Pour chaque candidat obtenu par la recomposition, on ajoute tous les suffixes listés pour obtenir toutes les possibilités. Pareillement, ces formes sont validées si elles peuvent être étiquetées dans le corpus par Unitex. Le processus qui servirait à dériver les noms de métiers est montré dans la Figure 38. Dans la Figure 39, on donne deux exemples sur le processus de dérivation.

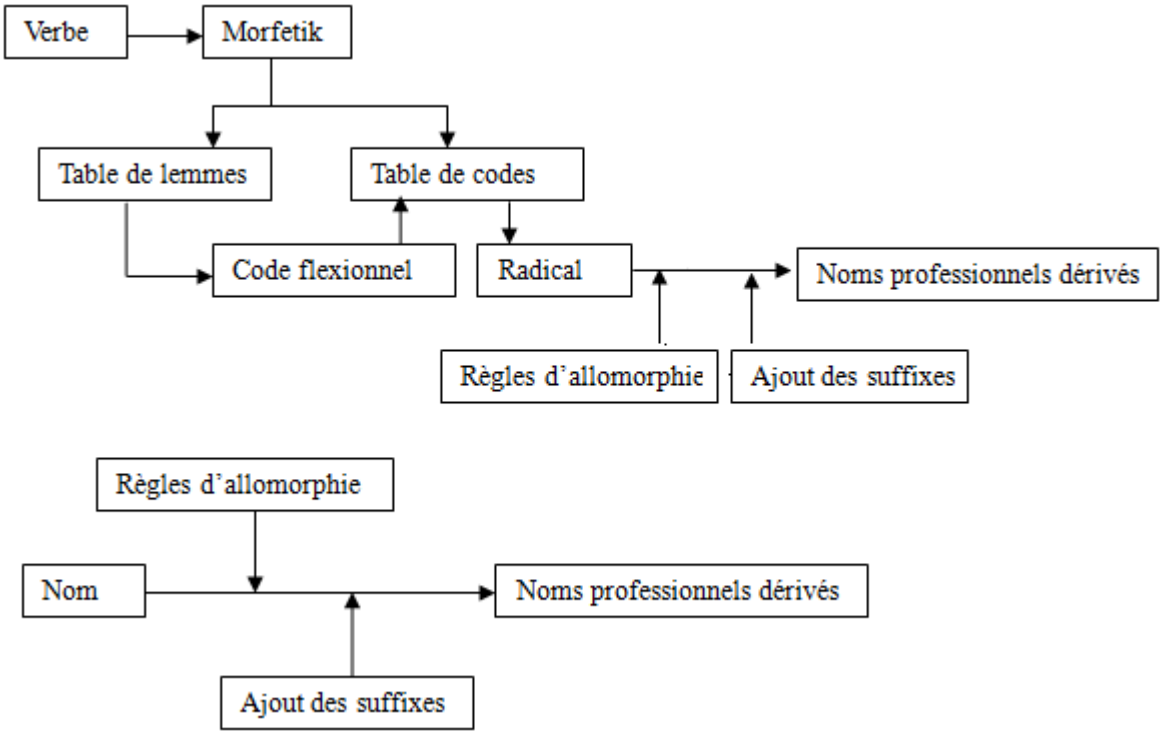

Figure 38 Processus de derivation

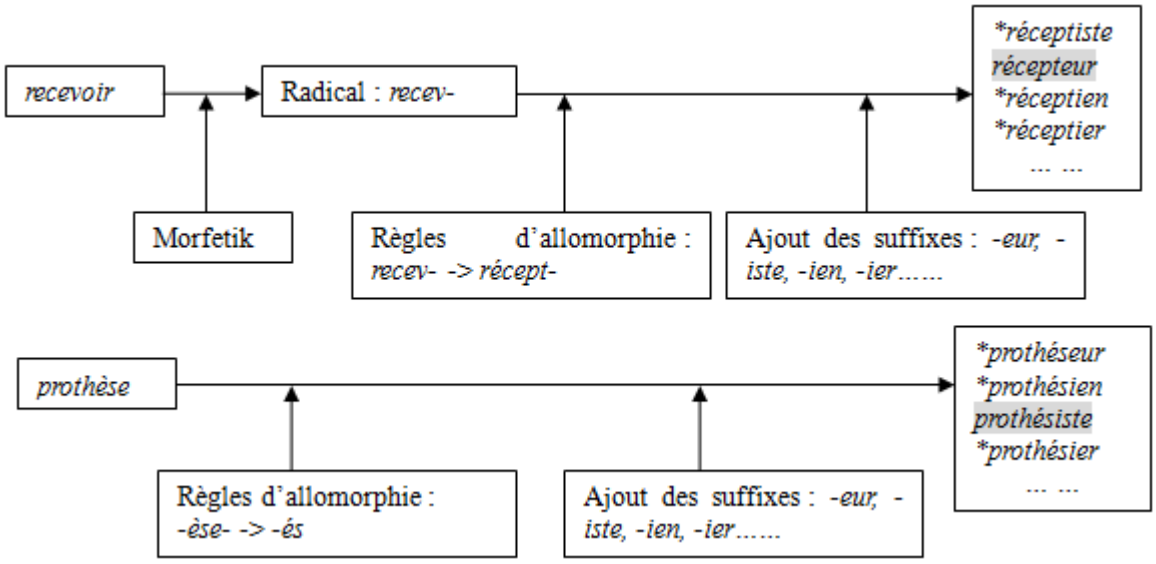

Figure 39 Exemple de dérivation

2.3.2. Construction des graphes et extraction des candidats-termes composés

La construction des graphes consiste à représenter les structures internes des noms de métiers par les grammaires locales afin d'identifier les candidats-termes. La construction des graphes est composée de deux parties : la construction des graphes pour reconnaître les noms de métiers (simples et dérivés) et la construction des graphes pour reconnaître les noms de métiers composés. Une liste de noms de métiers simples est fournie à l'avance et une liste de noms de métiers dérivés peut être obtenue par l'extension des termes monolexicaux. La construction des graphes pour les noms de métiers composés est réalisée à partir des noms de métiers monolexicaux (simples et dérivés).

Les graphes pour reconnaître les noms de métiers composés sont construits en nous fondant sur les patrons morphosyntaxiques. On ne considère que les noms de métiers composés de base de 1 à 5 grammes (i.e. 1 à 5 constituants). De la même façon, on ne considère que les groupes nominaux de 2 à 3 grammes (à savoir N+N, N+A et N+prep(Det)+N). Les noms de métiers composés complexes sont formés en général à partir des noms de métiers composés de base en mettant en œuvre une autre opération à l'aide d'opérateurs. On ne prend en compte que les noms de métiers composés complexes qui sont formés à partir des noms de métiers composés de base de 2 à 3 grammes à l'aide d'un seul opérateur et dont les éléments constitutifs introduits par l'opérateur ne dépassent pas 5 éléments constitutifs. La possibilité de combinaison entre les opérateurs est également prise en compte dans la construction des graphes sur les patrons morphosyntaxiques complexes.

On distingue deux groupes de graphes : il y a ceux qui représentent les patrons morphosyntaxiques simples et ceux qui représentent les patrons morphosyntaxiques complexes. Le premier groupe de graphes est construit à partir du graphe de noms de métiers monolexicaux (NMP.grf) et le deuxième l'est à partir du premier groupe de graphes. Chaque groupe est composé d'une série de graphes qui représentent respectivement les structures des noms de métiers de différents grammes (2, 3, 4, …) (cf. Annexe 1, section 7). Par exemple,

pour les patrons simples,

47) 2gramme_NMP.grf : ←NMP NMP.grf+NMP.grf
←N/GN NMP.grf+N
←A NMP.grf+A
… …
3gramme_NMP.grf : ←prep(Det)+N/GN(prep={de,à,sur,en,dans})
NMP.grf+prep(Det)+N
←N/GN NMP.grf+NN, NMP.grf+NA
… …

Pour les patrons complexes,

48) 2_3gramme1_NMP.grf : préfixe→ préfixe+2_3gramme_NMP.grf
←NMP 2_3gramme_NMP.grf+NMP.grf
… …
2_3gramme2_NMP.grf : ←N/GN 2_3gramme_NMP.grf+NN/NA
←NMPC 2_3gramme_NMP.grf+2gramme_NMP.grf
… …

Ensuite, on projette les graphes au corpus. Les candidats-termes monolexicaux, les candidats-termes composés reconnus et les noms de métiers monolexicaux à partir desquels les unités polylexicaux sont reconnues sont respectivement étiquetés. Le résultat enregistré est ensuite lemmatisé à l'aide de TreeTagger. On transforme la case des lettres en minuscule. Et puis, on extrait les candidats-termes à partir de ce résultat. Pour chaque candidat-terme composé, sa forme, son lemme et les termes monolexicaux à partir desquels ce terme composé est obtenu sont structurés selon le format suivant :

N_C	**N_S$_1$ N_S$_2$ …**
Lemme;Forme;	**Lemme:Forme Lemme:Forme**

Chaque ligne est composée de trois colonnes : la première colonne enregistre le lemme du candidat-terme composé (N_C), la deuxième colonne enregistre la forme du candidat-terme composé et la troisième colonne enregistre les informations sur les termes monolexicaux (N_S) à partir desquels le candidat-terme composé est obtenu. Dans la troisième colonne, le lemme et la forme de chaque candidat-terme monolexical sont liés par ":" et ces informations

(lemme et forme) des différents termes monolexicaux sont divisées par un espace. La Figure 40 est la capture d'écran d'une partie de ces informations.

```
ingénieur embarquer temps réel ;ingénieur embarqué temps réel ;ingénieur:ingénieur
ingénieur logiciel ;ingénieur logiciel ;consultant:consultants ingénieur:ingénieurs ingé
ingénieur logiciel ;ingénieur logiciel ;ingénieur:ingénieur
technicien itinérant ;technicien itinérant ;consultant:consultants ingénieur:ingénieurs
technicien itinérant ;technicien itinérant ;technicien:technicien
technicien itinérant ;technicien itinérant ;consultant:consultants technicien:technicien
chef de atelier ;chef d' atelier ;expert:expert chef:chef
gestionnaire administration du vente ;gestionnaire administration des ventes ;consultant
gestionnaire administration du vente ;gestionnaire administration des ventes ;gestionnai
gestionnaire administration du vente ;gestionnaire administration des ventes ;gestionnai
chef de secteur ;chef de secteur ;chef:chef
```

Figure 40 Informations structures sur les candidats-termes

2.3.3. Calcul de l'information mutuelle

L'extraction automatique des noms composés à l'aide de l'information mutuelle repose en général sur la définition du nom composé qui prend souvent en compte la cooccurrence des constituants du nom composé. L'information mutuelle a pour objet de calculer la dépendance entre deux variables (deux mots dans la méthode de l'extraction des noms composés) et permet de détecter la collocation. Cette mesure est définie par la formule suivante (Aggarwal et Zhai, 2012 : 89) :

$$IM\,(x,\,y) = \log_2 \frac{P(x,y)}{P(x)P(y)} \qquad (43)$$

dans laquelle (x, y) est un couple de variables, P(x) indique la probabilité de variable x, P(y) représente la probabilité de la variable y et P(x, y) signifie la probabilité de cooccurrence de x et y. Dans notre méthode, l'information mutuelle calcule le degré de dépendance des mots x et y composant la collocation.

Les candidats-termes composés sont extraits à l'aide des patrons morphosyntaxiques, mais cela ne garantit pas que tous les candidats-termes reconnus par les patrons morphosyntaxiques soient les noms composés. Certaines combinaisons de mots incluent un constituant d'une autre unité lexicale ou d'un autre élément syntaxique de la phrase et certaines combinaisons ne composent que les groupes nominaux au lieu de collocations. On fait appel au calcul d'information mutuelle pour détecter la collocation et éliminer les

candidats-termes qui ne sont pas des noms composés.

L'information mutuelle est calculée dans la méthode selon l'algorithme suivant :

- si le candidat-terme (CT) est de 2-grammes, l'IM est calculée entre les deux constituants du CT, et le candidat-terme est enregistré dans la liste de noms de métiers si la valeur d'IM du candidat-terme dépasse le seuil défini préalablement ;
- si le candidat-terme est de n-grammes (n>2), l'IM est d'abord calculée entre la sous-séquence de premiers (n-1) tokens et le dernier token ; si le résultat de calcul dépasse au seuil décidé, le CT est enregistré dans la liste de noms de métiers ; sinon, le dernier token est enlevé pour calculer l'IM entre la sous-séquence des premiers (n-2) tokens et le (n-1)^{ème} token, et cette étape sera répété itérativement jusqu'à ce qu'on trouve une collocation (à savoir quand l'IM dépasse le seuil)

L'étape pour calculer l'IM des n-grammes (n>2) est montré dans la Figure 41.

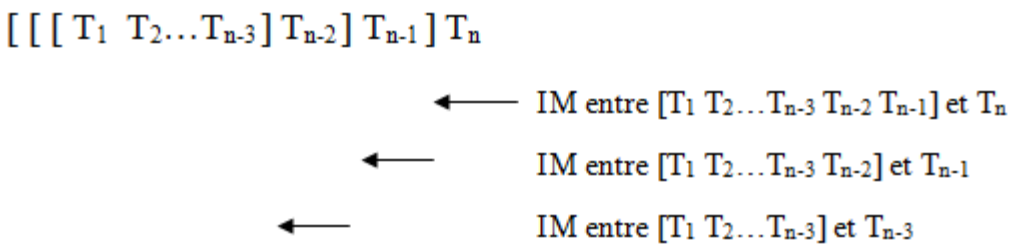

Figure 41 Méthode de calcul d'information mutuelle

La Figure 42 est la capture d'écran d'une partie du résultat final.

```
-2.77258872223978;assister commercial marchandise sinistré;assistant commercial marchandises
-2.83321334405622;chef de projet sur le dématérialisation;chef de projet sur la dématérialisa
-2.89037175789617;correspondre au cotisation patronal;correspondant aux cotisations patronale
-3.04452243772342;boulanger pâtissier et restaurateur;boulangers pâtissiers et restaurateurs;
-3.04452243772342;chef de le station de lavage;chef de la station de lavage;chef:chef
-3.04452243772342;technicien logistique de entreposage;technicien logistique d' entreposage;t
-3.09104245335832;ingénieur en aménagement paysager;ingénieurs en aménagement paysager;ingéni
-3.09104245335832;auteur de infraction sexuel;auteurs d' infractions sexuelles;auteur:auteurs
-3.2188758248682;ingénieur composant electrotechniques;ingénieur composants electrotechniques
-3.29583686600433;ingénieur commercial solution evénementielles;ingénieur commercial solution
-3.29583686600433;technicien de maintenance de parc eoliens;technicien de maintenance de parc
-3.3322045101752;vendeur livreur espacer;vendeur livreur espace;vendeur:vendeur
-3.36729582998647;chef être lauréat du espoir;chefs est lauréat des espoirs;chef:chefs
-3.40119738166216;chef de groupe de le comptabilité immobilisation;chef de groupe de la compt
-3.40119738166216;assister administratif _ assister trilingue;assistant administratif _ assis
-3.49650756146648;chef cuisinier ou mécano;chef cuisinier ou mécano;chef:chef
```

Figure 42 Résultat de calcul d'information

3. Évaluation

L'approche morphosémantique est un système de production dont les sorties sont un ensemble de termes. L'évaluation de la pertinence de ce type de sorties peut s'effectuer sur le plan de la qualité et sur le plan de la quantité. La quantité du résultat peut être évaluée en comparaison avec le standard en calculant le taux de rappel et le taux de précision. La qualité peut aussi être évaluée en comparaison avec un standard mais du point de vue de l'aspect qualitatif. Le standard est établi par une annotation manuelle en l'absence de ressources lexicales des termes composés. Dans cette section, on présente en détail respectivement les évaluations quantitative et qualitative des résultats de la méthode morphosémantique pour l'acquisition automatique des noms de métiers.

3.1. Évaluation quantitative

L'expérimentation de la méthode morphosémantique nous a permis de fournir une liste de termes simples et deux listes de bases (une liste de verbes et une liste de noms) permettant d'obtenir les termes dérivés par la dérivation morphologique. Les termes simples et les listes de bases donnés à l'avance comprennent environ 150 unités. Pour l'évaluation quantitative, on affiche d'abord tous les candidats-termes reconnus par les patrons morphosyntaxiques en leur associant la valeur de l'information mutuelle. On remarque que les erreurs commencent à apparaître fréquemment parmi les candidats-termes dont la valeur de l'information mutuelle est inférieure à -90. On définit ainsi un ensemble équivalent à un seuil de valeurs : -80, -85, -90, 95, -100 ; on applique ces valeurs pour obtenir cinq différents résultats. Et puis, on étiquette les termes dans le corpus avec cinq résultats différents. Ensuite, on sélectionne environ cinq cents textes de chaque résultat d'étiquetage au hasard et on rédige une annotation manuelle pour étiqueter les vrais positifs, les faux positifs et les faux négatifs. On évalue chaque résultat par l'annotation manuelle et on calcule le taux de précision, le taux de rappel et le taux de F-mesure. Les résultats d'évaluation sont listés dans le Tableau 25.

Valeurs de seuil	Précision	Rappel	F-mesure
-80	94.29%	76.91%	84.72%
-85	94.29%	77.76%	85.23%
-90	93.89%	78.10%	85.30%
-95	93.47%	78.35%	85.24%
-100	92.65%	78.61%	85.05%

Tableau 25 Résultats d'évaluation de la méthode morphosémantique

La Figure 43 montre la comparaison d'évaluation de ces cinq résultats. On voit bien que le taux de F-mesure est le plus élevé quand le seuil est proche de 90. Finalement, on retient le seuil 90 pour obtenir le résultat final.

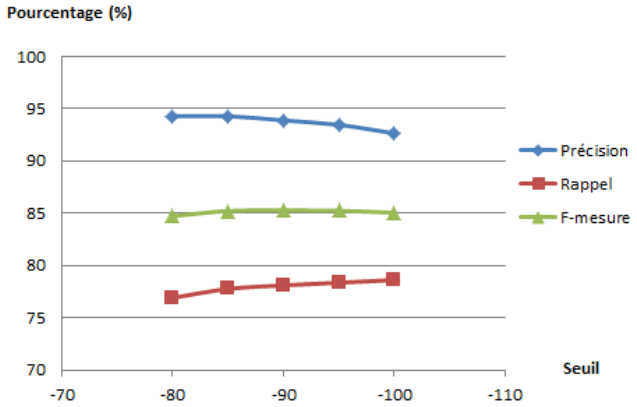

Figure 43 Comparaison des résultats d'évaluation obtenus avec les seuils différents

De plus, on calcule également le rappel du résultat obtenu par la liste des termes fournis préalablement. On étiquette le corpus avec les termes simples fournis à l'avance et les termes dérivés correspondant aux bases fournies de la même manière. Le rappel du résultat est de 31.89%. Par les processus morphologiques (recomposition des morphèmes et l'identification des termes composés), le rappel du résultat est augmenté de 31.89% à 78.35%.

3.2. Évaluation qualitative

On analyse les mille premiers termes de notre résultat : il s'y trouve 64 erreurs au total. Parmi ces erreurs, on recense 6 principaux types d'erreurs : type A, titre d'annonce d'emploi qui est une combinaison du nom de métier et du profil (par ex., *assistant administratif anglais courant*, *adjoint de direction futur directeur*, etc.) ; type B, construction verbale ou d'autres constructions syntaxiques apparentées (par ex., *correspondant aux attentes des recruteurs, correspondant au type de véhicule*, etc.) ; type C, termes composés

avec un ou plusieurs constituants superflus (par ex., *ingénieur développeur commercial métier, chef de secteur acquise,* etc.) ; type D, un groupe nominal au lieu d'un terme composé (par ex., *technicien de laboratoire bactériologie, chef de petite ou moyenne entreprise*, etc.) ; type E, changement de classe sémantique (par ex., *marin* est un nom de métier, mais *sous-marin* est un nom d'artefact) ; type F, ambiguïté morphosyntaxique (par ex., *ingénieur reste, experts dit vouloir*, etc.). On compte le nombre des différents types d'erreurs dans l'ensemble des erreurs trouvées et on calcule la proportion de chaque type d'erreur. Dans le Tableau 26, toutes ces informations sont énumérées. On voit bien que les groupes nominaux qui sont mal reconnus comme termes composés et les termes reconnus avec un ou plusieurs constituants superflus occupent la proportion la plus importante parmi tous les types d'erreurs. Le calcul d'information ne permet d'éliminer que certaines séquences qui ne sont pas des collocations. Cette méthode statistique est sensible à la quantité des informations de cooccurrences dans le corpus.

Types d'erreurs	Nombre	Proportion	Exemples
A	2	3.12%	*assistant administratif anglais courant, adjoint de direction futur directeur*
B	8	12.5%	*correspondant aux attentes des recruteurs, correspondant au type de véhicule, correspondant à la création, correspondant aux nouvelles exigences, correspondant aux postes susceptibles, correspondant à la tournée du chauffeur, correspondant aux évolutions ou corrections, correspondant aux cotisations*
C	27	42.20%	*ingénieur développeur commercial métier, chef de secteur acquise, gestionnaires des mises, vendeurs techniciens et passionnés, chefs de chantier à bac, ingénieur généraliste ou diplômé, etc.*
D	19	29.69%	*chef de petite ou moyenne entreprise, technicien de laboratoire bactériologie, dirigeant du groupe, ingénieurs jeunes diplômés, ingénieurs français, etc.*
E	2	3.12%	*sous-marin et bâtiment de surface, battant en polyuréthane*
F	5	7.81%	*ingénieur reste, expert dit vouloir, expert métier reconnu,* *conseiller clientèle repose sur la large, ingénieur suivi*
Total	64	100%	

Tableau 26 Types d'erreurs dans le résultat de la méthode morphosémantique

Ensuite, on étiquette le corpus avec le résultat obtenu (la liste de termes) par la méthode morphosémantique et on analyse 60 textes. On trouve au total 27 silences dont 5 sont présents en raison de leurs valeurs d'information mutuelle inférieures au seuil (notés comme type de silence A), 2 sont causés par le manque de patrons morphosyntaxiques (type de

silence B), 3 sont causés par les erreurs d'étiquetage morphosyntaxique (type de silence C) et 17 sont provoqués à cause de la liste incomplète de termes donnée à l'avance (type de silence D). En ce qui concerne Le type de silence B, certaines structures très longues ou séparées plusieurs fois par les séparateurs comme la structure [N+de+N+, +de+N+, +A+, +N+ou+N] (par ex., *assistant de planification, de gestion, commercial, affrètement ou logistique*) ou la structure [N+de+N+, +N/GN] (par ex., *chef d'atelier, de la planification des interventions de dépannage*) ne sont pas considérées dans la construction des patrons morphosyntaxiques pour éviter d'amener trop de bruits. Pour le type de silence C, certains mots avec les symboles comme *C++* ne sont pas entrés dans le dictionnaire et certains noms propres comme *GMS* dans *chef de secteur GMS* ne sont pas pris en compte par les patrons morphosyntaxiques. Pareillement, on compte le nombre de ces différents types de silences et on calcule la proportion de chaque type dans toutes les erreurs. Dans le Tableau 27, toutes les informations sur les types de silences recensés sont listées.

Types de silences	Nombre	Proportion	Exemples
A	4	14.81%	*animateur départemental* (-625.604817987331), *conseiller service client* (-122.277111070618), *assistant de gestion* (-602.809448072686), *technicien itinéran*t (-164.829915824731)
B	2	7.41%	*chef d'atelier*, *de la planification des interventions de dépannage* ; *assistant de planification*, *de gestion, commercial*, *affrètement ou logistique* ;
C	3	11.11%	*chef de secteur GMS*, *ingénieur logiciel C _ C++*, *ingénieur spécialisé C C++*
D	18	66.67%	*attaché au chargé d'affaires des filiales de Charente et de Dordogne, attaché au responsable des ventes, responsable de magasin, spécialiste, commerciale, métreur, opérateur,* etc.
Total	27	100%	

Tableau 27 Types de silences dans le résultat de la méthode morphosémantique

Les silences provoqués par la liste incomplète de termes donnée à l'avance occupent la proportion la plus importante parmi les différents types de silences recensés. Cela nous renvoie à la méthode distributionnelle qui permet de récupérer plus de termes à partir d'une liste de termes fournie à l'avance. Il est donc possible que la méthode qui combine la méthode distributionnelle et la méthode morphosémantique permette de donner un résultat amélioré, ce que nous allons voir dans le prochain chapitre.

Chapitre 3 Méthode combinatoire

1. Corpus adopté

La méthode combinatoire est une méthode qui associe l'approche distributionnelle et l'approche morphosémantique. Le corpus adopté pour la méthode combinatoire doit permettre de réaliser à la fois le fonctionnement des deux méthodes. La méthode distributionnelle (tant supervisée que semi-supervisée) se fonde sur l'exploitation des structures prédicat-argument pour réaliser l'acquisition automatique des termes, alors que la méthode morphosémantique consiste à exploiter les structures internes des unités lexicales pour identifier automatiquement les termes. Pour la méthode combinatoire, un corpus enrichi de structures prédicat-argument et d'informations morphosyntaxiques est nécessaire.

La méthode combinatoire s'applique à deux vocabulaires spécifiques : le vocabulaire des noms d'artefacts et le vocabulaire des noms de métiers. Pour l'étude des noms d'artefacts, on adopte le même corpus que celui de la méthode distributionnelle qui atteint 22,858 Ko. Il concerne plusieurs thèmes de la vie quotidienne (le transport, la mode, le ménage, le bricolage, etc.) et comprend sept genres de textes issus du web. Les textes de différents thèmes ont environ le même volume dans le corpus total. Les textes occupent également presque la même proportion dans le corpus total. Le corpus pour l'étude du vocabulaire des noms de métiers est le même que celui de la méthode morphosémantique. Il atteint 23,571 Ko et comprend trois genres de textes issus du web. Les textes sont tous adaptés à la thématique métier ou emploi. Les textes de différents genres dans le corpus de noms de métiers occupent à peu près la même proportion.

2. Méthode

La méthode combinatoire consiste à combiner la méthode distributionnelle et la méthode

morphosémantique pour profiter des avantages de chaque méthode et éviter les inconvénients de chacune afin d'optimiser la pertinence du résultat de l'acquisition automatique du vocabulaire. La méthode combinatoire est appliquée à deux vocabulaires : les noms d'artefacts et les noms de métiers. Pour les noms d'artefacts, la méthode morphosémantique est intégrée comme un module complémentaire dans la méthode distributionnelle. Pour les noms de métiers, cette méthode assure un rôle de filtrage sémantique pour la méthode morphosémantique. La combinaison des deux méthodes pour les noms de métiers doit être dotée au préalable d'un module de filtrage. Dans cette section, on présente respectivement le moyen de combinaison des deux méthodes pour les noms d'artefacts et les noms de métiers.

2.1. Méthode combinatoire pour les noms d'artefacts

Pour les noms d'artefacts, la combinaison de la méthode distributionnelle et de la méthode morphosémantique consiste à intégrer la méthode morphosémantique pour récupérer non seulement les unités monolexicales mais aussi les unités polylexicales. L'intégration de la méthode morphosémantique s'effectue à la fois dans la méthode distributionnelle supervisée et la méthode distributionnelle semi-supervisée. Dans ce qui suit, on présente d'abord les analyses morphosémantiques des noms d'artefacts. Ensuite, on présente le moyen d'intégration de la méthode morphosémantique dans la méthode distributionnelle. Finalement, on présente le résultat obtenu par la méthode combinatoire.

2.1.1. Analyses morphosémantiques des noms d'artefacts

L'analyse morphosémantique des noms d'artefacts consiste à analyser les structures internes des noms d'artefacts en tenant compte de la relation sémantique entre les éléments constitutifs. Pareillement, on ne considère que les composés graphiques (ceux qui sont articulés par des séparateurs, tels que le trait d'union, l'apostrophe ou l'espace) comme unités polylexicales. Les composés sans séparateurs (*magnétoscope, aérographe, amblyoscope*), les dérivés (*compteur, lanterneau, cafetière*) et les mots simples (*accessoire, aiguille, affiche*) sont tous considérés comme unités monolexicales. On analyse d'abord la morphologie des noms d'artefacts monolexicaux, y compris la morphologie dérivationnelle des noms

d'artefacts dérivés et la relation sémantique entre les bases et leurs dérivés. Ensuite, on analyse la morphologie des noms d'artefacts polylexicaux ainsi que la relation sémantique entre les termes polylexicaux et leurs constituants. Bien qu'un nom composé soit sémantiquement figé (i.e. la combinaison sémantique des constituants de ce nom composé n'est pas égale à son sens), la sémantique des éléments constitutifs n'ont pas aucun rapport avec le sens de ce nom composé. Par exemple, *porte-manteau* est un nom d'artefact dont les deux constituants (le verbe *porter* et le complément d'objet *manteau*) décrivent la fonction de l'artefact.

Par rapport au vocabulaire des noms de métiers, celui des noms d'artefacts possède moins de dérivés. La forme d'une unité monolexicale de nom d'artefact est plus abstraite. Le suffixe le plus souvent utilisé pour former un nom d'artefact dérivé est *–eur*. Par exemple, *accélérateur, absorbeur,* et *accumulateur* tous des noms terminés en *–eur* sont souvent dérivés de verbes décrivant les fonctions des artefacts par l'ajout du suffixe *–eur*. Certains noms d'artefacts sont dérivés de verbes en ajoutant *–euse*, tels qu'*agrafeuse, assembleuse, assortisseuse*, etc. Il existe également d'autres suffixes dans le vocabulaire des noms d'artefacts, tels que *–ette (camion->camionnette, corne->cornette, chemise->chemisette), -ière (café->cafetière, cuisine->cuisinière, genou->genouillère)* et *-eau (pointe->pointeau, arc->arceau, traîner->traîneau, chant->chanteau).* Néanmoins, *-ette* est plutôt un suffixe pour exprimer la petitesse. Il peut aussi être utilisé pour former les unités lexicales des autres classes sémantiques, par exemple, *fille -> fillette*. Il n'est pas un suffixe permettant de définir la classe sémantique d'une unité lexicale.

Pour les noms d'artefacts dérivés se terminant en *–eur* ou en *-euse*, la relation sémantique entre les bases (Y) et les dérivés (X) peut être expliquée par « X est l'artefact qui a pour fonction de+Y(verbe) », par exemple, *accélérateur* est l'artefact qui a pour fonction de faire *accélérer*. Pour ceux qui se terminent en *–ière*, la relation sémantique entre les bases et les dérivés peut être paraphrasée par « X est l'artefact qui V+Y(objet) », par exemple, *genouillère* est l'artefact qui protège les *genoux*. Pour ceux qui se terminent en *–eau*, les relations sémantiques entre les bases et les dérivés sont diverses ; par exemple, la relation sémantique

entre *traîner* et *traîneau* peut être paraphrasée par « X est l'artefact qui a pour fonction de+Y(verbe) » ; la relation sémantique entre *arc* et *arceau* peut être paraphrasée par « X est l'artefact qui est sous forme de Y » ; la relation sémantique entre *pointe* et *pointeau* peut être paraphrasée par « Y est une partie de X». Dans le Tableau 28, on liste tous les principaux suffixes que les noms d'artefacts dérivés peuvent prendre et les relations sémantiques entre les bases et leurs dérivés. Différentes de celles des noms de métiers, la plupart des bases des noms d'artefacts dérivés sont simples et ne peuvent plus être décomposées en morphèmes. Cependant, il existe de nombreux noms d'artefacts composés sans séparateurs, tels que *magnétoscope (magnéto-, scope), aérographe (aéro-, graphe), amblyoscope (amblyo-, scope), anémographe (anémo-, graphe), audiophone (audio-, phone), autocuiseur (auto-, cuiseur),* etc.

Suffixes	**Types sémantiques**	**Exemples**
-eur	X est l'artefact qui a pour fonction de+Y(verbe)	*basculeur* (*basculer*), *batteuse/batteur* (*battre*), *verseur/verseuse* (*verser*), *mesureur* (*mesurer*), *brûleur* (*brûler*), *compteur* (*compter*), etc.
-ière	X est l'artefact qui V+Y(objet)	*cafetière* (*café*), *cuisinière* (*cuisine*), *genouillère* (*genou*), *cartouchière* (*cartouche*), *chaudière* (*chaude*), *chiffonnière* (*chiffon*), etc.
-eau	X est l'artefact qui a pour fonction de+Y(verbe) ;	*traîneau* (*traîne*), etc.
	X est l'artefact qui est sous forme de Y ;	*arceau* (*arc*), *créneau* (*cran*), *lanterneau* (*lanterne*), etc.
	Y est une partie de X ;	*pointe (pointeau), etc.*
	X est un type de Y ;	*cordeau* (*corde*), *cuveau* (*cuve*), *écriteau* (*écrit*), etc.
-ette	X est le petit de Y	*camionnette* (*camion*), *cornette* (*corne*), *chemisette* (*chemise*), etc.
-elle	X est le petit de Y	*poutrelle* (*poutre*)

Tableau 28 Typologie de relations sémantiques internes des noms d'artefacts

La désignation d'un nom d'artefact est souvent définiee en combinant un prédicat qui décrit une action et un nom qui réfère à l'objet de cette action. Cette structure verbe+C.O.D. (complément d'objet direct) décrit la fonction du nom d'artefact. Par exemple, dans le nom d'artefact composé *allume-cigare, allume* est la conjugaison du verbe *allumer* et *cigare* est le nom qui indique l'objet de l'action allumer ; dans *brosse à dents*, *dents* est l'objet de l'action *brosser* et *brosser les dents* indique la fonction de l'artefact ; pareillement, *abat-jour* est aussi un nom d'artefact qui correspond à la structure verbe+C.O.D.

L'expansion qui décrit la fonction de l'objet indiqué par le substantif-tête peut être un

nom ou un groupe nominal introduit par la préposition *de* (par ex., dans *ailette de refroidissement, refroidissement* décrit la fonction de l'ailette) et un verbe infinitif ou une structure verbe-objet introduit par une préposition *à* (par ex., dans *boîte à déjeuner*, *déjeuner* décrit la fonction ou l'usage de la boîte ; dans *appareil à battre les collets, battre les collets* est la structure verbe-objet et il désigne la fonction de l'objet *appareil*).

De plus, un préfixe est souvent ajouté à une unité lexicale pour décrire la fonction ou la propriété de l'artefact concerné. Par exemple, dans *anti-vol*, *anti-* signifie « contre » et l'objectif de l'artefact *anti-vol* est de lutter contre le vol ; dans *pare-soleil*, *pare-* a aussi le sens « contre, éviter ou empêcher » et la combinaison du préfixe *pare-* avec l'unité lexicale *soleil* permet de décrire la fonction de l'artefact pare-soleil ; dans *auto-cuiseur*, auto- a pour objet de décrire la propriété d'automatisation de l'artefact.

L'expansion au sein d'un nom d'artefact composé peut décrire non seulement la fonction de l'artefact concerné mais aussi d'autres propriétés associées à l'artefact, telles que la taille, la couleur, la matière, certaines parties de l'artefact, ou même la source d'énergie d'où l'artefact doit s'alimenter pour fonctionner, etc. Toutes les propriétés associées à un artefact peuvent être employées pour former un artefact composé. Les expansions décrivant les propriétés de l'artefact peuvent être un nom qui modifie le substantif-tête du nom d'artefact, un adjectif (ou plusieurs adjectifs) ou une structure introduite par une préposition comme de+N, à+N, avec/sans+N ou en+N. Par exemple, dans *boîte aux lettres, lettres* est introduit par la préposition *à* et spécifie l'usage de l'artefact ; dans *chapeau sans bord, sans bord* désigne la forme du chapeau ; dans *accessoire pour chaussures*, *pour chaussures* précise l'objet auquel l'accessoire est réservé ; dans *pull en laine (wool pullover), en laine* désigne la matière du pull. Dans *appareil électro-acoustique, voiture éolienne, ampoule globe opale,* etc., les propriétés associées aux artefacts sont indiquées par les adjectifs.

Une expansion peut être ajoutée successivement au sein d'un nom d'artefact composé pour former d'autres noms d'artefacts. Par exemple, à partir du nom d'artefact *écrou hexagonal*, une expansion peut être ajoutée pour obtenir *écrou hexagonal auto freiné* et l'ajout d'une expansion peut être continué pour obtenir un autre nom d'artefact composé *écrou*

hexagonal auto freiné à anneau tout métal ou *écrou hexagonal auto freiné à anneau tout métal avec fente.* Certains noms d'artefacts composés sont formés en ajoutant une expansion à un autre nom d'artefact composé. Les expansions des noms d'artefacts peuvent être un adjectif, un nom, un groupe nominal, une séquence prépositive ou un autre nom d'artefact composé. Par exemple, pour le nom d'artefact composé *abat-jour en laine*, *abat-jour* est un nom d'artefact qui correspond à la structure Vp(verbe au présent)N et l'expansion *en laine* est surajoutée pour former un autre nom d'artefact composé *abat-jour en laine*. Le patron morphosyntaxique qui représente la structure interne de *abat-jour en laine* est (VpN1)enN2. Un autre exemple, *commande de température par cadran* est un nom d'artefact formé en ajoutant l'expansion *par cadran* à partir de *commande de température.*

La structure de *commande de température par cadran* peut être représentée par le patron morphosyntaxique (N1deN2)parN3. Tant *abat-jour en laine* que *commande de température par cadran*, leurs patrons morphosyntaxiques comprennent un autre patron morphosyntaxique de nom d'artefact composé. Les patrons morphosyntaxiques de ces noms d'artefacts composés sont appelés **les patrons morphosyntaxiques complexes**. Les patrons morphosyntaxiques complexes contiennent au moins un autre patron morphosyntaxique (simple ou complexe) de nom composé. **Le patron morphosyntaxique simple** ne comprend pas un autre patron morphosyntaxique de nom composé et il représente les structures de noms composés les plus petites et basiques. (cf., Annexe 1, sections 8 & 9.)

On recense un ensemble d'opérateurs qui indiquent les opérations à effectuer pour former les noms d'artefacts composés : ←N/GN/NAF (par ex., [NAF← N/GN/NAF]->[NAF+NAF] : *diffuseur vidéo*), Prefixe→ (par ex., [Prefixe→NAF]->[Préfixe+NAF] : *auto-cuiseur*), V→ (par ex., [V→N]->[V+N] : *garde-robe*), ←prep+N/GN/NAF(prep={de, à, dans, pour, par, sans, avec, sur}) (par ex.,[N←prep+N/GN/NAF]->[N+prep+N/GN/NAF] : *aimant de retenue, crème pour les mains, commande par bouton poussoir, boîte aux lettres, appareil avec halogène, abat-jour sur pc*), ←A (par ex., [N←A]->[N+à+A] : *aiguille hypodermique*), ←à+V (par ex., [N←à+V]->[N+à+V] : *boîte à déjeuner*), ←à+V+N/GN/NAF (par ex., [N←à+V+N/G]->[N+à+V+N/GN] : *appareil à battre les collets*) et ←Vpp+de+N/GN/NAF (par ex., [N←Vpp+de+N/GN/NAF])->[N+Vpp+de+N] : *bac garni de sac*). Ces opérations

désignent respectivement les constituants à ajouter. La liste des opérateurs établie représente les opérations possibles pour former les noms d'artefacts composés. Les noms composés de base peuvent être considérés comme les unités lexicales formées à partir des noms monolexicaux par une opération indiquée. Par exemple,

49) *ailette de refroidissement* : N1 -> [[N1] ←de+N2]

50) *porte-bébé* : N -> [Vp→ [N]]

51) *appareil à grand capteur* : N1->[[N1] ←GN2, GN2=A+N2]

Les noms d'artefacts composés complexes peuvent être considérés comme les noms formés à partir des noms d'artefacts composés de base en surajoutant une ou plusieurs opérations. Par exemple,

52) *brosse nettoyante visage* (N_1+A+N_2): [[[N_1] ←A] ←N_2]

53) *arbre porte galet* (N_1+Vp+N_2): [[N_1] ←NAF, NAF=Vp+N_2]

54) *bonde pare bruit* (N_1+Préfixe+N_2) : [[N_1] ←NAF, NAF=Préfixe+N_2]

2.1.2. Mise en place du module complémentaire

L'intégration de la méthode morphosémantique dans la méthode distributionnelle s'effectue en deux étapes : la construction des graphes pour l'identification des noms d'artefacts composés et l'extension des termes en exploitant les structures prédicat-argument extraites et les termes reconnus. Les graphes construits sont intégrés dans les graphes d'étiquetage de structures prédicat-argument. Les structures prédicat-argument extraites sont exploitées pour étiqueter les noms d'artefacts du type V+N. Les noms d'artefacts reconnus sont utilisés pour identifier d'autres noms d'artefacts composés, puisque de nombreux noms d'artefacts composés sont formés à partir d'un nom d'artefact simple et des noms d'artefacts composés complexes sont formés à partir des noms d'artefacts composés de base. L'extension des termes est effectuée en s'appuyant sur les analyses morphosémantiques et elle permet d'étiqueter plusieurs termes qui ne se trouvent pas dans les structures prédicat-argument. Par rapport à la méthode distributionnelle, la méthode combinatoire met en place d'un module

complémentaire.

2.1.2.1. Construction des graphes

La construction des graphes pour l'identification des noms d'artefacts composés se divise en deux étapes : la construction des graphes pour l'identification des noms d'artefacts composés de base en s'appuyant sur les patrons morphosyntaxiques simples (Termes_composés_de_base.grf) et la construction des graphes pour l'identification des noms d'artefacts composés complexes à partir des graphes de noms d'artefacts composés de base (Termes_composés_complexes.grf). La méthode de construction des graphes est démontrée par le schéma de la Figure 44.

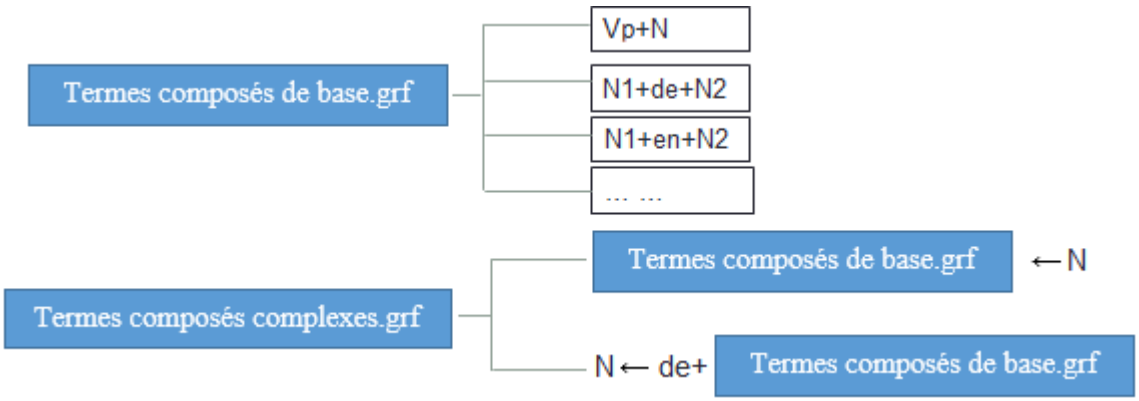

Figure 44 Méthode de construction des graphes

Un graphe qui fait appel à Termes_composés_de_base.grf et Termes_composés_complexes.grf est établi. On l'intègre dans la méthode distributionnelle. Pour la méthode supervisée, ce graphe est intégré à la position d'arguments dans le graphe qui est établi pour étiqueter les structures prédicat-argument. Pour la méthode semi-supervisée, il est intégré dans le graphe qui a pour objet d'identifier les structures prédicat-argument à partir des arguments donnés à l'avance et dans celui qui consiste à étiqueter les structures prédicat-argument à partir des prédicats récupérés. Les autres étapes de la méthode combinatoire pour l'acquisition automatique des noms d'artefacts sont pareilles à celles que nous utilisons dans la méthode distributionnelle présentée (tant supervisée que semi-supervisée). Finalement, on obtient une liste de termes qui comprend à la fois les noms d'artefacts monolexicaux et les noms d'artefacts polylexicaux.

2.1.2.2. Extension des termes

L'extension des termes est réalisée de deux façons : l'extension des termes monolexicaux et l'extension des termes polylexicaux. L'extension des termes monolexicaux est effectuée par la recomposition morphématique et la dérivation morphologique des termes monolexicaux reconnus. L'extension des termes polylexicaux est réalisée en exploitant les structures prédicat-argument extraites et les termes reconnus à partir desquels on peut former d'autres termes composés par des opérations morphologiques.

A. Extension des termes monolexicaux

Pour les termes monolexicaux obtenus, on applique la recomposition morphématique après leur segmentation en morphèmes par un script perl. L'objectif est d'obtenir les nouvelles combinaisons qui sont potentiellement des noms d'artefacts. Ensuite, on vérifie la présence des nouvelles bases dans le corpus à l'aide d'Unitex. Les bases candidates sont validées si elles sont étiquetées dans le corpus. À partir de la nouvelle liste de termes obtenue, on effectue la dérivation morphologique pour former les noms d'artefacts dérivés, par exemple,

55) corne ->cornette
56) plaque->plaquette
57) lanterne -> lanterneau
58) câble -> câbleau

Pareillement, les candidats termes dérivés sont aussi validés par leur présence dans le corpus. La dérivation avec le suffixe *–eur* à partir des prédicats extraits permet également d'obtenir des noms d'artefacts (par ex., *afficher->afficheur, modeler->modeleur, vibrer->vibreur*, nettoyer->nettoyeur, allumer->allumeur, etc.). Cependant, cette opération amène facilement des bruits (tels que *vider->videur, concevoir->concepteur, conduire->conducteur, prêter->prêteur, casser->casseur, presser->presseur*, etc.) par rapport à la dérivation morphologique à partir des noms, parce que les prédicats extraits ne sont pas tous les prédicats décrivant la fonction des artefacts. De plus, il est parfois difficile de décider si un prédicat décrit la

fonction des artefacts ou simplement une action ou une activité professionnelle (par ex., *presser* désigne une fonction d'un artefact ou une activité professionnelle ? *vider* indique la fonction d'un artefact ou simplement une action ?). La dérivation avec le suffixe *–eur* à partir des prédicats extraits ne sont pas pris en compte pour éviter trop d'erreurs.

B. Extension des termes polylexicaux

On peut exploiter certaines structures prédicat-argument extraites pour l'extension des termes polylexicaux du type V+N, par exemple,

59) chauffer des pains -> chauffe-pain [V+N]

60) *couper des jambons -> coupe-jambon* [V+N]

De plus, les termes composés (de base ou complexes) peuvent également être enrichis à partir des termes monolexicaux reconnus à l'aide d'une série d'opérations (telles que, Prefixe→, ←prep+N/GN/NAF, ←à+V, etc.), par exemple,

61) four-> four à micro-ondes [[N] ←à+N]

62) rouge à lèvres -> fixateur de rouge à lèvres [[N] ←de+NAF]

On intègre donc la liste de termes obtenue (par la méthode distributionnelle) dans la position des termes monolexicaux dans le graphe construit pour l'identification des noms d'artefacts composés (cf. Figure 44, p. 231). Ensuite, on relance le graphe et on essaie d'étiqueter les noms d'artefacts composés qui ne sont pas reconnus par l'étiquetage des structures prédicat-argument. Finalement, les candidats-termes composés obtenus sont validés s'ils sont présents dans le corpus et on calcule respectivement l'information mutuelle de chaque candidat-terme composé validé pour sélectionner les collocations.

2.2. Méthode combinatoire pour les noms de métiers

La combinaison de la méthode distributionnelle avec la méthode morphosémantique pour l'acquisition automatique des noms de métiers nécessite la mise en place d'un module de filtrage. La méthode distributionnelle s'exécute comme un module de filtrage sémantique

pour la méthode morphosémantique en vue de fournir une liste de bases permettant de dériver les termes de la classe sémantique de noms de métiers. Dans ce qui suit, on présente en détail les analyses des prédicats appropriés des noms de métiers et le moyen de mise en place du module de filtrage pour obtenir la méthode combinatoire pour l'acquisition automatique de noms de métiers.

2.2.1. Analyse des prédicats appropriés des noms de métiers

Un nom de métier réfère aux personnes qui participent régulièrement à une activité spécifique de travail. Les prédicats appropriés des noms de métiers peuvent être divisés en quatre classes en fonction de leurs distributions syntactico-sémantiques. La distribution syntactico-sémantique de la première classe Classe_1 (*embaucher, recruter, former,* etc.) est représentée par le patron syntaxique Nc/NMP+V+NMP (Nc réfère aux noms d'autres classes sémantiques et NMP désigne les noms de métiers. Par exemple, *embaucher* (*embaucher un chercheur*), *former* (*former les managers*), *encadrer* (*encadrer un stagiaire*), etc. La distribution syntactico-sémantique des prédicats appropriés de la deuxième classe Classe_2 est représentée par le patron syntaxique NMP+V. Les prédicats appropriés de la deuxième classe n'ont pas la forme passive. Par exemple, *avoir des compétences* (*un informaticien doit avoir des compétences de programmer*), *avoir pour mission* (*un designer a pour mission de « désigner »*), (*être*) *compétent* (*pour*) (*un vendeur est compétent pour promouvoir des produits*), (*être*) *responsable* (*de*) (*le chef de projet est responsable de projets*), etc.

Les prédicats appropriés de la troisième classe Classe_3, tel que *étudier* (*Le chercheur doit toujours étudier les nouveaux phénomènes dans son domaine.*), *gérer* (*Le manager gère tout le département de recherche.*), *optimiser* (*On cherche un ingénieur qui sera chargé d'optimiser la performance de notre outil.*), *animer* (*On a besoin d'un animateur pour animer une émission.*), *encadrer* (*Le maréchal encadre des troupes.*), etc. décrivent souvent une mission ou une activité spécifique exigée par une profession. La distribution syntactico-sémantique des prédicats appropriés de la troisième classe correspond au patron syntaxique NMP+V+N. Le patron syntaxique des prédicats appropriés de la quatrième classe concerne le complément introduit par une préposition. La quatrième classe se divise en deux sous-

classes : Classe_4a et Classe_4b. Classe_4a comprend les prédicats appropriés comme *être en contact avec qqn, prendre un rendez-vous avec qqn, avoir un entretient avec qqn, être en collaboration avec qqn,* etc. Le patron syntaxique des prédicats appropriés de Classe_4a est NMP/Nc+V+prep+NMP/Nc. Les prédicats de Classe_4a n'ont pas la forme passive non plus. Les prédicats appropriés de Classe_4b ont la distribution syntactico-sémantique NMP/Nc+V+Nc+prep+NMP/Nc. Par exemple, pour le prédicat *demander*, on peut avoir la phrase *Le secrétaire demande de l'aide auprès du technicien* dans laquelle *secrétaire* et *technicien* sont tous les noms de métiers. Un autre exemple, le prédicat *enseigner* dans la phrase *L'institutrice de l'école enseigne à des élèves* a deux noms de métiers comme argument. Le Tableau 29 liste toutes les informations sur l'analyse syntactico-sémantique des prédicats appropriés des noms de métiers.

Information / **Classes**	**Prédicats appropriés**	**Distribution syntactico-sémantique**
Classe_1	*charger, recruter, embaucher, former, employer, initier, passionner, chercher, engager*, etc.	Nc/NMP+V+NMP
Classe_2	*avoir des compétences de, être responsable de, être qualifié de, avoir pour mission de*, etc.	NMP+V
Classe_3	*optimiser, définir, animer, réaliser, encadrer, noter, gérer, diriger, étudier*, etc.	NMP+V+N
Classe_4		
Classe_4a	*être en contact, prendre contact, avoir un entretien, travailler, collaborer, être en collaboration*, etc.	NMP/Nc+V+prep+NMP/Nc
Classe_4b	*enseigner, demander*, etc.	NMP/Nc+V+Nc+prep+NMP/Nc

Tableau 29 Classement des prédicats appropriés de noms de métiers

La construction des patrons syntaxiques dans la méthode distributionnelle supervisée est fondée sur les patrons syntaxiques de base recensés dans le Tableau 29. Pour la méthode distributionnelle semi-supervisée, on prévoit quatre patrons syntaxiques de base en fonction de la position syntaxique de la distribution d'arguments à partir des prédicats appropriés des noms de métiers : V+NMP, NMP+V+N, N+V+prep+NMP et N+être+Vpp+par+NMP (V indique le verbe, prep réfère à la préposition sauf si c'est *par* et NMP signifie les noms de métiers). Chaque patron syntaxique prévu contient au moins une position syntaxique à laquelle les éléments linguistiques sont toujours les noms de métiers. Avec les patrons syntaxiques prévus et les arguments donnés à l'avance, les prédicats appropriés correspondants peuvent être localisés. D'autres patrons syntaxiques (tels que V+ADV+NMP,

NMP+V+ADV+ADV+N, NMP+être+Vpp, etc.) sont produits en plus à partir des quatre patrons syntaxiques de base.

2.2.2. Mise en place du module de filtrage

Pour combiner la méthode distributionnelle et la méthode morphosémantique pour l'acquisition automatique des noms de métiers, les étapes de la méthode distributionnelle sont aussi toutes appliquées et les patrons morphosyntaxiques sont intégrés dans les graphes d'étiquetage de structures prédicat-argument. Ensuite, les structures prédicat-argument extraites et les termes reconnus sont exploités pour enrichir le vocabulaire à l'aide des opérations morphologiques. Finalement, on calcule l'information mutuelle de chaque candidat-terme composé pour sélectionner les collocations. Dans ce qui suit, on détaille respectivement la méthode combinatoire qui se base sur la méthode distributionnelle supervisée et la méthode combinatoire qui se base sur la méthode distributionnelle semi-supervisée pour l'acquisition automatique des noms de métiers.

2.2.2.1. Méthode combinatoire basée sur l'apprentissage supervisé

A. Construction des graphes et extraction des termes

La construction des graphes dans le cadre de l'apprentissage supervisé est fondée sur le classement des prédicats appropriés des noms de métiers (cf. Tableau 29, p. 235). Deux graphes sont établis pour l'identification des structures prédicat-argument : le graphe qui a pour objet d'étiqueter les prédicats en faisant la désambiguïsation morphosyntaxique en fonction du contexte ; le graphe qui consiste à repérer les structures prédicat-argument en fonction de la distribution syntactico-sémantique des prédicats des noms de métiers. Le graphe de patrons morphosyntaxiques des noms de métiers MS_NMP.grf est intégré à la position d'arguments dans le graphe d'étiquetage de structures prédicat-argument. On applique ensuite les graphes au corpus et extrait les structures prédicat-argument étiquetées. Finalement, on calcule l'intersection sur les arguments pour obtenir les unités appartenant à classe sémantique de noms de métiers.

B. Extension des termes

Dans la méthode morphosémantique, on a établi une série de types de relations sémantiques entre les noms de métiers dérivés et leurs bases. Parmi les structures prédicat-argument du type NMP+V+N, on distingue ainsi trois sous-types : NMP+V+N (V=*avoir pour mission de* V_2*, avoir la compétence de* V_2*, être responsable de* V_2, etc.) dont V_2 peut être la base des noms de métiers en fonction du type de relation sémantique «X est la personne qui Y(verbe) », NMP+V+N (V=*étudier, conseiller, développer*, etc.) dont V peut être la base des noms de métiers en fonction du type de relation sémantique «X est la personne qui Y(verbe) » et NMP+V+N (V=*être spécialisé en, être dans le domaine de*, etc.) dont N peut être la base des noms de métiers en fonction du type « X est la personne qui est spécialiste en Y (discipline/domaine) ». Ces bases récupérées à partir des structures prédicat-argument extraites permettent de produire les noms de métiers en s'appuyant sur la morphologie dérivationnelle. Par exemple,

63) *La coordinatrice logistique de la formation a pour mission de gérer l'ensemble des locaux : gérer -> gérant*

64) *L'agence recrute un plieur numérique spécialisé en mécano-soudure : soudure ->soudeur -> mécano-soudeur*

65) *La spécialiste de la mode conseille et habille l'homme urbain : conseille->conseiller ; habiller ->habilleur*

Pour l'extension des termes, on récupère d'abord les bases permettant de faire la dérivation à partir des types de structures prédicat-argument choisis. Ensuite, on fait une segmentation morphématique des bases par un script perl dans le but de recomposer automatiquement les morphèmes obtenus pour enrichir la liste de bases. Et puis, la dérivation morphologique est effectuée pour former les dérivés. Les candidats-termes obtenus sont validés s'ils sont présents dans le corpus.

À partir des termes monolexicaux obtenus, on peut également poursuivre la détection des termes composés à l'aide des patrons morphosyntaxiques établis dans la méthode

morphosémantique (cf. Annexe 1, section 7), puisque de nombreux noms de métiers composés sont formés à partir des noms de métiers monolexicaux (tels que *batteur->batteur à œufs, batteur sur socle, accessoire de jeu, armoire à glace,* etc.). On intègre la liste de termes obtenue dans le graphe construit pour l'identification des noms de métiers composés et on le relance pour étiqueter plus de noms de métiers. Finalement, on calcule l'information mutuelle de chaque candidat-terme composé pour sélectionner les collocations. 90 est également décidé comme seuil.

2.2.2.2. Méthode combinatoire en s'appuyant sur l'apprentissage semi-supervisé

Dans la méthode combinatoire qui s'appuie sur l'apprentissage semi-supervisé pour l'acquisition automatique des noms de métiers, les structures prédicat-argument sont extraites à partir d'un ensemble d'arguments. Pareillement, un calcul probabiliste est intégré pour sélectionner le patron syntaxique adéquat pour chaque prédicat. Après le calcul de l'intersection appliquée aux candidats prédicats appropriés, une autre extraction de structures prédicat-argument est effectuée à partir des prédicats acquis. Les patrons morphosyntaxiques sont intégrés dans les graphes qui sont chargés d'étiqueter les structures prédicat-argument. Finalement, l'intersection est appliquée aux candidats noms de métiers reconnus et la méthode morphosémantique est appliquée de nouveau en exploitant les noms de métiers récupérés et les structures prédicat-argument extraites pour enrichir la liste de termes.

A. Extraction automatique des structures prédicat-argument

Le graphe qui consiste à étiqueter les structures prédicat-argument à partir des noms de métiers fournis à l'avance représente les quatre possibilités de distribution syntactico-sémantique de prédicat : V+NMP, NMP+V+N, N+V+prep+NMP et N+V+par+NMP. Les patrons syntaxiques recensés comprennent de nombreux patrons syntaxiques dérivés (tels que Vpassif+par+GNMP, GNMP+Vadjectif, GNMP+Va, etc.) à partir des quatre patrons syntaxiques de base V+NMP, NMP+V+N, N+V+prep+NMP et N+V+par+NMP. De plus, les patrons morphosyntaxiques pour reconnaître les noms de métiers composés sont intégrés à la position d'arguments dans les graphes établis pour étiqueter les structures prédicat-argument.

La Figure 45 est la capture d'écran d'une partie du résultat d'extraction des structures prédicat-argument pour les noms de métiers.

Figure 45 Résultat d'extraction des structures prédicat-argument

B. Intersection des candidats-prédicats et calcul des patrons syntaxiques

Dans les structures prédicat-argument extraites, les prédicats, les arguments et les patrons syntaxiques étiquetés sont extraits et structurés. Ensuite, l'intersection est appliquée aux candidats-prédicats reconnus.

Un calcul probabiliste est également effectué pour les patrons syntaxiques des noms de métiers extraits. Pour chaque prédicat, on calcule respectivement la probabilité d'avoir la distribution de noms de métiers à la position du sujet *P(s)*, la probabilité d'avoir la distribution de noms de métiers à la position du complément d'objet direct *P(cod)*, la probabilité d'avoir la distribution de noms de métiers à la position du complément introduit par la préposition *par P(par)* et la probabilité d'avoir la distribution de noms de métiers au complément introduit par la préposition *à, de* ou *avec P(p)*. Le calcul probabiliste est effectué à partir des informations extraites de patrons syntaxiques (telles que *s=vactif_gnmp, s=gnmp_va, s=gnmp_vpassif,* ect.). *P(s)* est calculé comme suit :

$$P(s) = \frac{c(sgnmp)}{c(s)} \tag{44}$$

c(sgnmp) réfère à la fréquence d'occurrence de patrons syntaxiques qui contiennent *gnmp* à la position du sujet. *c(s)* désigne la fréquence d'occurrence totale de tous les patrons syntaxiques

décrivant un prédicat. *P(cod)* est calculé selon l'équation suivante :

$$P(cod) = \frac{c(cgnmp)}{c(s)} \quad (45)$$

c(cgnmp) désigne la fréquence d'occurrence de patrons syntaxiques qui comprennent *gnmp* à la position du complément d'objet. *P(par)* est calculé par la formule suivante :

$$P(par) = \frac{c(pargnmp)}{c(s)} \quad (46)$$

c(pargnmep) indique la fréquence d'occurrence des patrons syntaxiques qui ont *gnmp* à la position du complément introduit par la préposition *par*. *P(p)* est calculé en fonction de l'équation suivante :

$$P(p) = \frac{c(pargnmp)}{c(s)} \quad (47)$$

c(pgnmp) est la fréquence d'occurrence des patrons syntaxiques qui contiennent *gnmp* à la position du complément introduit par la préposition *à, de* ou *avec*. Par exemple, dans la ligne :

66) charger:9;s=gnmp_vactif:6;s=gnmp_va:5;s=gnmp_vpassif:1

9 est la fréquence d'intersection du prédicat *charger*, 6, 5 et 1 indiquent respectivement la fréquence d'occurrence des patrons syntaxiques *s=gnmp_vactif*, *s=gnmp_va* et *s=gnmp_vpassif*. Pour le prédicat *charger*, *P(s)*, *P(cod)*, *P(par)* et *P(p)* sont respectivement calculés comme suit :

$$P(s) = \frac{c(sgnmp)}{c(s)} = \frac{6\ (s{=}gnmp\_vactif)}{6\ (s{=}gnmp\_vactif){+}5\ (s{=}gnmp_{va}){+}1\ (s{=}gnmp\_vactif)} = 0.5$$

$$P(cod) = \frac{c(pargnmp)}{c(s)} = \frac{5\ (s{=}gnmp\_va){+}1\ (s{=}gnmp\_vpassif)}{6\ (s{=}gnmp\_vactif){+}5\ (s{=}gnmp_{va}){+}1\ (s{=}gnmp\_vactif)} = 0.5$$

$$P(par) = \frac{c(pargnmp)}{c(s)} = \frac{0\ (s{=}par\_gnmp)}{6\ (s{=}gnmp\_vactif){+}5\ (s{=}gnmp_{va}){+}1\ (s{=}gnmp\_vactif)} = 0$$

$$P(p) = \frac{c(pgnmp)}{c(s)} = \frac{0\ (s{=}p\_gnmp)}{6\ (s{=}gnmp\_vactif){+}5\ (s{=}gnmp_{va}){+}1\ (s{=}gnmp\_vactif)} = 0$$

La Figure 46 est la capture d'écran du résultat de calcul probabiliste. Dans le résultat, le chiffre figurant devant la première virgule indique la fréquence d'intersection du prédicat ; le chiffre après la première virgule désigne la probabilité d'avoir la distribution de noms de métiers à la position du complément d'objet direct. Le troisième chiffre signifie la probabilité d'avoir la distribution de noms de métiers à la position du sujet. Le quatrième chiffre indique la probabilité d'avoir la distribution de noms de métiers à la position du complément introduit par la préposition *à, de* ou *avec* et le cinquième réfère à la probabilité d'avoir la distribution à la position du complément introduit par la préposition *par*. La dernière valeur est associée au quatrième chiffre et elle est composée de deux parties. Si la première partie égale False, la probabilité d'avoir la distribution de noms de métiers à la position du complément introduit par la préposition *à, de* ou *avec* égale 0. Sinon, la deuxième partie indiquera le nombre de prépositions différentes que le prédicat possède dans ses patrons syntaxiques.

```
[rechercher:26,1,0,0,0,False 0]
[recruter:19,0.947368421052632,0.0526315789473684,0,0,False 0]
[assister:11,0.727272727272727,0.272727272727273,0,0,False 0]
[confirmer:10,1,0,0,0,False 0]
[spécialiser:9,0.545454545454545,0.454545454545455,0,0,False 0]
[charger:9,0.5,0.5,0,0,False 0]
[industriel:7,0,1,0,0,False 0]
[chercher:6,1,0,0,0,False 0]
[expérimenter:6,0.375,0.625,0,0,False 0]
[commercial:6,0,1,0,0,False 0]
[travailler:6,0.555555555555556,0.444444444444444,0,0,False 0]
[immobilier:6,0,1,0,0,False 0]
[informatique:6,0,1,0,0,False 0]
```

Figure 46 Résultat du calcul probabiliste

Finalement, les prédicats sont classés en quatre groupes : les prédicats du premier groupe ont comme patron syntaxique V+NMP, les prédicats du deuxième groupe correspondent au patron syntaxique NMP+V, le patron syntaxique du troisième groupe est prep+NMP et celui du quatrième group est *par*+NMP. Si la probabilité d'avoir la distribution de noms de métiers à la position du complément introduit par la préposition *par* n'égale pas 0, le prédicat est classé dans le groupe *par*+NMP ; si la probabilité d'avoir la distribution de noms de métiers à la position du complément introduit par la préposition *à, de* ou *avec* n'égale pas 0, le prédicat est classé dans le groupe prep+NMP ; si la probabilité d'avoir la

distribution de noms de métiers à la position du complément d'objet direct est supérieur à la probabilité d'avoir la distribution de noms de métiers à la position du sujet, le prédicat est classé dans le V+NMP, sinon il est classé dans le groupe NMP+V.

C. Intégration des patrons morphosyntaxiques et extraction automatique des structures prédicat-argument avec des prédicats acquis

Les prédicats sélectionnés sont étiquetés en leur associant les patrons syntaxiques correspondants avec une série de balises (comme <v_nmp></v_nmp>, <nmp_v></nmp_v>, <par_nmp></par_nmp> et <p_nmp></p_nmp>). Le graphe établi pour l'identification des structures prédicat-argument comprend quatre types de structures prédicat-argument : V+NMP, NMP+V+N, N+V+prep+NMP et N+V+par+NMP qui sont étiquetées respectivement par <v_nmp></v_nmp>, <nmp_v></nmp_v>, <par_nmp></par_nmp> et <p_nmp></p_nmp>. Ensuite, les patrons morphosyntaxiques des noms de métiers sont intégrés à la position d'arguments dans les graphes établis pour étiqueter les structures prédicat-argument. Finalement, tous les noms de métiers identifiés sont extraits et l'intersection est appliquée aux candidats-termes pour obtenir le résultat final.

D. Extension des termes

Pareillement, pour les noms de métiers, les structures prédicat-argument sont également exploitées pour obtenir les bases permettant de faire la dérivation. La segmentation et la recomposition morphématique sont appliquées aux bases obtenues pour enrichir la liste de bases. Et puis, la dérivation morphologique sert à former les dérivés. Les termes composés sont également enrichis à partir des termes obtenus. Finalement, on fait appel au calcul de l'information mutuelle pour sélectionner les collocations parmi candidats-termes composés.

3. Évaluation

L'évaluation de la méthode combinatoire s'effectue en tenant compte des aspects quantitatif et qualitatif. La méthode combinatoire est appliquée à deux vocabulaires : les noms de métiers et les noms d'artefacts. Ainsi, une évaluation quantitative et une évaluation

qualitative sont faites respectivement pour les noms d'artefacts et les noms de métiers. Pour l'évaluation quantitative, le résultat est évalué en comparant avec le standard. Le standard est établi par l'annotation manuelle. Pour l'évaluation qualitative, on analyse respectivement les types d'erreurs et les types de silences des résultats.

2.3. Évaluation quantitative

3.1.1. Évaluation quantitative pour les noms d'artefacts

On a respectivement expérimenté la méthode combinatoire basée sur l'apprentissage supervisé et la méthode combinatoire basée sur l'apprentissage semi-supervisé pour l'acquisition automatique des noms d'artefacts. Une évaluation du résultat obtenu par l'apprentissage supervisé et une évaluation du résultat obtenu par l'apprentissage semi-supervisé sont respectivement effectuées. On annote les résultats manuellement pour étiqueter les faux positifs, les vrais positifs et les faux négatifs. Ensuite, on calcule respectivement le taux de précision, le taux de rappel et le taux de F-mesure. Pour la méthode basée sur l'apprentissage supervisée, on fournit préalablement la même liste de prédicats utilisée dans l'expérimentation de la méthode distributionnelle supervisée. Pour la méthode basée sur l'apprentissage semi-supervisée, on fournit à l'avance les mêmes trois classes d'arguments utilisées dans l'expérimentation de la méthode distributionnelle semi-supervisée.

On fait le même test que dans la méthode distributionnelle pour décider le seuil d'intersection d'arguments. Finalement, le nombre 5 est décidé comme seuil d'intersection dans la méthode combinatoire basée sur l'apprentissage supervisé. Ce seuil permet d'obtenir un meilleur résultat. Dans la méthode combinatoire basée sur l'apprentissage semi-supervisé, on choisit le nombre 2 comme le seuil d'intersection d'arguments des trois classes sémantiques. 3, 3 et 2 sont respectivement sélectionnés comme les nombres d'itérations de la classe sémantique des contenants d'arrangement, de la classe sémantique des appareils de cuisson et de la classe sémantique des moyens de transport. Le Tableau 30 et le Tableau 31 fournissent respectivement une liste des résultats d'évaluation de la méthode combinatoire basée sur l'apprentissage supervisé et de la méthode combinatoire basée sur l'apprentissage

semi-supervisé.

Classes d'arguments	Taux de précision	Taux de rappel	F-mesure
Noms d'artefacts	90.01%	76.20%	82.53%

Tableau 30 Résultat d'évaluation de la méthode combinatoire supervisée pour les noms d'artefacts

Classes d'arguments	Taux de précision	Taux de rappel	F-mesure
Moyen_Transport_Routier	62.79%	59.65%	61.18%
Appareil_Cuisson	77.18%	79.33%	78.24%
Contenant_Rangement	87.27%	84.46%	85.84%

Tableau 31 Résultat d'évaluation de la méthode combinatoire semi-supervisée pour les noms d'artefacts

3.1.2. Évaluation quantitative pour les noms de métiers

La même méthode d'évaluation quantitative est appliquée pour la méthode combinatoire basée sur l'apprentissage supervisé et celle basée sur l'apprentissage semi-supervisé pour l'acquisition automatique des noms de métiers. Pour l'expérimentation de la méthode combinatoire basée sur l'apprentissage supervisé, on donne à l'avance une quarantaine de prédicats de la classe sémantique de noms de métiers. Le nombre 2 est décidé comme seuil d'intersection après le test. Pour l'expérimentation de la méthode combinatoire basée sur l'apprentissage semi-supervisé, on fournit une centaine de noms de métiers à l'avance. On choisit respectivement 2 et 3 comme le seuil d'intersection et le nombre d'itération après plusieurs tests. Dans les Tableaux 32 et 33, on liste respectivement les résultats d'évaluation de la méthode combinatoire basée sur l'apprentissage supervisé et ceux de la méthode combinatoire basée sur l'apprentissage semi-supervisé.

Classes d'arguments	Taux de précision	Taux de rappel	F-mesure
Noms de métiers	86.51%	75.32%	80.03%

Tableau 32 Résultat d'évaluation de la méthode combinatoire supervisée pour les noms de métiers

Classes d'arguments	Taux de précision	Taux de rappel	F-mesure
Noms de métiers	81.07%	80.97%	81.02%

Tableau 33 Résultat d'évaluation de la méthode combinatoire semi-supervisée pour les noms de métiers

3.2. Évaluation qualitative

3.2.1. Évaluation qualitative pour les noms d'artefacts

Pour analyser les types d'erreurs dans le résultat de la méthode combinatoire basée sur l'apprentissage supervisé, on a choisi 150 termes au hasard dans le résultat final et on a relevé 14 erreurs. On recense essentiellement cinq types d'erreurs : le type A regroupe les termes dont les référents varient selon les contextes, tels que *type, version*, etc. ; le type B comporte les erreurs dépendant de certains prédicats qui ne sont pas les prédicats appropriés des noms d'artefacts mais qui sont quand même utilisés pour repérer les arguments afin d'augmenter le rappel du résultat, par exemple, dans *laver les pommes, laver les mai*ns*, laver la vaisselle* et *laver les vêtements*, le prédicat *laver* est utilisé pour repérer les noms d'artefacts comme *la vaisselle* et *les vêtements*, alors que *les pommes* et *les mains* sont reconnus à tort comme noms d'artefacts ; le type C regroupe les erreurs provoquées parce que les groupes nominaux sont reconnus à tort comme termes composés, par exemple, *voiture sur la bande gauche* (-6.08449941307517); le type D regroupe les erreurs obtenues en repérant des constituants syntaxiques à tort, par exemple, *maison à chèque cadeau* (-5.77144112313002) est reconnu à partir de la phrase *ça va de linge de maison à chèques cadeaux* ; le type E regroupe les erreurs produites par l'extension des termes composés à partir des termes d' autres classes sémantiques, par exemple, le terme *résistance de forme circulaire* (-4.17438726989564) est formé à partir de *résistance* qui est reconnu à tort comme nom d'artefact. Dans le Tableau 34, on liste les types d'erreurs recensés.

Types d'erreurs	**Nombre**	**Proportion**	**Exemples**
A	3	21.43%	*type* (*assemble de type*), *version* (*version fermée*), *reste* (*remplir le reste*)
B	4	28.57%	*pomme* (*laver les pommes*), *eau* (*verser de l'eau /retirer l'eau des tomates farcies*), *résistance* (*bouger des résistances mécaniques*), *discussion* (*poser une question dans une nouvelle discussion*)
C	1	7.14%	*voiture sur la bande gauche* (-6.08449941307517)
D	1	7.14%	*maison à chèque cadeau* (-5.77144112313002, *ça va de linge de maison à chèques cadeaux*)
E	5	35.71%	*résistance de forme circulaire* (-4.17438726989564), *versions non critiques* (-5.26785815906333), *version rocks concept* (-6.92461239604856), *partie raccord de départ* (-5.94017125272043), *reste des légères taches* (-6.4377516497364)
Total	14	100%	

Tableau 34 Types d'erreurs dans le résultat de la méthode combinatoire supervisée pour les noms d'artefacts

Ensuite, on étiquette le corpus avec le résultat obtenu (la liste de termes récupérés) et on analyse une soixantaine de textes contenant 30 silences. Les silences dans la méthode combinatoire basée sur l'apprentissage supervisé dépendent principalement de quatre causes : l'intersection des arguments supprime certains termes qui doivent être reconnus (type de silence A) ; le calcul de l'information mutuelle peut également éliminer les termes qui doivent être reconnus (type B) ; les termes qui ne se trouvent pas dans des structures prédicat-argument ou qui coexistent souvent avec les prédicats basiques (comme *avoir, mettre, trouver,* etc.) sont en général absents dans le résultat (type C) ; la coexistence fréquente avec les prédicats appropriés qui ne sont pas donnés à l'avance conduirait aussi aux silences (type D). Pour le type A, on distingue encore quatre raisons (cf. Tableau 35).

Types de silences	**Nombre**	**Proportion**	**Exemples**
A	22	70.97%	*sphère, capteur, catalyseur, amortisseur, courroie, bougie, afficheur, autoradio, ecran, Berlingo, cartouche, coffre, eprom, vérin, ambulance, corbillard, rétro, saxo, flaque, capuchon, chiffon, rotule*
Raison 1 : présence fréquente dans les structures non prédicat-argument			*saxo* (*achat saxo 1.5l d _ citroën _ forum marques/problème direction assisté saxo 1.5d 2001 _ saxo _ citroën _ forum marques*), *flaque* (*Pas énormément, mais une flaque/Je démarre et la un grosse flaque d'huile K*), *capuchon* (*un système de ventilation sur le capuchon/cerise sur le gâteau, ou capuchon sur le stylo*), etc.
Raison 2 : coexistence fréquente avec les prédicats basiques			*saxo* (*acquérir un/une saxo/touche les saxos/parler de ta saxo*), *flaque* (*se trouve la flaque d'eau/ça fait une jolie flaque/il y a fréquemment une petite flaque d'eau), rotule* (*tu mets de graisse sur les rotules/commander une rotule de suspension*), *capuchon* (*remettre le capuchon/la présence de capuchon*), *chiffon* (*n'oublier pas de mettre un chiffon dessous*), etc.
Raison 3 : liste incomplète de prédicats appropriés données à l'avance			*rotule* (*déboiter des rotules*), *capuchon* (*dévisser le capuchon*), etc.
Raison 4 : manque de patrons syntaxiques (ou morphosyntaxiques)			*rotule* (*tu as une référence de rotule/la coquille s'est fondue à l'enclenchement sur la rotule/j'ai vu le prix de la rotule*), *chiffon* (*étaient parfaitement récollées, d'un chiffon humide/essuyer la carrosserie à l'aide du chiffon/un petit coup d'un chiffon sec enlève toute bidons ou des chiffons*), etc.
B	1	3.22%	*auto car* (-17.4303698617868)
C	1	3.22%	*Citroën*
D	7	22.58%	*plip, filtre, boulon, chauffe, Ulysse, coque, étrier*
Total	31	100%	

Tableau 35 Types de silences dans le résultat de la méthode combinatoire supervisée pour les noms d'artefacts

Pour la méthode combinatoire semi-supervisée, on sélectionne respectivement 60 termes au hasard dans le résultat de la classe sémantique des contenants d'arrangement, de la classe sémantique des appareils de cuisson et de la classe sémantique des moyens de transport. Il y a trois explications pour les erreurs de ces trois classes sémantiques : les erreurs peuvent être produites par les prédicats basiques récupérés à partir des arguments donnés à l'avance, par exemple, à partir du nom d'artefact *caisse* (*torpiller la caisse),* on obtient le bruit *travail (torpiller le travail*) (type d'erreur A) ; les erreurs peuvent être les termes dont les référents varient en fonction du contexte (type B) ; les erreurs peuvent également produites par l'extension des termes (type C). Les erreurs causées par l'extension des termes dépendent des raisons suivantes : les groupes nominaux sont reconnus à tort comme termes composés ; les termes composés sont repérés avec des constituants reconnus à tort ; les termes composés des autres classes sémantiques sont obtenus parce qu'ils sont formés à partir des termes des

autres classes sémantiques. Dans le Tableau 36, on liste en détail les informations sur les erreurs de chaque classe sémantique.

Classes sémantiques	**Types d'erreurs**	**Nombre**	**Proportion**	**Exemples**
Contenants d'arrangement				
	A	5	41.67%	*stylo* (*trimbaler la trousse -> se trimbaler avec plein de stylos à la poche), travail (torpiller la caisse -> torpiller le travail*), etc.
	B	1	8.33%	*objet*
	C			
	Raison1: groupes nominaux	2	16.67%	*poches jetables ultra* (-6.12249280951439), etc.
	Raison2: avec des constituants reconnus à tort	2	16.67%	*frigo marche nickel* (*fait le frigo marche nickel enfin il congelait au début que je l avais mais ça s est ensuite arreté*), etc.
	Raison3: reconnus à partir de termes d'autres classes sémantiques	2	16.67%	*stylo à encre gel colorée* (-6.96224346426621), marqueur véléda (-3.93182563272433)
	Total	12	100%	
Appareils de cuisson				
	A	4	28.57%	*batterie*/*cercle* (*réchauffer le poêle/four -> réchauffer la batterie/le cercle*), etc.
	B	4	28.57%	*maximum* (*bronzer le maximum*), *gamme* (*compléter la gamme*), etc.
	C			
	Raison1: groupes nominaux	4	28.57%	*plateau acier amovible* (-4.30406509320417), *cuisine outre atlantique* (-2.19722457733622),…
	Raison3: reconnus à partir de termes d'autres classes sémantiques	2	14.28%	*batterie à électrolyte liquide* (-6.20455776256869), console de microsoft (-21.9552044192077)
	Total	14	100%	
Moyens de transport				
	A	16	66.67%	*stylo* (*manier la voiture -> manier le stylo*), *maison* (*louer une voiture -> louer une maison*), *présentation*, *site, carte, ville, message*, etc.
	B	2	8.33%	*version, objet*
	C			
	Raison1: groupes nominaux	2	8.33%	*voiture avec un logiciel de retouche* (-4.70048036579242), etc.
	Raison3: reconnus à partir de termes des autres classes sémantiques	4	16.67%	*stylo à encre gel coloré* (-6.96224346426621), stylo en chocolat noire (-7.68662133494462), etc.
	Total	24	100%	

Tableau 36 Types d'erreurs dans le résultat de la méthode combinatoire semi-supervisée pour les noms d'artefacts

Pour analyser les types de silences, on évalue respectivement une cinquantaine de textes et on trouve respectivement 6, 8 et 11 silences pour la classe sémantique des contenants d'arrangement, la classe sémantique des appareils de cuisson et la classe sémantique des moyens de transport. L'intersection (type de silence A), le calcul de l'information mutuelle (type B) et la présence fréquente dans les structures non prédicat-argument ou la coexistence fréquente avec les prédicats basiques (type C) sont les raisons principales pour les silences. Dans le Tableau 37, on liste les types de silences de chaque classe sémantique avec les exemples correspondants et la proportion de chaque type de silence.

Classes sémantiques	**Types de silences**	**Nombre**	**Proportion**	**Exemples**
Contenants d'arrangement				
	A	2	33.33%	*cocotte*, *sceau*
	B	1	16.67%	*frigo congélateur* (-7.33926198292358)
	C	3	50.00%	*cellier* (*j'ai bien une prise femelle de ce type mais dans mon cellier mais*), *solive* (*solives qui reposent sur deux bastaings de 90x90*), *cloison* (*as _tu prévu une cloison ?*)
	Total	6	100%	
Appareils de cuisson				
	A	2	25.00%	*bol de lait*, *casserole*
	C	6	75.00%	*sautoir* (*mettre les dolmades dans un sautoir), yaourtière /crêpière* (*a cuisine avec les yaourtières, crêpières*), *grille* (*nettoyer le filtre noir des grilles metaliques*), *Moulinex* (*acheter une Moulinex*), *brûleur gaz* (*je suis impressionné par la faculté des bruleurs gaz*)
	Total	8	100%	
Moyens de transport				
	B	1	9.09%	*auto car* (-17.4303698617868)
	C	10	90.91%	*camion* (*mettre sur le camion/j'ai un camion branché*), *camionnette* (*si la jolie camionnette bleue passe dans votre ville*), *Fiat* (*j'ai déja eu une FIAT*), *citroën* (*titre:temps d` attente livraison _ citroën _ forum marques*), etc.
	total	11	100%	

Tableau 37 Types de silences dans le résultat de la méthode combinatoire semi-supervisée pour les noms d'artefacts

3.2.2. Évaluation qualitative pour les noms de métiers

Dans le résultat de la méthode combinatoire supervisée pour les noms de métiers, on sélectionne 150 termes au hasard et on trouve 18 erreurs. Il y a quatre types d'erreurs : le type A comporte les erreurs produites par la présence de certains prédicats qui ne sont pas les

prédicats appropriés des noms de métiers mais qui sont utilisés quand même pour augmenter le rappel du résultat ; le type B regroupe les erreurs provoquées du fait que les groupes nominaux sont reconnus à tort comme des termes composés; le type C regroupe les erreurs obtenues en repérant des constituants syntaxiques à tort, par exemple, dans un terme reconnu à tort *auteur du rapport avance des recommandations*, *avance* est le verbe du sujet *auteur du rapport* au lieu du constituant du nom composé . Dans le Tableau 38, on liste la proportion de chaque type d'erreur et les exemples associés.

Types d'erreurs	**Nombre**	**Proportion**	**Exemples**
A	5	26.31%	*poste* (*poste recherché*), *secteur* (*secteur spécialisé*/*chargé*), *métier* (*métier spécialisé*/*recherché*), *semaine* (*semaine chargée/formation de semaine*), *personne* (*engagement des personnes/les personnes expérimentées/embauchées*)
B	9	47.37%	*animateur interne* (-13.7295386405493), *auditeur expérimenté* (-9.59601085275617), *chefs d'entreprise sur les enjeux de l'emploi*, *ingénieur jeune diplômé* (-8.85865296954849), *gestionnaire rigoureux* (-37.7380765658223), *gestionnaire paie en avril* (-6.01859321449623), *personne handicapée* (-42.1212185205515)
C	6	31.58%	*auteur du rapport avance des recommandations*, *adjoint de direction futur directeur*, *chef de chantier à bac*, *auditeurs justifiant*, *expert métier reconnu*, *libraire doit*
Total	19	100%	

Tableau 38 Types d'erreurs dans le résultat de la méthode combinatoire supervisée pour les noms de métiers

Pour analyser les types de silences dans la méthode combinatoire supervisée, on étiquette le corpus avec le résultat obtenu (la liste de termes récupérés) et on analyse une soixantaine de textes qui contiennent environ 32 silences. Les silences sont causés par l'intersection des arguments (type de silence A), par le calcul de l'information mutuelle (type B), par la présence fréquente dans les structures non prédicat-argument ou les structures prédicat-argument non reconnues en l'absence de patrons syntaxiques (type C) et par la coexistence fréquente avec les prédicats basiques (comme *avoir, mettre, trouver,* etc.) ou les prédicats appropriés qui ne sont pas donnés à l'avance (type D). De plus, il existe aussi cinq raisons différentes pour expliquer les silences de type A. Dans le Tableau 39, on liste tous les types de silences, la proportion de chaque type de silence et les exemples correspondants.

Types de silences	**Nombre**	**Proportion**	**Exemples**
A	15	46.87%	*acheteur, dirigeant, conseiller clientèle, graphiste, directeur, conseiller service client à distance, développeur, téléacteur, secrétaire médicale, photographe, spécialiste, négociant, téléconseiller, gestionnaire de paie, saisonnier*
Raison 1 : présence fréquente dans les structures non prédicat-argument			*acheteur, dirigeant, conseiller clientèle, directeur, conseiller service client à distance, artiste, photographe, spécialiste, gestionnaire de paie*
Raison 2 : coexistence fréquente avec les prédicats basiques			*graphiste (un graphiste avait imaginé, un graphiste qui ne vous est pas inconnu), secrétaire médical (Muriel est devenu secrétaire médical), photographe(c'est fait par un photographe)*
Raison 3 : manque de prédicats appropriés dans la liste donnée à l'avance			*artiste, gestionnaire de paie*
Raison 4 : manque de patrons syntaxiques (ou morphosyntaxiques) pour repérer les structures prédicat-argument (ou les termes composés)			*développeur, téléconseiller, saisonnier*
Raison 5 : coexistence avec un seul prédicat approprié			*téléacteur(former le téléacteur) , éducateur (spécialisé), négociant (négociant spécialisé)*
B	4	12.50%	*chef de projet déploiement* (-90.7472471973139), *gestionnaire de clientèle* (-93.0442456563559), *gestionnaire de sinistre* (-99.072018869574), *chef du département* (-96.6468537534106)
C	5	15.63%	*masseur-kinésithérapeute*, *concepteur-vendeur*, *tuteur*, *technicien supérieur agricole*, *technicien de maintenance*
D	8	25.00%	*administrateur des bases de données*, *administrateur*, *administrateur de réseaux*, *assistant de gestion PME PMI*, *assistant manager*, *assistant*, *électromécanicien*, *mécanicien carrossier*
Total	32	100%	

Tableau 39 Types de silences dans le résultat de la méthode combinatoire supervisée pour les noms de métiers

Dans le résultat de la méthode combinatoire semi-supervisée pour les noms de métiers, on sélectionne 150 termes au hasard à partir du résultat et trouve 27 erreurs. Elles sont les erreurs occasionnées par les prédicats basiques récupérés à partir des arguments fournis préalablement (type d'erreur A), les erreurs provoquées du fait que les groupes nominaux sont reconnus à tort comme termes composés et les erreurs produites par les constituants syntaxiques reconnus à tort. Dans le Tableau 40, on liste les trois types d'erreurs, la proportion de chaque type et les exemples correspondants.

Types d'erreurs	**Nombre**	**Proportion**	**Exemples**
A	7	25.92%	*secteur* (*secteur spécialisé*/*chargé*), *métier* (*métier spécialisé*/*recherché*), *semaine* (*semaine chargée*/*formation de semaine*), *personne* (*engagement des personnes*/*les personnes expérimentées*/*embauchées*), *carrière*, *social*, *collègue*
B	11	40.74%	*animateur interne* (-13.7295386405493), *auditeur expérimenté* (-9.59601085275617), *gestionnaire paie en avril* (-6.01859321449623), *gestionnaire rigoureux* (-37.7380765658223), *personne handicapée* (-42.1212185205515), *expert reconnu* (-28.3628404604782), *cuisinier nomade* (-5.12396397940326), *libraire indépendant* (-7.60688453121963), *ingénieur jeune diplômé* (-8.85865296954849)
C	9	33.33%	expert dédié, auteur du rapport avance des recommandations, adjoint de direction futur directeur, chef de chantier à bac, auditeurs justifiant, expert métier reconnu, libraire doit, chef de projet mise, chef lundi (-13.2759767700615)
Total	27	100%	

Tableau 40 Types d'erreurs dans le résultat de la méthode combinatoire semi-supervisée pour les noms de métiers

Pour analyser les types de silences, on étiquette le corpus avec le résultat obtenu et analyse une soixantaine de textes dans lesquels on trouve 19 silences. L'intersection des arguments, le calcul de l'information mutuelle et la fréquente présence dans les structures non prédicat-argument ou la coexistence avec les prédicats basiques sont également les causes principales des silences. Dans le Tableau 41, on liste les informations sur chaque type de silence.

Types de silences	**Nombre**	**Proportion**	**Exemples**
A	7	36.84%	*couvreur*, *conseiller de vente*, *employeur*, *entrepreneur*, *ouvrier*, *directeur*, *opérateur*
B	4	21.05%	*chef de projet déploiement* (-90.7472471973139), *gestionnaire de clientèle* (-93.0442456563559), *gestionnaire de sinistre* (-99.072018869574), *chef du département* (-96.6468537534106)
C	8	42.11%	*masseur-kinésithérapeute*, *concepteur-vendeur*, *tuteur*, *technicien supérieur agricole*, *technicien de maintenance*, *administrateur des bases de données*, *administrateur*, *administrateur de réseaux*, *assistant de gestion PME PMI*, *assistant manager*, *assistant*, *électromécanicien*, *mécanicien carrossier*
Total	19	100%	

Tableau 41 Types de silences dans le résultat de la méthode combinatoire semi-supervisée pour les noms de métiers

Troisième partie : Présentation des résultats

Chapitre 1 Analyses des résultats et comparaison des méthodes

Dans ce chapitre, on présente une comparaison des résultats obtenus à partir de différentes méthodes et on présente les principales raisons qui expliquent ces différents résultats et quelle est la différence entre les différentes méthodes. À travers ces analyses, on essaie de mieux montrer les différents fonctionnements des trois méthodes présentées dans cette thèse et les avantages ainsi que les inconvénients de chaque méthode. On présentera en premier lieu la comparaison des résultats d'évaluation quantitative. Ensuite, on présentera celle des résultats d'évaluation qualitative.

1. Comparaison des résultats d'évaluation quantitative

La précision de la méthode distributionnelle supervisée peut atteindre 89.40% et le rappel est de 71.78%. Par rapport à la méthode distributionnelle supervisée, la précision de la méthode distributionnelle semi-supervisée est plus basse mais le rappel est beaucoup augmenté. L'avantage de la méthode distributionnelle semi-supervisée est qu'elle permet d'augmenter la capacité du système à donner toutes les solutions pertinentes (i.e. le rappel du résultat). Le rappel de la méthode distributionnelle supervisée est dépendant de la quantité des prédicats appropriés donnés à l'avance. Plus le rappel qu'on veut obtenir est haut, plus il y aura de prédicats appropriés qu'on doit fournir à l'avance. Néanmoins, la méthode distributionnelle semi-supervisée permet de repérer plus de prédicats appropriés et d'arguments à partir d'un nombre limité d'arguments.

Par rapport à la méthode distributionnelle, la précision de la méthode morphosémantique est beaucoup plus élevée ; elle peut atteindre 93.89%. La méthode morphosémantique est plutôt une méthode d'extension des termes. Elle enrichit la liste de termes donnée à l'avance par la recomposition morphématique, la dérivation et la formation

des noms composés. Ces opérations morphologiques sont effectuées au sein des unités lexicales et elles occasionnent moins de bruits par rapport à la méthode distributionnelle qui consiste à identifier les structures syntaxiques dans tous les textes. Cependant, le rappel du résultat obtenu par la méthode morphosémantique est également déterminé par la quantité de termes fournis à l'avance.

La comparaison avec la méthode combinatoire permet de mieux visualiser les fonctionnements et les intérêts de chaque méthode (méthode distributionnelle, méthode morphosémantique et méthode combinatoire). Pour les noms d'artefacts, la méthode combinatoire basée sur l'apprentissage supervisé augmente la précision du résultat jusqu'à 90.01% et le rappel jusqu'à 76.20%, alors que la précision et le rappel du résultat obtenu par la méthode distributionnelle supervisée est seulement de 89.40% et de 71.78%. Pareillement, la méthode combinatoire basée sur l'apprentissage semi-supervisé pour l'acquisition automatique des noms d'artefacts augmente également la précision et le rappel du résultat. La précision et le rappel sont respectivement augmentés environ de 7% et de 3% pour la classe sémantique des appareils de cuisson et la classe sémantique des contenants d'arrangement. Pour la classe sémantique des moyens de transport, la précision est seulement augmentée de 0.33% et le rappel est seulement augmenté de 1.12% du fait qu'il existe moins de termes composés dans la classe sémantique des moyens de transport. Bref, l'intégration de la méthode morphosémantique dans la méthode distributionnelle dans le cadre de l'acquisition automatique des noms d'artefacts permet d'augmenter non seulement le rappel par l'extension des termes mais aussi la précision en considérant les termes composés.

Pour les noms de métiers, la précision et le rappel du résultat obtenu par la méthode combinatoire basée sur l'apprentissage supervisé sont respectivement de 86.51% et de 75.32%. La précision et le rappel du résultat obtenu par la méthode combinatoire basée sur l'apprentissage semi-supervisé sont de 81.07% et de 80.97%. Par rapport à la méthode morphosémantique (dont la précision et le rappel sont de 93.89% et de 78.10%), le taux de précision de la méthode combinatoire est beaucoup baissé (environ de 10%) mais le rappel est augmenté (environ de 3%). Dans la méthode morphosémantique, on ne peut étiqueter que 31.89% de noms de métiers avec la liste de bases donnée à l'avance (environ 150 termes),

alors que la méthode morphosémantique permet d'enrichir les termes jusqu'à 78.10% de rappel. En combinant la méthode distributionnelle, le rappel de la méthode combinatoire pour l'acquisition automatique des noms de métiers est encore augmenté de 3%.

La méthode morphosémantique permet d'augmenter davantage le rappel du résultat des noms de métiers que celui des noms d'artefacts du fait que le vocabulaire des noms de métiers est plus riche de caractéristiques morphologiques par rapport au vocabulaire des noms d'artefacts. La baisse de précision de la méthode combinatoire pour les noms de métiers est causée par l'intégration de la méthode distributionnelle qui produit plus de bruits par rapport à la méthode morphosémantique. Dans le Tableau 42, on présente une liste les résultats d'évaluation quantitative de toutes les méthodes pour comparaison.

Méthodes			**Précision**	**Rappel**	**F-mesure**
Méthode distributionnelle (NAF)	**Supervisée**		89.40%	71.78%	79.63%
	Semi-supervisée	Moyens de transport	62.46%	58.53%	60.43%
		Appareils de cuisson	70.14%	76.87%	73.35%
		Contenants d'arrangement	81.34%	81.02%	81.20%
Méthode morphosémantique (NMP)			93.89%	de 31.89% à 78.35%	85.30%
Méthode combinatoire					
NAF	**Supervisée**		90.01%	76.20%	82.53%
	Semi-supervisée	Moyens de transport	62.79%	59.65%	61.18%
		Appareils de cuisson	77.18%	79.33%	78.24%
		Contenants d'arrangement	87.27%	84.46%	85.84%
NMP	**Supervisée**		86.51%	75.32%	80.03%
	Semi-supervisée		81.07%	80.97%	81.02%

Tableau 42 Comparaison des résultats d'évaluation des trois méthodes

2. Comparaison des résultats d'évaluation qualitative

Pour la méthode distributionnelle supervisée, la plupart des erreurs dépendent des prédicats qui ne sont pas appropriés mais sont utilisés quand même pour constituer des prédicats appropriés définitionnels afin d'augmenter le taux de rappel du résultat. Pour la méthode distributionnelle semi-supervisée, la plupart des erreurs sont produites par les

prédicats non appropriés récupérés à partir des arguments donnés à l'avance. Le fait que certains constituants syntaxiques sont reconnus à tort est aussi une raison majeure qui explique les erreurs dans la méthode distributionnelle supervisée. Néanmoins, dans la méthode distributionnelle semi-supervisée, c'est plutôt la sélection des patrons syntaxiques basée sur un calcul probabiliste qui provoque la reconnaissance d'erreurs touchant certains constituants syntaxiques. Pour la méthode distributionnelle (supervisée et semi-supervisée), il existe trois raisons principales pour les silences : il arrive souvent que les termes qui doivent être reconnus ne se trouvent pas fréquemment dans les structures prédicat-argument ; les termes qui doivent être reconnus coexistent souvent avec les prédicats basiques ; le manque de patrons syntaxiques (ou de patrons morphosyntaxiques) pour repérer les structures prédicat-argument (ou les termes composés). La pertinence de la méthode distributionnelle dépend largement de la pertinence de l'identification des structures prédicat-argument. L'identification ainsi que la sélection des prédicats appropriés définitionnels et des patrons syntaxiques correspondants déterminent directement la précision du résultat.

Par rapport à la méthode distributionnelle, la plupart des erreurs dans la méthode morphosémantique provient de l'extension des termes composés, tels que les termes composés reconnus à tort à cause de l'ambiguïté morphosyntaxique, les groupes nominaux mal reconnus comme termes composés, les termes composés reconnus avec des constituants en trop, etc. Le calcul d'information mutuelle ne permet d'éliminer qu'une partie des bruits. En ce qui concerne les types de silences, ils sont plus variables. Cependant, parmi tous les types de silences, les listes incomplètes de termes données à l'avance sont les causes les plus importantes des silences .

La méthode combinatoire combine la méthode distributionnelle et la méthode morphosémantique. Les types d'erreurs dans la méthode combinatoire comprennent ceux qui sont causés par l'identification des structures prédicat-argument et ceux qui sont causés par l'extension des termes composés. Dans la méthode combinatoire (tant supervisée que semi-supervisée) pour l'acquisition automatique des noms de métiers, les erreurs provoquées du fait que les groupes nominaux sont reconnus à tort comme termes composés occupent la proportion la plus importante (47.37% pour la supervisée et 40.74% pour la semi-supervisée).

Néanmoins, dans la méthode combinatoire basée sur l'apprentissage supervisé pour l'acquisition automatique des noms d'artefacts, les erreurs produites par les prédicats non appropriés de la classe sémantique et les erreurs provoquées du fait que les groupes nominaux sont reconnus à tort comme des termes composés occupent les plus importantes proportions (28.57% et 35.71% respectivement).

Dans la méthode combinatoire basée sur l'apprentissage semi-supervisé pour l'acquisition automatique des noms d'artefacts, la proportion du type d'erreur causée par les prédicats non appropriés et celle du type d'erreur causée par l'extension des termes composés (y compris les groupes nominaux mal reconnus comme termes composés, les termes composés reconnus à tort à cause de l'identification erronée des constituants et les termes composés formés à partir des termes d'autres classes sémantiques) sont respectivement de 33.33% et de 33.34% pour la classe sémantique des contenants d'arrangement. Ils sont respectivement de 28.57% et de 42.85% pour la classe sémantique des appareils de cuisson. Pour la classe de moyens de transport, ils sont 58.33% de et de 25.00%.

Parmi les erreurs causées par l'extension de termes composés, les trois types d'erreurs (les groupes nominaux, les termes composés avec des constituants reconnus à tort et les termes composés formés d'autres classes sémantiques) se produisent dans environ les mêmes proportions. Cette comparaison prouve que la délimitation entre les noms de métiers composés et les groups nominaux dont les noyaux sont les noms de métiers est plus difficile à faire par rapport à celle entre les noms d'artefacts et les groupes nominaux dont les noyaux sont les noms d'artefacts, puisque la plupart des noms de métiers composés est formée à partir des noms de métiers monolexicaux en ajoutant les expansions pour spécifier l'activité professionnelle.

De plus, pour l'acquisition automatique des noms d'artefacts fondée sur l'apprentissage supervisé, l'intersection (70.97% de proportion) est la raison principale des silences. Pour l'acquisition automatique des noms d'artefacts basée sur l'apprentissage semi-supervisé, le silence est principalement causé du fait que certains arguments se trouvent souvent dans les structures prédicat-argument pour la classe sémantique des contenants

d'arrangement (50.00%) et la classe sémantique des appareils de cuisson (75.00%), alors que la plupart des silences sont causés par l'intersection pour la classe sémantique des moyens de transport (90.91%). Pour l'acquisition automatique des noms de métiers basée sur l'apprentissage supervisé, les silences causés par l'intersection occupent la proportion la plus importante : 46.87%. Pour l'acquisition automatique des noms de métiers basée sur l'apprentissage semi-supervisé, il y a deux principales raisons : l'intersection des arguments et le fait que de nombreux arguments se trouvent souvent dans les structures non prédicat-argument. Les silences causés par ces deux raisons occupent respectivement 42.11% et 36.84%.

Finalement, on constate que l'amélioration du résultat en intégrant l'extension des termes composés est moins grande pour la classe sémantique des moyens de transport du fait que la classe sémantique des moyens de transport possède moins de noms composés par rapport à la classe sémantique des contenants d'arrangement et à la classe sémantique des appareils de cuisson. Enfin, il faudrait souligner que la quantité des prédicats appropriés définitionnels de la classe sémantique des moyens de transport est moins grande que celle de la classe sémantique des contenants d'arrangement et que celle de la classe sémantique des appareils de cuisson. Cela rend la méthode distributionnelle moins performante pour la classe sémantique des moyens de transport. Dans le Tableau 43, on liste les types d'erreurs et les types de silences principaux de chaque méthode pour une comparaison.

Méthodes			**Types d'erreurs**	**Types de silences**
Méthode distributionnelle (NAF)	**Supervisée**		1. utilisation des prédicats qui ne sont pas les prédicats appropriés pour augmenter le taux de rappel du résultat (31.25%) 2. constituants syntaxiques reconnus à tort (25.00%)	1. présence fréquente dans les structures non prédicat-argument 2. coexistence fréquente avec les prédicats basiques (avoir, prendre, etc.) 3. absence de patrons syntaxiques
	Semi-supervisée	**Moyens de transport**	1. récupération des prédicats non appropriés à partir des arguments 2. constituants syntaxiques reconnus à tort en raison du calcul des patrons syntaxiques	
		Appareils de cuisson		
		Contenants d'arrangement		
Méthode morphosémantique (NMP)			1. extension des termes composés (constituants reconnus à tort, groupes nominaux reconnus comme termes composés)	1. suppression par le calcul de l'information mutuelle 2. complétude des listes de termes données à l'avance
Méthode combinatoire				
NAF	**Supervisée**		1. utilisation des prédicats qui ne sont pas les prédicats appropriés pour augmenter le taux de rappel du résultat (28.57%) 2. les groupes nominaux reconnus à tort comme termes composés (35.71%)	1. suppression par l'intersection (70.97%)
	Semi-supervisée	**Moyens de transport**	1. récupération des prédicats non appropriés (58.33%) 2. extension des termes composés (25.00%)	1. suppression par l'intersection (90.91%)
		Appareils de cuisson	1. récupération des prédicats non appropriés (28.57%) 2. extension des termes composés (42.85%)	1. présence fréquente dans les structures non prédicat-argument (75.00%)
		Contenants d'arrangement	1. récupération des prédicats non appropriés (33.33%) 2. extension des termes composés (33.34%)	1. présence fréquente dans les structures non prédicat-argument (50.00%)
NMP	**Supervisée**		1. groupes nominaux reconnus à tort (47.37%)	1. intersection (46.87%)
	Semi-supervisée		1. groupes nominaux reconnus à tort (40.74%)	1. intersection (42.11%) 2. présence fréquente dans les structures non prédicat-argument (36.84%)

Tableau 43 Comparaison des types d'erreurs et de silence des trois méthodes

Chapitre 2 Analyses

Dans ce chapitre, on présente respectivement les analyses à propos des structures syntaxiques et lexicales, les classes sémantiques ainsi que les langues de spécialité dans le traitement automatique des langues. On tente d'expliquer plus profondément le mécanisme des méthodes développées dans la thèse et les résultats obtenus par ces méthodes. Premièrement, on explique la représentation des structures syntaxiques et celle des structures lexicales présentée dans la thèse. Ensuite, on présente une analyse des classes sémantiques dans le traitement automatique des langues. Finalement, nous fournissons une analyse sur les classes sémantiques appliquée aux langues de spécialité.

1. Représentation des structures syntaxiques et des structures lexicales

1.1. Représentation des structures syntaxiques

Les phrases peuvent être considérées comme la production d'un nombre limité d'opérations à partir d'un nombre limité de phrases de base. Par exemple,

67) *La négociation n'a abouti à aucun résultat. -> La négociation (<- +Modifieur, Modifieur=Proposition relative) n'a abouti à aucun résultat. -> La négociation qui a eu lieu à Luxembourg n'a abouti à aucun résultat.*

68) *Le chef de cuisine a coupé le poulet en morceaux. -> Le chef de cuisine a coupé le poulet en morceaux.* (<- Transformation au passif) -> *Le poulet a été coupé en morceaux par le chef de cuisine.*

69) *Il faut éteindre la lumière avant de partir. -> Il faut éteindre la lumière* (<- Pronominalisation) *avant de partir. -> Il faut l'éteindre avant de partir.*

Si l'on exécute successivement l'opération β_1 (β_1= ←de(Dét)+N, ←= ajouter les constituants

suivants à gauche) au sujet W (W=*le rapport*) de la phrase Z (Z= *Le rapport a été distribué*), l'opération β_2 (β_2= ←apposition) au sujet W et l'opération β_3 (β_3= ←C.O.I.) au prédicat P (P=*distribuer*) de la phrase Z, on peut obtenir une série de phrases suivantes :

70) *Le rapport a été distribué. -> le rapport de la conférence a été distribué. -> Le rapport de la conférence a été distribué aux participants. -> Le rapport de la conférence, imprimé hier soir, a été distribué aux participants.*

Pour relier les constructions de ces phrases, on utilise la règle :(Z, W←β_1, P) -> (Z', W'←β_2, P,) -> (Z'', W', P←β_3) -> (Z''', W', P') permettant de passer d'une structure à une autre.

Néanmoins, toutes les opérations ne peuvent se combiner pour produire les phrases, par exemple,

71) *Marc a cassé la table* (<- Pronominalisation). -> *Marc l'a cassée. -> Marc l'a cassée* (<- +Modifieur, Modifieur=Proposition relative) -> ***** *Marc l'a cassée dont il a achetée l'année dernière.*

Une même opération ne peut non plus être appliquée à toutes les phrases, par exemple,

72) *J'ai récuré* (<-+C.M., C.M.=avec+NAF) *mon lavabo. -> *J'ai récuré mon lavabo avec le récurant.*

73) *Le secrétaire a pris contact avec le nouveau stagiaire.* (<-Transformation en passif) -> ***** *Le nouveau stagiaire a été pris en contact par le secrétaire.*

L'étiquetage des structures prédicat-argument présenté dans la thèse est réalisé en se fondant sur les patrons syntaxiques. Dans la construction des patrons syntaxiques, on a essayé de représenter les structures syntaxiques les plus susceptibles de subir des traitements automatiques. Par exemple, quatre patrons syntaxiques de base sont établis pour représenter les distributions syntactico-sémantiques des prédicats de noms d'artefacts : V+NAF, V+NAF+prep+NAF, V+Nc+prep+NAF et V+prep+NAF. Ensuite, nous avons décrit les observations faites après application des règles suivantes :

74) V+NAF *(éteindre l'ordinateur)*->V+ADV+NAF (*éteindre soudain l'ordinateur*)

V+NAF (*éteindre l'ordinateur*)->NAF+être(se faire)+Vpassif (*l'ordinateur est éteint*)

V+NAF->NAF+Vpp (*l'ordinateur éteint*)

V+NAF+prep+NAF (*nettoyer la casserole avec la liquide vaisselle*)->PRON+V+prep+NAF (*les nettoyer avec la liquide vaisselle*)

… …

Les patrons syntaxiques qui se trouvent à droite de la flèche sont appelés les patrons syntaxiques dérivés ; ils illustrent la relation entre les constructions qu'ils représentent et les constructions représentées par les patrons syntaxiques de base. Les patrons syntaxiques dérivés peuvent être considérés comme la production d'un nombre limité d'opérations à partir d'un nombre limité de patrons syntaxiques de base.

1.2. Représentation des structures lexicales

Nous avons également essayé de représenter les structures internes des unités lexicales en nous fondant sur les observations de milliers de noms (les noms de métiers et les noms d'artefacts). L'objectif est d'établir un ensemble de patrons morphosyntaxiques permettant de représenter le plus complètement possible les structures lexicales dans le but d'un traitement automatique.

Les unités polylexicales sont considérées comme la production d'un nombre limité d'opérations à partir d'une unité monolexicale. Pour les noms de métiers et les noms d'artefacts, nous avons établi respectivement un ensemble d'opérateurs (tels que ←N/GN, ←de+N/GN/NAF, ←NMP, etc.) qui indiquent les opérations à effectuer pour passer d'une unité monolexicale à une unité polylexicale. Ces opérations désignent les éléments constitutifs à ajouter pour représenter la structure d'une unité polylexicale à partir d'une unité monolexicale. Par exemple,

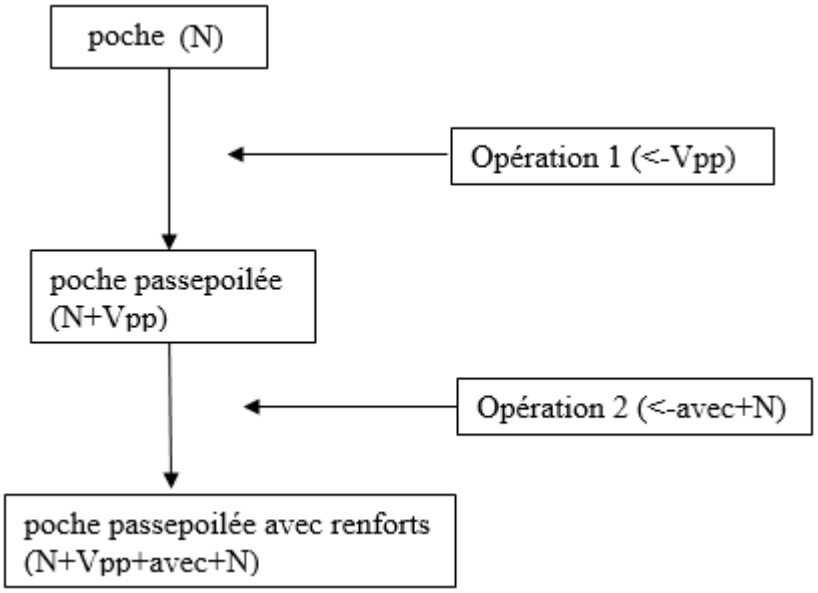

Figure 47 Représentation de la structure interne du nom composé *poche passepoilée avec renforts*

Cette représentation lie la structure d'une unité monolexicale (*poche*) à celle d'un nom composé de base (*poche passepoilée*) et lie la structure du nom composé de base (*poche passepoilée*) à celle d'un nom composé complexe (*poche passepoilée avec renforts*).

Cependant, toutes les opérations ne peuvent être appliquées à toutes les unités lexicales, par exemple, pour l'opérateur ←et/ou+A,

75) *technicien* +et+A->**technicien et électronique*
76) *technicien mécanique*+et+A ->*technicien mécanique et électronique*

De plus, toutes les opérations ne sont pas combinables. Par exemple, l'opérateur ←de(Dét)+N ne peut pas se combiner avec l'opérateur ←et/ou+A, mais l'opérateur ←A est combinable avec ←et/ou+A,

77) *technicien+de(Dét)+N -> technicien du bâtiment ->technicien du bâtiment+et+A->*technicien du bâtiment et électronique*
78) *technicien+A->technicien mécanique->technicien mécanique+et+A->technicien mécanique et électronique*

Nous avons présenté avec plus de détails une analyse sur la possibilité de combinaisons entre les différentes opérations (cf., Tableau 24, p. 211) pour les unités polylexicales de noms de

métiers.

Finalement, dans la plupart des cas, la relation sémantique entre le nom de métier composé (ou le nom d'artefact composé) et le nom de métier (ou le nom d'artefact) monolexical contenu dans sa structure interne peut être représentée par un processus de spécification par les fonctions professionnelles (les propriétés ou les fonctions d'un artefact). Par exemple,

79) chimiste-> bio-chimiste (spécifie la fonction professionnelle)

80) technicien -> technicien du bâtiment -> technicien du bâtiment en énergies renouvelables (spécifie la fonction professionnelle)

81) bonde -> bonde pare bruit (*pare bruit* spécifie la fonction de bonde)

82) appareil -> appareil à grand capteur (*à grand capteur* indique la propriété d'appareil)

Cette analyse concernant la relation sémantique interne permet de mieux comprendre la formation et la composition des unités polylexicales étudiées. Cela est indispensable pour les analyses des structures internes des unités lexicales.

2. Classes sémantiques dans le traitement automatique des langues

La recherche d'un nombre minimum de prédicats permettant de définir une classe sémantique est distincte du recensement de l'ensemble des prédicats appartenant en propre à cette classe sémantique (à savoir les prédicats appropriés de cette classe sémantique). Dans le premier cas, chaque prédicat n'est pas obligatoirement un prédicat approprié de la classe sémantique, alors que chaque prédicat doit être le prédicat approprié de la classe sémantique dans le second cas. (Gross G., 2012 :75-76) Par exemple, pour la classe sémantique « MOYEN_DE_TRANSPORT », *conduire* et *réparer* suffisent à circonscrire la classe sémantique « MOYEN_DE_TRANSPORT » et ils constituent un nombre minimum de prédicats permettant de définir la classe sémantique donnée, alors que *réparer* n'est pas le

prédicat approprié de la classe sémantique « MOYEN_ DE_TRANSPORT », car on peut dire non seulement réparer + « VEHICULE » mais aussi réparer + « ELECTROMENAGER », réparer + « CHAUSSURE », etc. Néanmoins, l'ensemble des prédicats appropriés de la classe sémantique « MOYEN_DE_TRANSPORT » comprend les prédicats : *rouler, conduire, garer*, etc. qui appartiennent tous en propre à la classe sémantique. Dans le traitement automatique des langues, c'est l'ensemble des prédicats appropriés de la classe sémantique qu'on doit recenser pour identifier les entités désirées. Cet ensemble de prédicats appropriés constituent les prédicats appropriés définitionnels de la classe sémantique.

Dans un corpus collecté, il est difficile de garantir que tous les arguments d'une classe sémantique apparaissent et que tous les arguments de cette classe sémantique ne coexistent qu'avec les prédicats appropriés définitionnels de cette classe sémantique. Certains arguments de la classe sémantique donnée ne coexistent qu'avec certains prédicats appropriés définitionnels. Certains coexistent à la fois avec certains prédicats basiques (ou certains prédicats appropriés d'une classe sémantique plus générale) et avec certains prédicats appropriés définitionnels de la classe sémantique donnée. Certains ne coexistent qu'avec les prédicats basiques ou les prédicats appropriés de la classe sémantique plus générale. Les arguments qui coexistent avec tous les prédicats appropriés définitionnels de la classe sémantique dans un corpus sont souvent très peu nombreux. Ainsi, dans le processus d'intersection, on a défini la fréquence d'intersection pour sélectionner non seulement les arguments en commun de l'ensemble des prédicats appropriés définitionnels mais aussi ceux qui sont très partagés par les prédicats dans le but d'augmenter le rappel du résultat. Cela conduit au fait que la précision du résultat baisse avec l'abaissement du seuil décidé pour la fréquence d'intersection et augmente avec l'augmentation du seuil de la fréquence d'intersection. De plus, dans la méthode distributionnelle supervisée, certains prédicats qui ne sont pas les prédicats appropriés de la classe sémantique donnée mais qui contribuent également à définir une classe sémantique sont aussi pris en compte afin d'augmenter le rappel du résultat.

Cependant, recenser tous les prédicats appropriés d'une classe sémantique nécessite beaucoup de temps et d'efforts. Souvent, il est difficile d'obtenir une liste exhaustive de

prédicats appropriés d'une classe sémantique donnée. Ainsi, nous avons développé la méthode semi-supervisée pour repérer plus de prédicats appropriés à partir d'un ensemble d'arguments donné à l'avance. Néanmoins, les prédicats basiques ont un spectre sémantique général et ils ont souvent les classes sémantiques des prédicats appropriés. En conséquence, la distribution d'arguments d'une classe sémantique donnée ne se limite pas dans les contextes des prédicats appropriés de cette classe sémantique mais aussi dans les contextes des prédicats basiques. C'est la raison pour laquelle l'identification des prédicats appropriés à partir des arguments nous amène de nombreux prédicats basiques. Dans la méthode semi-supervisée, l'étape d'élimination des prédicats basiques ne permet pas d'éliminer tous les prédicats basiques reconnus et les prédicats basiques qui restent amènent très probablement les arguments d'autres classes sémantiques. Cela est la raison principale de l'abaissement de la précision du résultat de la méthode semi-supervisée par rapport à la méthode supervisée.

Le nombre des prédicats appropriés définitionnels des différentes classes sémantiques n'est pas pareil dans un corpus collecté. Les prédicats appropriés de certaines classes sémantiques sont en général en grande quantité et les prédicats appropriés de certaines classes sont relativement moins nombreux. Dans la méthode semi-supervisée, si les prédicats appropriés d'une classe sémantique sont très peu nombreux, on risque de récupérer plus de prédicats basiques à partir des arguments fournis à l'avance et de baisser la pertinence du résultat, puisque la distribution d'arguments d'une classe sémantique ne se limite aux contextes des prédicats appropriés de cette classe sémantique mais aussi aux contextes des prédicats basiques. Cela permet d'expliquer pourquoi la méthode distributionnelle semi-supervisée est moins performante pour la classe sémantique « MOYEN_DE_TRANSPORT » dont les prédicats appropriés sont moins nombreux par rapport à ceux des deux autres classes sémantiques « APPAREIL_DE_CUISSON » et « CONTENANT_DE_ARRANGEMENT ».

3. Classes sémantiques dans les langues de spécialité

La langue de spécialité est un sous-langage d'un domaine, d'un thème ou d'une situation de communication qui se caractérise par un lexique limité et par l'existence de schémas de phrases en nombre restreint. Le corpus spécialisé est à l'origine des langues de spécialité.

D'après Harris (2007 : 37-38), la contrainte de la combinaison entre les mots s'explique par la probabilité d'occurrence des mots. Certains mots ont une probabilité d'occurrence exceptionnellement élevée ou un statut spécial dans une position donnée.

En réalité, dans les différents corpus spécialisés, la probabilité d'occurrence de certains mots dans une position donnée peut être changée. Par exemple, le prédicat *acheter* a un spectre sémantique général et presque tous les types d'arguments peuvent être combinés avec ce prédicat : les arguments d' « ALIMENT » (*pain, vin, fromage, foie gras*, …), les arguments de « VETEMENT » (*robe, chemise, pantalon, linge*, ….), les arguments de « MOYEN_DE_TRANSPORT » (*tracteur, voiture, tramway, train*, …), etc., La probabilité d'occurrence des arguments de « MOYEN_DE_TRANSPORT » apparaît relativement plus élevée en position de complément du prédicat *acheter* par rapport aux autres classes sémantiques si le corpus est constitué des textes de la thématique moyens de transport, alors que la probabilité d'occurrence des arguments de « MOYEN_DE_TRANSPORT » devient moins probante si le corpus est de la thématique mode ou gastronomie. Le corpus spécialisé privilégie certaines classes sémantiques des prédicats ayant un spectre sémantique général.

Les trois méthodes présentées dans la thèse sont appliquées au corpus spécialisé. Pour la méthode morphosémantique, le corpus spécialisé permet de fournir une quantité suffisante de termes d'informations morphologiques. Pour la méthode distributionnelle, le corpus spécialisé permet de limiter la probabilité d'occurrence des mots de certaines classes sémantiques dans des positions données et cela permet d'augmenter la précision de l'identification des arguments de la classe sémantique donnée. Pour la méthode combinatoire qui associe les deux premières approches, le corpus spécialisé favorise la mise en valeur des deux approches à la fois. Dans un corpus spécialisé, les unités lexicales concernant le thème ou le domaine du corpus présentent une grande fréquence et la classe sémantique d'arguments concernant le thème ou le domaine a une grande probabilité d'occurrence dans une position donnée. Néanmoins, cela ne signifie pas qu'il n'existe pas de langue générale dans les corpus spécialisés.

Chapitre 3 Résultats obtenus et leur utilisation

Les résultats obtenus ne sont pas des mots en vrac. Ils consistent en des listes de termes regroupés par classes sémantiques. Les termes sont extraits avec leurs contextes qui sont les prédicats coexistant avec eux dans le corpus spécialisé. Les résultats sont plutôt les ressources lexicales structurées sémantiquement. L'acquisition automatique des termes pourrait par la suite être appliquée aux autres types de vocabulaires. Les résultats obtenus pourraient contribuer à l'élaboration de relevés de toutes espèces : dictionnaires, glossaires, lexiques spécialisés, etc. Ils pourraient également être très intéressants pour tous les travaux de traduction humaine, assistée ou automatique. Les autres applications du traitement automatique des langues pourraient aussi être favorisées par l'exploitation de ces résultats. Dans ce qui suit, on présente en détail le type de résultats obtenus et leur utilisation.

1. Type de résultats obtenus

1.1. Termes regroupés par classes sémantiques

Les résultats obtenus sont les listes de termes regroupés par classes sémantiques. Ils comprennent les listes de termes de différentes classes sémantiques, telles que la classe sémantique des moyens de transport, la classe sémantique des appareils de cuisson et la classe sémantique des arrangements de contenant. Ces classes sémantiques constituent des sous-ensembles de l'hyperclasse : noms d'artefacts. De plus, les résultats obtenus contiennent également une liste de termes de noms de métiers et une liste de termes de l'hyperclasse noms d'artefacts dans laquelle les différentes classes sémantiques ne sont pas distinguées. Ces résultats sont les ressources lexicales qui comprennent à la fois les unités monolexicales et les unités polylexicales. Ces ressources lexicales sont structurées par divisions sémantiques.

1.2. Présentation des termes dans les résultats

Dans la méthode distributionnelle, les termes extraits sont présentés dans un ordre décroissant selon la probabilité les rattachant à la classe sémantique indiquée. Plus un terme est partagé par les prédicats d'une classe sémantique, plus il appartient probablement à cette classe sémantique. Notre algorithme consiste à proposer les termes les plus susceptibles d'être de la classe sémantique donnée et présente les termes dans l'ordre décroissant selon le nombre de prédicats qui les partagent (noté comme la fréquence d'intersection). À la fin de la liste de termes, on trouve souvent les noms d'autres types. Cependant, dans le résultat de la méthode combinatoire, de nombreux noms composés de la classe sémantique donnée se trouvent également à la fin de la liste du fait que leurs occurrences sont relativement moins élevées par rapport aux termes monolexicaux. Nous avons défini un seuil pour préciser la fréquence d'intersection dans le but de sélectionner directement les résultats les plus pertinents, mais il existe quand même des termes à récupérer parmi les unités qui n'ont pas été sélectionnées.

Le résultat de la méthode morphosémantique est composé de deux parties : une liste de termes enrichie par les opérations morphologiques et une liste de noms composés associant à chacun une valeur d'information mutuelle. Les noms composés sont présentés dans l'ordre décroissant selon leurs valeurs d'information mutuelle. Plus la valeur d'information mutuelle est élevée, plus le candidat est probablement un nom composé. Nous avons également défini un seuil pour la valeur d'information mutuelle afin de sélectionner les collocations.

Le résultat de la méthode combinatoire est une combinaison des deux premiers résultats présentés. Chaque liste de termes présentés dans l'ordre en fonction de la probabilité les rattachant à la classe sémantique indiquée est accompagnée d'une liste de termes enrichie par les opérations morphologiques et une liste de noms composés avec leurs valeurs d'information mutuelle.

1.3. Termes extraits avec les contextes

Les termes sont extraits avec leurs contextes. Ces contextes sont plutôt les prédicats qui coexistent avec les termes identifiés. Nous avons étiqueté les prédicats appropriés à partir d'un ensemble d'arguments fourni à l'avance. Ces prédicats appropriés récupérés font partie de la classe sémantique des arguments donnés. Cela permet de prédire quels substantifs (i.e. les substantifs de certaines classes sémantiques) sont possibles en position de complément pour engendrer des phrases possibles construites autour de ces prédicats. Par exemple, les prédicats de la classe sémantique de moyen de transport *conduire*, *garer*, etc. ont en général les arguments de moyen de transport (tels que, *voiture, bus, Citroën*, *camping-car*, etc.) en position de compléments au lieu des arguments d'autres types (tels que *cul de poule, cuiseur*, *mijoteuse*, etc.).

Les résultats intermédiaires obtenus par l'intersection des prédicats et l'élimination des prédicats basiques comprennent les listes de prédicats appropriés des différentes classes sémantiques extraits à partir des arguments donnés. Ces résultats intermédiaires pourraient être les ressources lexicales importantes pour la génération automatique des langues naturelles, car la syntaxe seule ne permet pas de générer les phrases toujours acceptables. Or, le regroupement des prédicats par classes sémantiques permet de fournir les informations sémantiques nécessaires pour la génération des phrases.

2. Utilisation des résultats

2.1. Élaboration des ressources lexicales

Une contribution importante de l'ère actuelle est celle de fournir à tous l'accès à une information à l'échelle mondiale de plus en plus rapide avec l'évolution du web. Il existe un énorme dépôt de documents sous forme électronique. L'exploitation de ces milliards de données n'a pas commencé d'hier. Des groupes de recherche spécialisés ont déjà travaillé au début de l'ère électronique. Néanmoins, jusqu'à maintenant, les ressources lexicales des unités spécifiques sont encore très peu exploitées. Certains types de vocabulaires ne sont pas suffisamment traités, tels que les noms d'artefacts, les noms de métiers, etc. De plus, ces ressources lexicales sont en général moins informatisées pour le traitement automatique des

langues. La pertinence de notre recherche dans cette thèse est d'avoir déterminé et documenté une façon originale d'effectuer la saisie de milliers de données triées sous l'étiquette « langues de spécialité ».

Une des activités humaines associée à la collection des données consiste à créer des dictionnaires, des glossaires, etc. À vue d'œil, les demandeurs et utilisateurs de ces produits pourraient être des industries spécialisées, des groupes de recherche pointue, des agglomérats pharmaceutiques, des institutions documentaires ou culturelles, des organes de diffusion journalistique et télévisuelle, etc. L'extraction des unités lexicales spécifiques pourrait par la suite être effectuée en les divisant selon des activités humaines, par exemple, disons d'une manière non détaillée, la mécanique des sols, les métiers d'art, la physique nucléaire, les travaux de forge, la reliure, etc. Nous pouvons appliquer notre méthode aux autres types de vocabulaires et récupérer les classes sémantiques de termes aussi diversifiées que possible. Ces vocabulaires peuvent être les lexiques de différents domaines, les lexiques demandés dans les buts spécialisés ou les lexiques des utilisateurs de différents métiers.

Les résultats obtenus contribuent à l'élaboration des ressources lexicales. Un champ d'exploitation privilégié pourrait être celui de la mise en publication de toutes espèces, dictionnaires, glossaires, vocabulaires et lexiques spécialisés, etc. Toutes ces ressources lexicales peuvent également être exploitées pour les travaux de traduction (humaine, assistée par la machine ou automatique) ou pour les autres applications du traitement automatique des langues.

2.2. Traduction humaine, assistée ou automatique

Les ressources lexicales spécialisées sont indispensables dans tous les types de traductions de la langue de spécialité. Les collections de termes résultant des saisies électroniques de données pourraient faciliter davantage les travaux de traduction humaine, assistée ou automatique dans les langues de spécialité. On pourrait citer, par exemple, les traductions des noms de produits et celles des noms de métiers ou de profession. Notamment, les résultats obtenus avec les contextes sémantiques permettent de prendre en compte la

polysémie dans la traduction assistée ou automatique. Notre recherche sur l'acquisition automatique des unités spécifiques pourra aussi être effectuée sur les vocabulaires des autres langues pour l'élaboration des ressources lexicales multilingues.

2.3. Autres applications du traitement automatique des langues

Les termes collectionnés seraient également signifiants pour les autres applications du traitement automatique des langues, par exemple, dans un système de dialogue homme-machine, l'intégration des listes de termes regroupés par classes sémantiques permet une analyse plus pertinente des textes en langue naturelle. Par rapport à une application plus près du dialogue homme-machine, la génération des langues naturelles, toutes les hypothèses peuvent être envisagées. Ces activités sont favorisées du fait que les unités lexicales ciblées sont regroupées par domaines, par activités humaines et par divisions sémantiques.

Conclusion

Dans cette thèse, nous avons étudié la fonction argumentale dans le but de proposer une méthode pertinente pour l'acquisition automatique des termes. Trois méthodes ont été développées en se fondant sur les caractéristiques morphologiques des unités lexicales et sur la relation d'appropriation entre les prédicats appropriés et leurs arguments. Nous avons commencé par exploiter les structures prédicat-argument pour repérer les arguments d'une classe sémantique (la méthode distributionnelle). Les structures internes des unités lexicales ont ensuite été analysées dans le but d'enrichir la liste des termes (la méthode morphosémantique). Les deux premières approches ont été enfin combinées afin d'améliorer la pertinence des résultats (la méthode combinatoire). En même temps, nous avons également évalué le travail de profilage du corpus et celui de la construction du corpus web pour le traitement automatique des langues.

La méthode distributionnelle est mise en place au moyen de deux stratégies : la méthode distributionnelle supervisée et la méthode distributionnelle semi-supervisée. La méthode distributionnelle a été appliquée au vocabulaire des noms d'artefacts. La méthode distributionnelle supervisée consiste à reconnaître les noms d'artefacts en repérant les structures prédicat-argument à partir des prédicats donnés à l'avance à l'aide des patrons syntaxiques. La méthode distributionnelle semi-supervisée a pour objet spécifiquement d'identifier les prédicats appropriés en repérant les structures prédicat-argument à partir d'une classe sémantique d'arguments fournie à l'avance. Et puis, les patrons syntaxiques des prédicats reconnus sont appris automatiquement par un calcul probabiliste pour obtenir plus de noms d'artefacts. Par rapport à la méthode distributionnelle supervisée, la méthode distributionnelle semi-supervisée permet d'augmenter le rappel du résultat. Néanmoins, la précision de cette dernière méthode est moins élevée en raison de la récupération des prédicats basiques et de l'impossibilité de prévoir la distribution syntactico-sémantique des prédicats à partir des arguments. La méthode distributionnelle permet d'étiqueter du

vocabulaire en repérant la distribution d'arguments et cela permet de prendre en compte les éléments polysémiques, les néologismes, les fautes d'orthographe et les abréviations. Cependant, pour les arguments qui ne se trouvent pas fréquemment dans les structures prédicat-argument ou qui coexistent souvent avec les prédicats basiques dans un corpus collecté, la méthode distributionnelle apparaît moins efficace pour les récupérer. Enfin, il faudrait souligner qu'il manque de mécanismes pour l'identification des unités polylexicales dans la méthode distributionnelle.

La méthode morphosémantique consiste à enrichir le vocabulaire en exploitant les structures internes des unités lexicales. La méthode morphosémantique est expérimentée à partir du vocabulaire des noms de métiers. On fournit à l'avance une liste de noms de métiers simples et une liste de bases à partir desquelles les noms de métiers dérivés peuvent être obtenus. Ensuite, la segmentation morphématique et la recomposition des morphèmes sont effectuées pour obtenir de nouvelles bases potentielles. Et puis, la dérivation morphologique est appliquée à toutes les bases pour obtenir les noms de métiers dérivés. Finalement, les patrons morphosyntaxiques sont construits en s'appuyant sur les analyses des structures internes des unités polylexicales dans le but d'étiqueter les noms composés à partir des noms monolexicaux. La précision du résultat obtenue par la méthode morphosémantique est assez élevée, mais le rappel est dépendant de la quantité de termes fournis préalablement. De plus, la méthode morphosémantique se fonde sur les analyses morphologiques des unités lexicales et elle devient moins performante si les vocabulaires à traiter sont pauvres du point de vue des caractéristiques morphologiques.

La méthode combinatoire associe la méthode distributionnelle et la méthode morphosémantique. Elle consiste à mettre en valeur les deux méthodes afin d'augmenter la pertinence du résultat. La méthode combinatoire est appliquée aux vocabulaires des noms d'artefacts et des noms de métiers. Pour les deux vocabulaires, les patrons morphosyntaxiques établis pour l'identification des unités polylexicales sont intégrés à la position des arguments dans les graphes de structures prédicat-argument. Cela permet de compléter un mécanisme d'identification des unités polylexicales à la méthode distributionnelle. Ensuite, les structures

prédicat-argument extraites par la méthode distributionnelle sont exploitées pour obtenir une liste de bases à partir desquelles la liste de termes reconnus peut être enrichie par les opérations morphologiques. Finalement, à partir des unités monolexicales obtenues, plus de noms composés peuvent être repérés à l'aide des patrons morphosyntaxiques. D'un côté, la méthode distributionnelle permet de fournir une liste de termes à enrichir pour la méthode morphosémantique. D'un autre côté, la méthode morphosémantique complète la méthode distributionnelle par un mécanisme d'identification des unités polylexicales et surmonte la limite de la méthode distributionnelle sur le plan de la reconnaissance des termes qui ne se trouvent pas fréquemment dans les structures prédicat-argument. Par rapport à la méthode distributionnelle et à la méthode morphosémantique, la méthode combinatoire permet d'obtenir un résultat plus pertinent.

A travers l'analyse des résultats et leur comparaison, on a pu effectuer une série d'analyses sur la représentation des structures syntaxiques et celle des structures lexicales, sur les classes sémantiques et les langues de spécialité dans la perspective du traitement automatique des langues. Nous avons tenté d'expliquer quelle est l'influence du corpus spécialisé sur l'occurrence des phénomènes langagiers ; c'est pourquoi nous adoptons l'opération d'intersection et le calcul d'élimination des prédicats basiques amenés à partir des arguments donnés à l'avance dans la méthode distributionnelle semi-supervisée. Bien que les patrons syntaxiques et les patrons morphosyntaxiques établis ne soient pas exhaustifs, cette méthode permet quand même de représenter la plupart des structures syntaxiques et lexicales dans le cadre du projet de cette thèse.

Les trois méthodes d'acquisition automatique des termes peuvent être appliquées aux autres types de vocabulaires. Cela permet une grande ouverture à l'élaboration des ressources lexicales, notamment des ressources lexicales spécialisées. Les résultats regroupés par divisions sémantiques favorisent tous les travaux de traduction (humaine, assistée ou automatique) et aux autres applications du traitement automatique des langues, telles que la recherche dans le cadre du dialogue homme-machine, la génération automatique des langues naturelles, etc.

Bibliographie

Abney S. P., 1991, “Parsing by chunks”, in Berwick, R., Abney S., Tenny C. (eds) *Principle-based parsing*, Dordrecht, Kluwer Academic Publishers, p. 257-278.

Aggarwal C. C. et Zhai C. X., 2012, “A Survey of Text Clustering”, in C. C. Aggarwal, C. Zhai (eds), *Mining Text Data*, US, Springer, p. 77-131.

Apothéloz D., 2002, *La construction du lexique français, Principes de morphologie dérivationnelle*, Paris, Ophrys.

Baroni M., Bernardini S., Ferraresi A. et Zanchetta E., 2009, “The WaCky Wide Web: A Collection of Very Large Linguistically Processed Web-Crawled Corpora”, *Language Resources and Evaluation*, Vol 43(3), Pays-Bas, Kluwer, p. 209-226.

Baroni M., Chantree F., Kilgarriff A. et Sharoff S., 2008, “Cleaneval: a Competition for Cleaning Web Pages”, in *LREC'08*, European Language Resources Association (ELRA).

Benson M., Benson, E. et Ilson, R., 1986, *The BBI Combinatory Dictionary of English: A Guide to Word Combinations*, Pays-Bas, John Benjaminsn.

Benveniste E., 1967, « Fondements syntaxiques de la composition nominale », *Bulletin de la Société de Linguistique de Paris*, LXII/1, Paris, Klincksieck, p. 15-31.

Benveniste E., 1974, *Problèmes de linguistique générale II*, Paris, Gallimard.

Bescherelle L.-N. et Bescherelle H., 1852, *Le Véritable Manuel des conjugaisons, ou la Science des conjugaisons mise à la portée de tout le monde*, Paris, Dépôt Central des Publications Classiques.

Biber D., 1994, “Representativeness in corpus design”, in *Linguistica Computazionale*, IX-X, p. 377–408.

Biskri I., Meunier J.-G. et Joyal S., 2004, « L'extraction des termes complexes : une approche modulaire semi-automatique », dans *7es Journées internationales d'Analyse statistique des Données Textuelle*, France, p.192-201.

Boulaknadel S., Daille B. et Aboutajdine D., *Acabit : un outil d'extraction des termes complexes*, « http://tal.ircam.ma/conference/docs/ticam2008/6_boulaknadel.pdf », visité le 9 juillet 2013.

Boulbry G., Kercadio Y. D., 2004, « De l'analyse lexicale à la construction d'Echelles psychométriques : Application à la mesure du tempérament nostalgique », dans P. Robert-Demontrond (Ed.), *L'analyse de discours*, France, p.154-179.

Bourigault D., Gonzalez-Mullier I. et Gros C., 1996, " LEXTER : a Natural Language Tool for Terminology Extraction ", in *Proceedings of the 7th EURALEX International Congress*, Göteborg, p. 771-779.

Buvet P.-A., 2009a, « Des mots aux emplois : la représentation lexicographique des prédicats », Le Français Moderne, vol. 77, n° 1, pp. 83-96.

Buvet P.-A., 2009b, « Les dictionnaires électroniques du modèle des classes d'objets », *Langages*, Larousse, p.63-79.

Cabré M. T., 2007, « Constituer un corpus de textes de spécialité », *Cahier du CIEL 2007-2008*, p. 37-56.

Cavnar W.B. et Trenkle J. M., 1994, "N-gram-based Text Categorization". In *Proceedings of SDAIR-94*, Las Vegas, Nevada, U.S.A., UNLV Publications/Reprographics, p. 161-171.

Charaudeau P., 2009, « Dis-moi quel est ton corpus, je te dirai quelle est ta problématique », *Revue Corpus n°8*, Nice, p. 37-66.

Cho J., et Garcia-Molina H., 1999, "The Evolution of the Web and Implications for an Incremental Crawler", in *Proceeding VLDB '00 Proceedings of the 26th International Conference on Very Large Data Bases*, Stanford, p. 200-209.

Chomsky N., 1969, *Structures Syntaxiques*, Paris, Seuil.

Church K. W. et Hanks P., 1990, "Word Association Norms, Mutual Information, and Lexicography", in *Computational Linguistics*, 16(1), p. 22–29.

Condamines A., 2005, *Sémantique et corpus*, Paris, Lavoisier.

Constant M., Sigogne A. et Watrin P., 2012, « La reconnaissance des mots composés à l'épreuve de l'analyse syntaxique et vice-versa : évaluation de deux stratégies discriminantes », dans *Actes de la 19e conférence sur le Traitement Automatique des Langues Naturelles*, Grenoble, p. 57-70.

Corbin D., 1987, *Morphologie dérivationnelle et structuration du lexique*, Tübingen, M. Niemeyer.

Cornuéjols A et Miclet L., 2010, *Apprentissage Artificiel. Concepts et algorithmes*, Paris, Eyrolles, p. 113-124.

Cusin-Berche F., Moirand S., Rakotonoelina F., Reboul-Touré S., Bosredon B., 2004, *Les mots et leurs contextes*, Paris, Presses de la Sorbonne Nouvelle.

Daille B., 1994, "Study and implementation of combined techniques for automatic extraction of terminology", in *The Balancing Act: Combining Symbolic and Statistical Approaches to Language, Proceedings of the Workshop of the 32nd Annual Meeting of the ACL*, USA, p. 29-36.

Darmesteter A., 1875, *Traité de la formation des mots composés dans la langue française comparée aux autres langues romanes et au latin*, France, A. Franck.

David S. et Plante P., 1990, « De la nécessité d'une approche morphosyntaxique en analyse de textes », *Intelligence artificielle et sciences cognitives au Québec*, 3(3), p. 140-155.

Dubois J. et al., 1994, *Dictionnaire de linguistique et des sciences du langage*, Paris, Larousse, p. 400.

Dugas A. et Manseau H., 1996, *Les verbes logiques : guide pratique*, Montréal : Éditions logiques.

Dugas A., 2010, « L'exploitaion d'un corpus pour une description automatique des verbes du français », dans *Actes du Colloque international La linguistique entre recherche et application*, p. 238-9.

Dunning, T., 1993, "Accurate Methods for the Statistics of Surprise and coincidence", *Computational Linguistics*, 19(1), p. 61–74.

Efron B. et Tibshirani R. J., 1993, *An Introduction to the Bootstrap*, New York, Chapman & Hall.

Enguehard C., 1993, « Acquisition de terminologie à partir de gros corpus », Actes Informatique&Langue Naturelle, Nantes, p. 373-384.

Feldman R., Fresko M., Kinar Y., Lindell Y., Liphstat O., Rajman M., Schler Y. et Zamir O., 1998, "Text mining at the term level", *LNAI: Principles of Data Mining and Knowledge Discovery*, 1510(1), p. 65–73.

Fotopoulou A., 1996, « Analyse automatique des textes de spécialités et dictionnaires électroniques des termes de télécommunications : remarques sur la morphosyntaxe des termes composés », dans *Actes de colloque Lexique, syntaxe et analyse automatique des textes*, n°34-35, p. 89-95.

Fradin B., 2003, *Nouvelles approches en morphologie*, Paris, Presses universitaires de France.

Frérot C., Bourigault D. et Fabre C., 2003, « Marier apprentissage endogène et ressources exogènes dans un analyseur syntaxique de corpus. Le cas du rattachement verbal à distance de la préposition 'de' », dans *Traitement Automatique des Langues*, p. 44-3.

Green, S., de Marneffe, Bauer M.-C., , J. et Manning C. D., 2011, "Multiword Expression Identification with Tree Substitution Grammars : A Parsing *tour de force* with French". In *Proceedings of the conference on Empirical Method for Natural Language Processing (EMNLP'11)*, p.725-735.

Green S., Marneffe M.-C. D., Manning C. D., 2013, "Parsing Models for Identifying Multiword Expressions", *Computational Linguistics*, vol. 39, no 1, p. 195-227.

Grevisse M., 2008, *Le bon usage*, Belgique, Duculot.

Gross G., Chaurand J., Mathieu-Colas M., Vives R. et Billy P., 1986, *Typologie des noms composés*, Villetaneuse : Paris XIII.

Gross G., 2012, *Manuel d'analyse linguistique*, France, Presses Universitaires Septentrion.

Gross M., 1986, *Grammaire transformationnelle du français : Syntaxe du verbe ; Syntaxe du nom*, France, Cantilène.

Gross M., 1986, "Lexicon Grammar: the Representation of Compound Words". In *Proceedings of Computational Linguistics (COLING'86)*, p. 1-6.

Gross M., 1995, « Une grammaire locale de l'expression des sentiments », *Langue française*, vol. 105, no 1, p. 70-87.

Guilbert L., 1963, « De l'utilisation de la statistique en lexicologie appliquée », *Études de Linguistique appliquée*, Paris, Didier, p. 12-24.

Habert B., Nazarenko A., Salem A., 1997, *Les linguistiques de corpus*, Paris, Armand Colin/Masson.

Habert B., Fabre C. et Issac F., 1998, *De l'écrit au numérique : constituer, normaliser, exploiter les corpus électroniques*, Paris, Informatiques.

Hamon T. et Hû O., 1998, "How to Evaluate Necessary Cooperative Systems of Terminology Building?", *Third International Conference on Language Resources and Evaluation*, Spain, Las Palmas de Gran Canaria, p. 1543-1550.

Harris, Z. S., 1979, *Mathematical Structures of Language*, New York, R. E. Krieger.

Harris, Z. S., 1976, *Notes du cours de syntaxe*, France, Seuil.

Harris, Z. S., 2007, *Langue et information*, France, Cellule de recherche en linguistique.

Hearst M. A., 1992, “Automatic Acquisition of Hyponyms from Large Text Corpora”. In *Proceedings of the 14th conference on Computational linguistics*, vol. 2, p. 539-545.

Hamon T. et Nazarenko A., 2001, “Detection of Synonymy Links between Terms: Experiment and results”. In *Recent Advances in Computational Terminology*, John Benjamins.

Huot H., 2001, *Morphologie, forme et sens des mots du français*, A. Colin.

Isabelle P. et S. Warwick-Armstrong, 1993, « Les corpus bilingues : une nouvelle ressource pour la traduction », dans Bouillon P. et Clas A. (dir.) (1993), *La traductique*, Presses de l’Univeristé de Montréal, p. 288-306.

Ibekwe-Sanjuan F., 2007, *Fouille de textes*, Paris, Lavoisier.

Issac F. 2007. “Yet Another Web Crawler”. In Fairon C., Naets H., Kilgarriff A., De Schryver Grilles-Maurice, (eds), *Building and Exploring Web Corpora (WAC3-2007), Cahiers du CENTRAL*, n°4, . Louvain-la-Neuve, Presses universitaires de Louvain, p. 57-68.

Jacquemin C. 1991. *Transformations des noms composés*. Thèse en informatique fondamentale, Université de Paris 7.

Jacquemin C. et Bourigault D., 2003, “ Extraction and Automatic Indexing ”, in R. Mitkov (ed.), *The Oxford Handbook of Computational Linguistics*, Royaume-Uni, Oxford University Press, p. 600.

Jiang J., “ Information Extraction from Text ”, 2012, in C. C. Aggarwal, C. Zhai (eds), *Mining Text Data*, US, Springer, p. 11-35.

Jouis C. et ARC A3, 2000, “ARC A3: A Method for Evaluation Term Extracting Tools and/or Eemantic Eelations between Terms from Corpora”. In *Proceedings of the Second International Conference on Language Resources and Evaluation*, Athens, LREC.

Junghoo C. et Hector G.-M., 2000, “The Evolution of the Web and Implications for an Incremental Crawler”. In Proceeding *VLDB '00 Proceedings of the 26th International*

Conference on Very Large Data Bases, San Francisco, CA, USA, Morgan Kaufmann Publishers Inc., p. 200-209.

Kassel G., 2009, « Vers une ontologie formelle des artefacts », *20es Journées Francophones en Ingénierie des Connaissances*, p. 121.

Kodratoff Y., 1999, "Knowledge Discovery in Texts: A Definition, and Applications". In *LNAI: Proc. of the 11th Int'l Symp. ISMS'99*, vol 1609, Warsaw, Springer, p. 16–29.

Lafferty J., Mccallum A. et Pereira F, 2001, "Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data". In *Proceedings of the Eighteenth International Conference on Machine Learning (ICML 2001)*, p. 282–289.

Landauer T. K. et Dumais S. T., 1997, "A Solution to Plato's Problem: the Latent Semantic Analysis Theory of Acquisition, Induction and Representation of Knowledge", *Psychological Review*. Vol(104), p. 211-240.

Lauriston A., 1994, "Automatic Recognition of Complex Terms: problems and the TERMINO solution", *Terminology*, vol. 1, n°1, p. 147-170.

Lebarbé T., 2002, « Validation des relations de dépendance par la cooccurrence sur Internet : présentation critique », *communication présentée au colloque TALN, Corpus et Web 2002*, Villetaneuse, p. 371-387.

Manning C. D. et Schütze H., 1999, *Foundation of Statistical Natural Language Processing*, chapitre 5, p. 141-177.

Manning, C. D., Raghavan, P. et Schtze, H., 2008, *Introduction to Information Retrieval*, Cambridge University Press, New York, NY, USA.

Mathieu-Colas M., 2009, « Morfetik : une ressource lexicale pour le TAL », *Cahiers de lexicologie*, Centre National de la Recherche Scientifique, p.137-146.

Mathieu-Colas M., 2009, « Essai de typologie des noms composés français », *Cahiers de lexicologie*, Didier Erudition, p. 71-125.

Maurel D., 1993, « Reconnaissance automatique d'un groupe nominal prépositionnel. Exemple des adverbes de date », *Lexique,* vol. 1, n° 1, p. 147-161.

Meilland J.-C. et Bellot P., 2003, « Extraction automatique de terminologie à partir de libellés textuels courts », dans G. Williams (ed.), *Linguistique de corpus*, France, Presses Universitaires de Rennes, p. 81-91.

Mejri, S., 1997, *Le figement lexical*, France, Scentifiques.

Moirand Sophie, 2004, « L'impossible clôture des corpus médiatiques : la mise au jour des observables entre catégorisation et contextualisation », *Revue Tranel*, vol(40), Suisse, Neuchâtel, p71-92.

Morlane-Hondère F., 2012, *Une approche linguistique de l'évaluation des ressources extraites par analyse distributionnelle automatique*, PhD thesis, Université Toulouse le Mirail

Moscarola J. B., Chirac J., 1995, « Les mots d'une campagne : quelques exemples d'analyse lexicone avecle Sphinx », *cahier du GEREG*, Annecy / France, Université de Savoie.

Mourlhon-Dallies F., Rakotonoelina. F. et Reboul-Touré S., 2004, *Les discours de l'internet : nouveaux corpus, nouveaux modèles ?*, Paris, Presses Sorbonne Nouvelle.

Mustafa el Hadi W., Timimi I. et Dabbadie M., 2004, "EVALDA-CESART Project: Terminological Resources Acquisition Tools Evaluation Campaign", In *Proceedings of the Fourth International Conference on Language Resources and Evaluation*, Lisbon, Portuga, LREC.

Nakagawa H. et Mori T., 2003, "Automatic Term Recognition based on Statistics of Compound Nouns and Their Components". *Terminology*, 9(2), pp. 201-219.

Nazarenko A., Zargayouna H., Hamon O. et Puymbrouck J. v., 2009, « Évaluation des outils terminologiques : enjeux, difficultés et propositions », *TAL*, Vol. 50 – n° 1/2009, pp. 257- 281.

Plaisantin Alecu B. P., Thomas I. et Renahy J., 2012, « La "multi-extraction" comme stratégie d'acquisition optimisée de ressources terminologiques et non terminologiques », dans *Actes*

de la 19e conférence sur le Traitement Automatique des Langues Naturelles, Grenoble, pp. 511-518.

Planas E., 2012, « BiTermEx : un prototype d'extraction de mots composés à partir de documents comparables via la méthode compositionnelle », dans *Actes de la 19e conférence sur le Traitement Automatique des Langues Naturelles (TALN'2012)*, Grenoble, France, pp.415-422.

Popescu-Belis et A., 1999, « L'évaluation en génie linguistique : un modèle pour vérifier la cohérence des mesures », *Langues (Cahiers d'études et de recherches francophones)*, 2(2), pp. 151–162.

Quiniou S., Cellier P., Charnois T. et Legallois D., 2012, « What About Sequential Data Mining Techniques to Identify Linguistic Patterns for Stylistics? », *Computational Linguistics and Intelligent Text Processing*, Heidelberg, pp. 166-177.

Ramshaw L. et Marcus M., 1995, "Text Chunking Using Transformation-Based Learning". In *ACL 3rd Workshop on Very Large Corpora*, pp. 82–94.

Rennie J., Shih, L., Teevan, J. et Karger, D., 2003, "Tackling the Poor Assumptions of Naïve Bayes Text Classifiers". *Proc. of ICLM-2003*, p. 616-623.

Saussure F. d., 1982, *Cours de linguistique générale*, France, Payot.

Seeker W. et Kuhn J., 2013, « Morphological and Syntactic Case in Statistical Dependency Parsing », *Computational Linguistics*, vol. 39, no 1, pp. 23-55.

Serrano L., Charnois T., Brunessau S., Grilheres B. et Bouzid M., 2012, « Combinaison d'approches pour l'extraction automatique d'événements », dans *Actes de la 19e conférence sur le Traitement Automatique des Langues Naturelles*, Grenoble, pp. 57-70.

Schäfer R. et Bildhauer F., 2012, "Building Large Corpora from the Web Using a New Efficient Tool Chain", *European Language Resources Association (ELRA)*, pp. 486-493.

Sinclair J., 1996, “EAGLES Preliminary Recommendations on Corpus Typology”, EAG-TCWG-CTYP/P. Pisa : ILC-CNR.

Smadja F., 1993, “Retrieving collocations from text : Xtract”, *Computational Linguistics,* vol. (19), pp.143-177.

Sampson, G.R. et Babarczy A., 2003, “A Test of the Leaf-ancestor Metric for Parse Accuracy”, *J. of Natural Language Engineering,* Vol. 9, pp. 365-380.

Schone P. et Jurafsky D., 2001, “Is Knowledge-Free Induction of Multiword Unit Dictionary Headwords a Solved Problem?”. In *Proceedings of Conference on Empirical Methods in Natural Language Processing,* USA, Lillian Lee and Donna Harma, pp. 100-108.

Snow R., Jurafsky D. et Andrew Y. N., 2004, « Learning Syntactic Patterns for Automatic Hypernym Discovery », *Advances in Neural Information Processing Systems,* British, Columbia, pp. 1297-1305.

Sparck Jones, K. et Galliers, J. R. (éd.), 1996, “Evaluating Natural Language Processing Systems: An Analysis and Review”, *Lecture Notes in Computer Science,* vol. 1083, Springer, p. 336-338.

Spence G., Marie-Catherine de M. et Christopher D. M., “Parsing Models for Identifying Multiword Expressions”, *Computational Linguistics,* Vol. 39, n°1, Etats-Unis, p. 195-227.

Suchomel V., 2012, “Efficient Web Crawling for Large Text Corpora”. In *Proceedings of the seventh Web as Corpus Workshop (WAC7),* Lyon, pp.39-43.

Tellier I. et Tommasi, M., 2011, « Champs Markoviens Conditionnels pour l’extraction d’information », dans *Eric GAUSSIER et François YVON, éditeurs : Modèles probabilistes pour l’accès à l’information textuelle,* Hermès, p. 223-267.

Thelwall M., 2001, «A Web Crawler Design for Data Mining», *Journal of Information Science,* UK, University of Wolverhampton, pp.319–325.

Tsarfaty R., Nivre J. et Andersson E., 2011. “Evaluating Dependency Parsing: Robust and Heuristics-free Cross-framework Evaluation”. In *Proceedings of EMNLP*, UK, Edinburgh.

Timimi I., 2006, « Évaluation des systèmes d'acquisition de terminologie : nouvelles pratiques, nouvelles métriques », dans *8es Journées internationales d’Analyse statistique des Données Textuelles*, France, p. 895-906.

Williams G. (éd.) 2005, *La linguistique de corpus*, Presses Universitaires de Rennes.

Wolfgang S. et Jonas K., 2012, “Morphological and Syntactic Case in Statistical Dependency Parsing”, *Computational Linguistics*, Vol. 39, n°1, Stuttgart, p. 23-55.

Annexes

Table des matières

Annexe 1 : Analyses syntaxiques et morphosémantiques

1. Patrons syntaxiques établis pour étiqueter les structures prédicat-argument dans la méthode supervisée

Classes	Distribution syntactico-sémantique	Patrons syntaxiques dérivés
Classe_1a	V+NAF	V+ADV/(ADV+ADV)/(ADV+ADV+ADV)+NAF; PrédN+de+(Dét)+NAF ; PrédN+de+(Dét)+ADV/(ADV+ADV)/(ADV+ADV+ADV)+A/(A+A)/(A+A+A)+NAF; NAF+Vpp; NAF+ADV/(ADV+ADV)/(ADV+ADV+ADV)+A/(A+A)/(A+A+A)+Vpp; NAF+Vpassif; NAF, NAF+et/ou+NAF, qui+pouvoir/devoir+Vpassif ;
Classe_1b	V+dessus/dessous/derrière /devant+NAF	PrédN+dessus/dessous/derrière/devant+NAF ; V+ADV/(ADV+ADV)/(ADV+ADV+ADV)+dessus/dessous/derrière/ devant+NAF; NAF+Vpp; NAF+ADV/(ADV+ADV)/(ADV+ADV+ADV)+A/(A+A)/(A+A+A)+Vpp; NAF+Vpassif; NAF, NAF+et/ou+NAF, qui+pouvoir/devoir+Vpassif ;
Classe_1c	V+à/de (+Det)+NAF	V+ADV/(ADV+ADV)/(ADV+ADV+ADV)+à/de(+Det)+NAF ;
Classe_2a	V+NAF+de/avec/par+ NAF	V+ADV/(ADV+ADV)/(ADV+ADV+ADV)+NAF+de/avec/par+NAF; PrédN+de+(Dét)+NAF+de/avec/par+NAF ; PrédN+de+(Dét)+ADV/(ADV+ADV)/(ADV+ADV+ADV)+A/(A+A)/(A+A+A)+NAF+de/avec/par+NAF; NAF+Vpp+de/avec/par+NAF; NAF+ADV/(ADV+ADV)/(ADV+ADV+ADV)+A/(A+A)/(A+A+A)+Vpp+de/avec/par+NAF; NAF+Vpassif+de/avec/par+NAF; NAF, NAF+et/ou+NAF, qui+pouvoir/devoir+Vpassif +de/avec/par+NAF;
Classe_2b	V+NAF/Nc+de/avec/par+ NAF	Idem
Classe_2c	V+NAF/Nc+de+NAF	Idem
Classe_3a	V+NAF/Nc+sous/sur/devant/derrière/au-dessus de/au-dessous de/ à la droite de…+NAF Prep= {sous, sur, devant, derrière, au-dessus de, au-dessous de, à la droite de…}	V+ADV/(ADV+ADV)/(ADV+ADV+ADV)+NAF+ prep+NAF; V+ADV/(ADV+ADV)/(ADV+ADV+ADV)+ prep+NAF+NAF; PrédN+de+(Dét)+NAF+prep+NAF ; PrédN+de+(Dét)+ADV/(ADV+ADV)/(ADV+ADV+ADV)+A/(A+A)/(A+A+A)+NAF+prep+NAF; NAF+Vpp+prep+NAF; NAF+ADV/(ADV+ADV)/(ADV+ADV+ADV)+A/(A+A)/(A+A+A)+Vpp+prep+NAF; NAF+Vpassif+prep+NAF; NAF, NAF+et/ou+NAF, qui+pouvoir/devoir+Vpassif +prep+NAF;
Classe_3b	V+NAF/Nc+en+NAF	Idem
Classe_3c	V+NAF/Nc+à+NAF	Idem
Classe_4a	V+NAF/Nc+dans+NAF Prep= {dans}	V+ADV/(ADV+ADV)/(ADV+ADV+ADV)+prep+NAF; V+ADV/(ADV+ADV)/(ADV+ADV+ADV)+Nc+prep+NAF; PrédN+prep+(Dét)+NAF; PrédN+prep+(Dét)+ADV/(ADV+ADV)/(ADV+ADV+ADV)+A/(A+A)/(A+A+A)+NAF; PrédN+prep+(Dét)+Nc+à+NAF; PrédN+de+(Dét)+ADV/(ADV+ADV)/(ADV+ADV+ADV)+A/(A+A)/(A+A+A)+Nc+prep+NAF; Vpp+prep+NAF ; Vpp+prep+ADV/(ADV+ADV)/(ADV+ADV+ADV)+A/(A+A)/(A+A+A)+NAF; Vpassif+prep+NAF; Vpassif+prep+ADV/(ADV+ADV)/(ADV+ADV+ADV)+A/(A+A)/(A+A+A)+NAF;
Classe_4b	V+Nc+avec/par/de+NAF	Idem
Classe_4c	V+Nc+à+NAF	Idem
Classe_4d	V+Nc+sur+NAF	Idem

Tableau 1 Patrons syntaxiques établis pour l'identification des structures prédicat-argument

2. Patrons syntaxiques établis pour étiqueter les structures prédicat-argument dans la méthode semi-supervisée

Patrons syntaxiques de base	Patrons syntaxiques dérivés
V+NAF	NAF+être+Vpassif NAF+Vpp NAF+être+ADV/(ADV+ADV)/(ADV+ADV+ADV)+Vpassif
V+NAF+prep+NAF	V+NAF, avec+NAF V+prep+NAF+NAF NAF+Vpp+prep+NAF NAF+Vpp, avec+NAF NAF+être+Vpassif+prep+NAF NAF+être+Vpassif, avec+NAF V+ADV/(ADV+ADV)/(ADV+ADV+ADV)+NAF,avec+NAF V+ADV/(ADV+ADV)/(ADV+ADV+ADV)+prep+NAF+NAF NAF+Vpp+ADV/(ADV+ADV)/(ADV+ADV+ADV), avec+NAF NAF+Vpp+ADV/(ADV+ADV)/(ADV+ADV+ADV)+prep+NAF NAF+être+ADV/(ADV+ADV)/(ADV+ADV+ADV)+Vpassif+prep+NAF NAF+être+ADV/(ADV+ADV)/(ADV+ADV+ADV)+Vpassif, avec+NAF
V+Nc+prep+NAF	V+Nc, prep+NAF être+Vpassif+prep+NAF être+Vpassif, prep+NAF V+ADV/(ADV+ADV)/(ADV+ADV+ADV)+Nc, prep+NAF être+ADV/(ADV+ADV)/(ADV+ADV+ADV)+Vpassif+prep+NAF être+ADV/(ADV+ADV)/(ADV+ADV+ADV)+Vpassif, prep+NAF V+prep+NAF V+ADV/(ADV+ADV)/(ADV+ADV+ADV)+prep+NAF
V+prep+NAF	V+ADV/(ADV+ADV)/(ADV+ADV+ADV)+prep+NAF

Tableau 2 Patrons syntaxiques établis dans la méthode distributionnelle semi-supervisée

3. Suffixes des noms de métiers

Suffixes	Exemples
-teur, -trice, deur, deuse, -drice, eur, -euse	*acteur, auteur, administrateur, ambassadeur, appareilleur, chanteur, vendeur, abatteur, assureur,* etc.
-iste	*aciériste, accessoiriste, archiviste, aubergiste, audioprothésiste, anesthésiste,* etc.
-ier, -ière	*métallier, ouvrier, mercier, banquier, batelier, bâtonnier, bijoutier, bobinier,* etc.
-ien, -ienne,	*chirurgien, électricien, clinicien, gardien, géophysicien, glypticien,* etc.
-aire	*mandataire, gestionnaire, libraire, lapidaire, manutentionnaire, bibliothécaire, magnétothécaire,* etc.
-eute	*musicothérapeute, psychothérapeute, thérapeute,* etc.
-ain, -aine	*écrivain, forain,* etc.
-cial, -ciale	*commercial,* etc.
-able	*comptable, responsable,* etc.
-ate	*acrobate,* etc.
-at	*avocat, magistrat,* etc.
-in, -ine	*laborantin,* etc.
-tre	*peintre,* etc.
-and, -ande	*marchand, tisserand,* etc.
-on, -onne	*tâcheron, forgeron,* etc.
-isan	*artisan,* etc.
-ant, -ante	*aspirant, assistant, commerçant, consultant, correspondant, enseignant,* etc.
-ard, -arde	*motard,* etc.
-er, -ère	*écailler, volailler, berger, boulanger, fromager, horloger, boucher, vacher,* etc.
-er (Empr. de l'ang.)	*doker, designer, skipper, speaker, schooper, manager,* etc.

Tableau 3 Suffixes des noms de métiers

4. Analyses morphématiques des noms de métiers

Morphèmes lexicaux	**Morphèmes lexicaux**	**Suffixes**	**Exemples**
musico-, psych(o)-	*thérapie*	*-eute*	*musicothérapeute,* etc.
audio-	*prothèse*	*-iste*	*audioprothésiste,* etc.
géo-	*physique*	*-ien*	*géophysicien,* etc.
Préfixes	**Morphèmes lexicaux**	**Suffixes**	**Exemples**
an-	*esthésie*	*-iste*	*anesthésiste,* etc.
ex-, in-	*-terne*		*externe, interne,* etc.
Morphèmes lexicaux	**Suffixes**	**Suffixes**	**Exemples**
compter	*-able*	*conversion*	*comptable,* etc.
commercer	*-cial*	*conversion*	*commercial,* etc.
informer	*-ique*	*-ien*	*informaticien,* etc.
Morphèmes lexicaux	**Morphèmes lexicaux**	**Morphèmes lexicaux**	**Exemples**
grapho-, psych(o)-, géo-, gemmo-, gynéco-, hydro-	*-logie*	*-logue*	*psychologue, géologue, graphologue, gemmologue, gynécologue, hydrogéologue, etc.*
bio-, géo-, hydro-,	*-graphie*	*-graphe*	*biographe, géographe, hydrographe,* etc.
psych(o)-, péd(i)-,	*-iatrie*	*-iatre*	*psychiatre, pédiatre,* etc.
homéo-, ostéo-,	*-pathie*	*-pathe*	*homéopathe, ostéopathe,* etc.
photogramme, géo-	*-métrie*	*-mètre*	*photogrammètre, géomètre*
agro-, astro-,	*-nomie*	*-nome*	*agronome, astronome,* etc.
pédi-, manu-		*cure*	*pédicure, manucure*
contre		*maître*	*contremaître,* etc.
bar- (Empr. de l'ang.)		*-maid,* (Empr. de l'ang.)	*barmaid,* etc.
bar-, camera-, perch-, ski- (Empr. de l'ang.)		*-man,* (Empr. de l'ang.)	*barman, cameraman, camerawoman, perchman, skipman,* etc.

Tableau 4 Analyses morphématiques des noms de métiers

5. Patrons morphosyntaxiques des noms de métiers composés

Patrons simples	**Relations sémantiques entre les constituants**	**Exemples**
NMP+N2/GN2	N2/GN2-> modifie N1 et indique le contenu du travail de N1	*assistant administration, producteur télé, garde-chasse, guide nature, graphiste multimédia, sage-femme,* etc.
N1/GN1(Emprunt)+NMP	N1/GN1(Emprunt)-> indique le contenu ou l'objet du travail de N2	*game designer, campus manager, web planner, data protection manager*, etc.
NMP1+NMP2	juxtaposition de N1 et N2	*chirurgien-dentiste, charcutier-traiteur, chauffeur-livreur*, etc.
Préfixe+NMP	Préfixe->modifie N et indique le contenu du travail de N1	*bio-informaticien, éco-développeur, médico-psychologique*, etc.
N1+A	A-> modifie N1 et indique le contenu du travail de N1	*animateur touristique, contrôleur budgétaire, médiateur culturel*, etc.
NMP+de(Det)+N2/GN2	N2->modifie N1	*directeur de thèse, maréchal de camp, accessoiristes de spectacles, agent des services techniques*, etc.
NMP+en+N2/GN2	N2/GN2-> modifie N1 et indique le domaine du travail de N1	*installateur en électroménager, peintre en carrosserie, technicien en réseau de communication*, etc.
NMP+à(Det)+N2/GN2	N2/GN2-> modifie N1 et indique le contenu du travail de N1	*accompagnateur à la mobilité, conseiller à l'emploi*, etc.
NMP+sur(Det)+N2/GN2	N2/GN2 -> modifie N1 et indique le contenu du travail de N1	*décorateur sur porcelaine, décorateur sur verre, graveur sur pierre, opérateur sur les marchés*, etc.
NMP+dans(Det)+N2/GN2	(dans(Det)+N2/GN2) -> modifie NMP1	*professeur dans l'enseignement agricole*, etc.
NMP+Vpp+en+N2/GN2	N2/GN2 -> modifie N1 et indique le domaine du travail de N1	*journaliste spécialisé en environnement*, etc.
Patrons complexes		
(NMP1+A)+de(Det)+N2/GN2	N2/GN2 -> modifie le nom de métier composé (N1+A)	*directeur financier du cabinet (chief financial officer (CFO)), etc.*
(NMP1+A1)+et/ou+A2	(et/ou+A2) -> surajouté sur le nom de métier composé (N1+A1) pour compléter le modifieur A1	*conseiller conjugal et familial, etc.*
(NMP1+N2/GN2)+et/ou + N3/GN3	(et/ou+N3/GN3) -> surajouté sur le nom de métier composé (N1+N2/GN2) pour compléter le modifieur N2/GN2	*expert incendies et risques divers, ingénieur calcul et appui scientifique, etc.*
(NMP1+prep(Det)+N2/GN2)+et/ou+de+N3/GN3	(et/ou+N3/GN3) -> surajouté sur le nom de métier composé (N1+prep(Det)+N2/GN2) pour compléter le modifieur (prep(Det)+N2/GN2)	*conducteur de grues et d'engins de levage, directeur de centre de vacances ou de loisirs, etc.*
(NMP1+NMP2)+NMP3	juxtaposition de (N1+N2) et N3	*maître-nageur sauveteur, etc.*
NMP1+(NMP2+NMP3)	juxtaposition de N1 et (N2+N3)	*infirmier sapeur-pompier, etc.*
PE1+(PE2+NMP)	préfixe1 ->surajouté sur le nom de métier (préfixe2+N)	*oto-rhino-laryngologiste, etc.*
(PE+NMP1)+en+N2	N2->indique le domaine du travail de (préfixe+N1)	*technico-commercial en agrofournitures, etc.*
(NMP1+et+NMP2)+de(Det)+N3/GN3	N3/GN3->indique le contenu du travail de (N1+et+N2)	*analyste et contrôleur des risques financiers, etc.*
(NMP1+de(Det)+N2/GN2)+en+N3/GN3	N3/GN3->modifie le nom de métier composé (N1+de(Det)+N2/GN2)	*technicien du bâtiment en énergies renouvelables, etc.*

Tableau 5 Patrons morphosyntaxiques des noms de métiers composés

6. Règles d'allomorphie

Règles d'allomorphie	Exemples
Règles pour les verbes	
recev- ->récept-	*recevoir -> recv- ->récept- -> récepteur*
cev- ->cept-	*concevoir -> concev- -> concept- -> concepteur*
str- ->strat-	*administrer ->administr- -> administrat- -> administrateur*
uis- ->uct-	*conduire -> conduis- -> conduct- -> conducteur*
dig- ->dact-	*rédiger ->rédig- -> rédact- -> rédacteur*
ment- -> mentat-	*alimenter -> aliment- -> alimentat- ->alimentateur*
alis- -> alisat-	*réaliser -> réalis- ->réalisat- -> réalisateur*
ilis- -> ilis-	*utiliser -> utilis- -> utilisat- -> utilisateur*
démontr- ->démonstrat-	*démontrer -> démontr- -> démonstrat- ->démonstrateur*
tiss- -> tisser-	*tisser -> tiss- -> tisser- -> tisserand*
agn- -> agnat-	*accompagner -> accompagn- -> accompagnat- ->accompagnateur*
ys- -> yste	*analyser -> analys- -> analyste*
g- -> geant	*diriger -> dirig- -> dirigeant*
c- -> çant	*commercer -> commerc- -> commerçant*
Règles pour les noms	
-c -> -ch-	*bouc -> bouch- -> boucher*
-èse -> -és-	*prothèse -> prothés- -> prothésiste*
-on -> -onn-	*gestion -> gestionn- -> gestionnaire*
-eau -> -el-	*bâteau -> batel- -> batelier*
-ou ->-out-	*bijou- -> bijout- -> bijoutier*
-e -> Ø (qui signifie vide)	*archive -> archiv- -> archiviste*
-et -> - étaire	*secret ->secrét- -> secrétaire*
-sophie- -> -sophe	*philosophie -> philosophe*
-nomie- -> -nome	*astronomie -> astronome*
-nautique -> -naute	*aéronautique -> aéronaute*
-métrie -> -mètre	*géométrie -> géomètre*
-iatrie -> -iatre	*pédiatrie -> pédiatre*
-yse- -> -yste	*analyse -> analyste*
-iologie -> -iologiste	*radiologie -> radiologiste*
-^[i]ologie -> -logue	*géologie -> géologue*
-thérapie -> -thérapeute	*ergothérapie -> ergothérapeute*
-thèque -> -thécaire	*bibliothèque -> bibliothécaire*
-pathie -> -pathe	*ostéopathie -> ostéopathe*
-ique -> -icien	*physique -> physicien*
-actoire -> -antin	*laboratoire -> laborantin*
-il -> -illeur	*outil -> outilleur*
engin -> ingénieur	*ingénieur*
moto- > motard	*motard*

Tableau 6 Règles d'allomorphie

7. Patrons morphosyntaxiques établis pour reconnaître les noms de métiers composés

Graphes	Opérateurs	Patrons morphosyntaxiques
2gramme_NMP.grf	←NMP	NMP+NMP
	←et/ou+NMP	NMP+et/ou+NMP
	←N/GN	NMP+N
	←A	NMP+A
	préfixe→	préfixe+NMP
	N/GNEmprunt→	NEmprunt+NMP
3gramme_NMP.grf	←prep(Det)+N/GN (prep={de,à,sur,en,dans})	NMP+prep(Det)+N
	←N/GN	NMP+NN NMP+NA
4gramme_NMP.grf	←prep(Det)+N/GN (prep={de,à,sur,en,dans})	NMP+prep(Det)+NN, NMP+prep(Det)+NA
	←Vpp+en+N/GN	NMP+Vpp+en+N
	←N/GN	NMP+N+prep(Det)+N
5gramme_NMP.grf	←Vpp+en+N/GN	NMP+Vpp+en+NN, NMP+Vpp+en+NA
	←prep(Det)+N/GN (prep={de,à,sur,en,dans})	NMP+prep(Det)+N+prep(Det)+N
2_3gramme1_NMP.grf	préfixe→	préfixe+2_3gramme_NMP
	N/GNEmprunt→	NEmprunt+2_3gramme_NMP
	←NMP	2_3gramme_NMP+NMP
	←NMPC	NMP+2_3gramme_NMP
	←N/GN	2_3gramme_NMP+N
	←A	2_3gramme_NMP+A
2_3gramme2_NMP.grf	←N/GN	2_3gramme_NMP+NN, 2_3gramme_NMP+NA
	←NMPC	2_3gramme_NMP+2gramme_NMP
	←et/ou+NMPC	NMP+et/ou+2_3gramme_NMP
	←prep(Det)+NMPC	NMP+prep(Det)+2_3gramme_NMP
	←prep(Det)+N/GN (prep={de,à,sur,en,dans})	2_3gramme_NMP+ prep(Det)+N
	←et/ou+N	2_3gramme_NMP+et/ou+N
	←et/ou+A,	2_3gramme_NMP+et/ou+A
2_3gramme3_NMP.grf	←et/ou+prep(Det)+N/GN	2_3gramme_NMP+et/ou+prep(Det)+N
	←et/ou+N/GN	2_3gramme_NMP+et/ou+NN 2_3gramme_NMP+et/ou+NA
	←prep(Det)+N/GN	2_3gramme_NMP+prep(Det)+NN 2_3gramme_NMP+prep(Det)+NA
	←N/GN	2_3gramme_NMP+N+prep(Det)+N
	←NMPC	2_3gramme_NMP+3gramme_NMP
	←et/ou+NMPC	2_3gramme_NMP+et/ou+2gramme_NMP
	←prep(Det)+NMPC	2_3gramme_NMP+prep(Det)+2gramme_NMP
	←Vpp+prep(Det)+N/GN	2_3gramme_NMP+Vpp+prep(Det)+N
2_3gramme4_NMP.grf	←Vpp+prep(Det)+N/GN	2_3gramme_NMP+Vpp+prep(Det)+NN 2_3gramme_NMP+Vpp+prep(Det)+NA
	←et/ou+N/GN	2_3gramme_NMP+et/ou+N+prep(Det)+N
	←prep(Det)+N/GN	2_3gramme_NMP+prep(Det)+N+prep(Det)+N
	←et/ou+prep(Det)+N/GN	2_3gramme_NMP+et/ou+prep(Det)+NN 2_3gramme_NMP+et/ou+prep(Det)+NA
	←et/ou+NMPC	2_3gramme_NMP+et/ou+3gramme_NMP
	←prep(Det)+NMPC	2_3gramme_NMP+prep(Det)+3gramme_NMP
2_3gramme5_NMP.grf	←et/ou+prep(Det)+N/GN	2_3gramme_NMP+et/ou+prep(Det)+N+prep(Det)+N
	←Vpp+prep(Det)+N/GN	2_3gramme_NMP+Vpp+prep(Det)+N+prep(Det)+N

Tableau 7 Patrons morphosyntaxiques pour reconnaître les noms de métiers composés

8. Patrons morphosyntaxiques simples des noms d'artefacts composés

Information / Patterns simples	Relations sémantiques entre les constituants	Exemples
VpN(noun)	N->l'objet qui reçoit l'action Vp	*allume-cigare, appui-pouce*, etc.
N1pourN2/GN2	N2->l'objet auquel N1 est réservé	*crème pour les mains, chaussure pour femme pied droit*, etc.
N1surN2	N2-> la propriété de N1	*batteur sur socle, batteur sur pied*, etc.
N1parN2/GN2	N2->l'outil de N1	*assemblage par boulons, commande par bouton poussoir*, etc.
N1avecN2	N2->une partie de N1	*appliques avec halogène*, etc.
N1deN2	N2->function modifying N1	*ailette de refroidissement, aimant de retenue*, etc.
N1deN2	N2->l'objet auquel N1 est réservé	*accessoire de jeu*, etc.
N1deN2	N1->une partie de N2	*anse de panier*, etc.
N1àN2	N2->une partie de N1	*armoire à glace, alternateur à turbine*, etc.
N1àN2/GN2	GN2->propriété de N1	*appareil à grand capteur, bol à bec verseur, brosse à tapis longs poils*, etc.
N1àN2	N2->l'objet qui reçoit l'action lancée par N1	*batteur à œufs*, etc.
N1àDetN2	N2->l'usage de N1	*boîte aux lettres*, etc.
N1àVN2	VN2(structure verbe-objet)->fonction de N1	*appareil à battre les collets*, etc.
N1en forme de N2	N2->propriété de N1	*chapeau en forme de champignon*, etc.
NVpp	Vpp->propriété de N	*accotement enherbé*, etc.
NàV	V->usgae de N	*aiguille à brider*, etc.
N1Vpp deN2	Vpp deN2->construction du verbe	*bac garni de sac*, etc.
N(PrefixA)	PréfixA modifies le N	*appareil électro-acoustique*, etc.
PrefixeN	Préfixe->fonction ou propriété du N	*anti-vol, auto-cuiseur*, etc.
N1N2	N2 modifies N1	*diffuseur video, appareil photo*, etc.
N1sansN2	N2 modifies N1	*chapeau sans bord*, etc.
NA	A->modifie N	*aiguille hypodermique, console numérique, etc.*
N1GN2	GN2->propriété de N1	*ampoule basse consommation*, etc.
N1NbrN2	NbrN2->modifie N1	*commutateur 10 positions*, etc.
AN1	A->propriété de N1	*flexible aspirateur*, etc.
N1N2	N2 est une unité lexicale anglaise	*filme bluray*, etc.
N1NbrdansNbr	NbrdansNbr->modifie N1	*combiné 2 dans 1*, etc.

Tableau 8 Patrons morphosyntaxiques simples des noms d'artefacts

9. Patrons morphosyntaxiques complexes des noms d'artefacts composés

Information / Patterns complexes	Relations sémantiques entre les constituants	Exemples
N1(VpN2)	(VpN2)->fonction de N1	*arbre porte galet*, etc.
(VpN1)enN2	N2 -> matière de (VpN1)	*abat-jour en laine*, etc.
(N1deN2)àN3	N3 -> modifie (N1deN2)	*boitier de direction à vis et galet*, etc.
(N1àV/N2)enN3	N3 modifies (N1àV/N2)	*boîte à déjeuner en néoprène, fard à joue en poudre*, etc.
(N1N2)àN3	N3->l'objet auquel (N1N2) est réservée	*brosse peigne à cils et sourcils*, etc.
(N1deN2)surN3	N3->nature de (N1deN2)	*accessoire de jeu sur pc*, etc.
(N1A1)A2	A1->propriété de N1; A2->propriété de (N1A1)	*plaque vitrocéramique radiant*, etc.
(NVpp1)Vpp2	Vpp1->propriété de N1; Vpp2->propriété de (NVpp1)	*poche plaquée surpiquée*, etc.
(N1N2) A	N2 modifie N1->la propriété ; A->modifie (N1N2)	*ampoule globe opale, appareil photo compact*, etc.
(N1Vpp)avecN2	N2->propriété de (N1Vpp)	*poche passepoilée avec renforts*, etc.
(N1A)avecGN2	GN2->propriété de (N1A)	*appliques murales avec halogène tibse blanc*, etc.
N1(PréfixeN2)	(PréfixeN2)->propriété de N1	*bonde pare bruit*, etc.
(N1A)deGN2	GN2->modifie (N1A)	*avertisseur sonore de sac plein*, etc.
(N1deN2)àN3	N ->l'objet auquel (N1deN2) est réservé	*bâton de rouge à lèvre*, etc.
(N1N2)N3	N2->matière de N1 ; N3->matière de (N1N2)	*bas nylon voile*, etc.
N1de(N2N3)	(N2N3)->l'objet de N1	*lecteur de carte mémoire*, etc.
(N1deN2)deN3	N3->modifie (N1deN2)	*bande d'arrêt d'urgence*, etc.
(N1àN2)àDetN3	N ->modifie (N1àN2)	*broyeur à interrupteur au couvercle*, etc.
(N1N2)NbrN3	NbrN3->(N1N2)	*carte sd 18 go, film caméscope 8mm*, etc.
(N1avecGN2)pourN3	N3->l'objet auquel (N1avecGN2) est réservé	*chaine avec station d'accueil pour ipad*, etc.
(N1deN2)parN3	N3->modifie (N1deN2)	*commande de température par cadran*
(N1A)pour(Det)N2	N2->l'objet ou l'usage auquel (N1A) est réservé	*crème lavante pour les mains, injecteurs élastiques pour lavement, kit d'enceintes pour baladeur*, etc.
(VpN1)àN2N3	N2N3->propriété de (VpN1)	*dégage lame à bouton poussoir*, etc.
(NA)dePn	Pn -> signifie prédicat nominal figé ; Pn->modifie (NA)	*dispositif automatique de remise en marche*
(N1V)N2	V et N2 sont tous en anglais ; N2->propriété de (N1V)	*enceinte surround back infinity*, etc.
(N1N2)(PréfixeN3)	N2, Prefixe et N3 en anglais ; (PréfixeN3)->modifie (N1N2)	*enceinte wireless stereo speaker*, etc.
(N1N2N3)Vpp	Vpp->modifie (N1N2N3)	*ensemble home cinéma intégré*, etc.

Tableau 9 Patrons morphosyntaxiques complexes des noms d'artefacts

Annexe 2 : Extraits des corpus utilisés pour l'acquisition automatique du vocabulaire

1. Extraits des corpus de noms d’artefacts

1.1. AutoCara_Forum_Automobile

(TITRE:Connection Bluetooth - C4 / C4 Picasso - Citroën - FORUM Marques:TITRE)

Iza6238rofil : nouveau membre posté le 20-09-2010 à 16:34:21

Publicité

Bonjour, j'ai une C4 coupé. J'essaie de connecter mon téléphone en bluetooth mais le menu "Rechercher un téléphone" est grisé et je ne peux pas le sélectionner. Une idée pour pouvoir enfin connecter mon téléphone ? Est ce que ça vient de ma voiture ou de mon tél ? Bonjour j'ai la meme chose, à mon avis sur nos versions le bleutooth n'est pas disponible.ar contre, j'ai une petite entre de carte SIM. Et je n'arrive pas à mettre ma carte, car cette entre est trop grande (Ma carte tombe à l'interieur et je suis oblige de prendre une pince pour la sortir). Donc auriez vous une idée, pour connecter un telephone à ma voiture. Merci za6238 a écrit : bonjour, j'ai une C4 coupé. J'essaie de connecter mon téléphone en bluetooth mais le menu "Rechercher un téléphone" est grisé et je ne peux pas le sélectionner. Une idée pour pouvoir enfin connecter mon téléphone ? Est ce que ça vient de ma voiture ou de mon tél ? Quelle finition ? Quelle année ? Quel type d'autoradio ? quelles options ? Bonjour, je sais que ce thread est vieux, mais une réponse peu aider d'autres utilisateurs avec le même problème. (je me suis inscrit expres pour partager)e viens d'acheter un c4 picasso, et je me suis retrouvé confronté au problème du menu "rechercher un téléphone" grisé.es dizaines de tentatives avant de pouvoir connecter mon téléphone (un android de base) et rien n'y faisait (y compris debrancher/rebrancher la batterie)n fait, je n'ai jamais réussi à avoir le menu accessible mais j'ai pu faire l'association dans l'autre sens. couper le contact tourner la clef de contact légèrement (sans mettre le contact) allumer l'autoradio sur le téléphone: recherche de périphérique bluetooth - un appareil nommé "CITROEN" est trouvé sur le téléphone: appairer le peripherique "CITROEN" sur l'affichage de l'autoradio : un message apparait: "autoriser la connexion avec nom telephone " - repondre oui avec la touche ok de l'autoradio. Sur le téléphone un message "appairé avec succès" apparaît mon téléphone s'ensuite connecté à l'autoradio en tant que "ressource audio du téléphone", puis un message "autoriser CITROEN à accéder au répertoire" est apparu. répondre oui (bien sûr) et voila, tout fonctionne maintenant, mon portable est reconnu quelques secondes apres etre entré dans la voiture, j'accède via les commandes au volant au répertoire et au journal des appels du téléphone.

(TITRE:Systéme anti pollution défaillant - C8 - Citroën - FORUM Marques:TITRE) phlag14

Profil : nouveau membre posté le 23-05-2010 à 09:20:09

Publicité bonjour à tous, je poséde un C8 2lHDI 107 de Déc 2003 avec 130000 km , cette semaine je l'ai mis au garage pour une distibution et le fap (1450€ !!!!!), j'ai récupéré ma voiture samedi matin, retour chez moi (10 km) RAS, puis en fin de journée je la prend pour aller faire des courses et là au bout de 30 km la voiture se coupe "défailance systéme anti pollution" rien à faire, elle redémarre puis se recoupe, quelqu'un a-t'il des infos ???? sachant qu'il m'ont aussi changé le filtre à gasoil, une prise d'air peut-elle provoquer ce genre de panne ???? pour info, je cherchai à savoir comment réparer l'éclairage de mon afficheur monochrome; et bien comme je m'en doutai, l'éclairage est intégré au circuit , il n'existe pas de lampe pour cela, il faut changer l'afficheur (~500€) ou trouver une boite d'électronique pour voir ce qu'il est possible de faire !!!

Je crois que je vais changer de marque, parce que citroen ca commence à bien faire. bon courage à tous

Le véhicule sort d'une 'grosse' intervention par un garage = C'est à lui de remettre tout en place. La probabilité que cela provienne de l'intervention est beaucoup plus élevé que cela soit une coincidence qu'un autre organe claque juste à ce moment ! Quelques pistes:- vérifier le bon branchement et l'étanchéité de toutes les durites du système de dépression (depuis pompe à vide, vers mastervac, les n electrovannes, les capsules de commande (EGR, volet air chaud, soupape turbo, swirl),- capsule de swirl: tige cassé, membrane percé,- les electrovannes (au moins 6),- le débimetre,- différents capteurs pression / temperature / ouverture reservoir ...u qu'il y a décision d'arret du moteur, l'électronique doit reporter des messages d'erreur assez précis sur les organes en cause. Le garage doit etre capable de les lire, les interpréter et remédier à la panne. Once au garage, c'est leur boulot surtout que le véhicule vient juste d'en sortir ! merci pour les infos, j'attend avec impatience l'ouverture du garage demain matin. salut , il y a 3 semaines d'ici g eu le meme message avec ma citroen c8 " filtre antipolution defaillllant" et maintenant plus rien indiquant, je me suis renseigné aupres des collegues, ils m'ont que j'allait consommmé plus, est ce vrais ? c koi exatement

(TITRE:Aide changement plaquette de freins av et ar Xantia - Xantia - Citroën - FORUM Marques:TITRE) thogie

Profil : nouveau membre posté le 15-06-2007 à 11:20:57

Publicité

Slt à tous, 'aurai besoin de conseils pour le changement des plaquettes de freins avant et arrière de ma Citroën XANTIA 1,6L tentation. Je voudrai savoir si je dois me procurer des outils spéciaux, les techniques (afin de me faciliter le tache !!)...par avance merci pour toutes vos réponses !!hogie salut a l'arriere rien de particulier pour l'avant il existe un outil pour repousser le piston (il faut le faire tourner en le repoussant)avec un peu de patience une pince multi et un tournevis peuvent faire l'affaire k, merci pour ta réponse nico2298 ! Éventuellement, y a il un site qui montre les étapes du changement ??? (avec photos si possible) il te fauit aussi une cle torx de 55 pour defaire l'etrier avant pour repousser le piston des etriers de frein avant , il faut faire tourner une clé allen dans le sens inverse des aiguilles d'une montre.'ai cherché et forcer avec un demonte pneu pendant des heures sans y arriver , avant de trouver le conseil dans la revue technique de la voiture. Un tournevis fait très bien l'affaire pour repousser les pistons a l'avant, faut juste un peu de patience. justement ca fait deux jours que je me bat avec les pistons alors qui faut tourner ce comme pour verifier le niveau de la boite toute la voiture bouge mais le boulon rien alut tu peux te servir d'un sert-joint qui visse si tu vois se que je veu dire.

Bin... c'est pas une Xantia ! Il y a un ressort et un amortisseur !

Nan nan mais c'est pour "l'exemple" j'aurai du le precisé en plus c'est un ressort jaune scad a écrit :alut tu peux te servir d'un sert-joint qui visse si tu vois se que je veu dire.http://images.forum-auto.com/mesimages/321082/dsc017355nz.jpgpeut tu me dire comment changer mon étrier de frein de ma xantia si tu peut merci d'avance. Bonjour à tous et Bonne année ...

Besoin d'un petit conseil sur changement plaquettes Avant sur 2,1 TD de 1999:je crois me rappeler qu'il faut positionner le piston d'une manière particulière, pour que le réglage automatique du frein à main se fasse? je n'ai rien vu à ce sujet sur les tutos consultés (notamment celui de françois2b) ???n'est-ce pas important? et si oui, est-ce que le petit trou sur le piston doit se trouver en haut lorsque l'étrier est remonté? avant de placer les plaquettes? Merci pour vos réponses. bonsoir, tu dois repousser le piston au maximum en le tournant, sinon tes plaquettes neuves risquent de ne pas rentrer. ensuite, tu règles le piston pour que l'ergot de la plaquette passe par un des trous du piston. Bonjour françois2b et merci pour ta réponse rapide. j'en conclus que le rattrappage du frein à main se met en place tout seul du fait que le piston est repoussé au maximum ?encore merci, notamment pour tous tes tutos : sacré boulot

… …

1.2. CommentCaMarche_Actualités_InfoJeuxImage

(TITRE:La Nuit des Musées - Nouvelle appli:TITRE)

Avec son application en réalité augmentée, le musée d'histoire de Marseille, ouvert depuis septembre 2013, invite ses visiteurs à emprunter la voie historique.

Sortir des murs par la réalité augmentée

La nouvelle application reprend le contenu des expositions, tout en sortant des murs du bâtiment pour se concentrer sur la plus ancienne ville de France et ses 26 siècles d'existence.

Par « extension numérique », l'utilisateur parcourt la cité en superposant le paysage actuel à celui de l'Antiquité.

Une balade interactive en 10 étapes

Le parcours s'étend du port antique jusqu'au MuCEM en 10 étapes, jalonnées de différentes ambiances sonores, par exemple, un char romain roulant sur les pavés. Des interviews d'historiens complètent également ce retour dans le passé, de même que des panneaux explicatifs avec des QR codes.

La technologie permet aux visiteurs de rentrer dans l'univers, de ressentir les émotions du passé et donc de se l'approprier via cette immersion dans le temps.

Massalia au temps de la Pax Romana

Les paysages initiaux sont restitués. Le promeneur, en orientant sa tablette et son smartphone vers le port aperçoit alors des épaves grecques ou romaines provenant des collections du musée. « Extension Numérique » est disponible gratuitement sur l'AppStore et

Google Play

Bonne balade à travers le temps !

Crédit photographique : marseille.fr

(TITRE:Secteur des télécoms : un décret permet au gouvernement d'encadrer les investissements étrangers:TITRE)

Jusque-là, les investissements étrangers n'étaient soumis à autorisation du gouvernement que pour certains secteurs d'activités stratégiques. Un décret, publié au journal Officiel ce 15 mai, vient d'élargir les secteurs dans lesquels le gouvernement pourra être actif et imposer ses conditions. Parmi eux, le marché des communications électroniques

Le ministre de l'Économie, du Redressement productif et du Numérique, Arnaud Montebourg n'a pas réussi à influencer la vente de SFR et peine à faire entendre sa voix pour celle d'Alstom.

Pour étendre le pouvoir du gouvernement sur les entreprises privées, le décret n° 2014-479 vient de modifier le code monétaire et financier

Journal Officiel ce 15 mai, il concerne les investissements étrangers qui sont soumis à autorisation préalable

Ce décret élargit les secteurs jugés stratégiques. Aujourd'hui, les marchés concernés sont :

- électricité, gaz, hydrocarbures ou autre source énergétique
- réseaux et services de transport
- approvisionnement en eau
- réseaux et services de communications électroniques- santé publique

Le ministre Arnaud Montebourg a déclaré qu'il « s'agit pour le gouvernement de s'assurer que ses objectifs légitimes seront pleinement pris en compte par les investisseurs étrangers, qu'ils soient issus de pays de l'Union européenne ou de pays tiers. Au besoin, le gouvernement pourra demander des engagements spécifiques ou imposer des conditions à la réalisation des investissements concernés, afin de garantir la préservation des intérêts du pays ». Le Premier ministre estime que : «la puissance publique doit avoir son mot à dire » sur les secteurs stratégiques.

Si ce décret a pour but de « lutter contre la fraude et l'évasion fiscale », il renforce le protectionnisme économique et pourrait limiter les investissements venus de l'étranger.

La commission européenne n'a pas tardé à réagir

Michel Barnier, commissaire chargé du Marché intérieur, a déclaré que : «l'objectif de protéger les intérêts essentiels stratégiques dans chaque État membre est essentiel dès qu'il s'agit de sécurité ou ordre public. C'est clairement prévu dans le traité. Mais (nous) devons vérifier si cet objectif est appliqué de manière proportionnée sinon cela reviendrait à du protectionnisme».

Crédit photo: Wikipedia

(TITRE:Skype : les appels vidéo de groupe redeviennent gratuits :TITRE)

Le célèbre logiciel de messagerie instantanée et d'appels vidéo par Internet

Skype annonce que son outil de visioconférence groupée réintègre les fonctionnalités gratuitement disponibles.

Cette évolution permet aux utilisateurs du logiciel d'effectuer des appels multi-participants gratuits sur les plateformes Windows, Mac, and box One, en attendant que la mise à niveau soit déployée sur d'autres plateformes : notamment l'application pour l'interface moderne de Windows (ex Metro app) Windows Phone, Android et iOS.

Conserver l'attrait de Skype dans un contexte de concurrence accrue

Désormais propriété de Microsoft, Skype avait commencé à facturer les appels groupés en janvier 2011.

Pour utiliser cette fonctionnalité, au moins une personne participant à l'appel devait avoir souscrit un compte Skype Premium (8,04 euros par mois en TTC, ou 4,01 euros par jour pour le forfait à la journée).

Ce "retour" au Freemium, pour une fonctionnalité essentielle à l'attrait de Skype, tient sans doute compte de la concurrence de Google Hangouts qui permet des appels vidéo gratuits rassemblant jusqu'à 10 participants.

En savoir plus

L'annonce de Skype

Crédit photo : Skype

(TITRE:Envoyer un mail au père Noël:TITRE)

Chaque année, des centaines de milliers de lettres sont envoyées en France au Père Noël. A l'ère du 2.0, le site du secrétariat du père noël permet aux enfants d'envoyer un mail au héros de cette fête.

Pour augmenter la magie des fêtes de fin d'année, pourquoi ne pas offrir aux plus jeunes la possibilité d'envoyer un mail au père noël ? C'est ce que propose le site du secrétariat du père noël. Musique douce et flocons de neige vous attendent dès votre entrée.

Avant d'écrire son message, il faudra remplir un formulaire d'adresse car une surprise est à la clef. Ensuite, place à la rédaction. Les plus jeunes devront faire appel à un adulte de leur entourage pour dicter leur message.

Listes de cadeaux, poésies : les enfants ne manqueront certainement pas d'imagination pour amadouer le bon vieil homme barbu afin qu'il les gâte. Une bonne occasion pour les parents ou autres membres de la famille de découvrir les bons cadeaux à déposer sous le sapin.

Précision importante : l'envoi d'email n'est possible que jusqu'au 20 décembre.

Jeux en ligne et carte de remerciement pour chaque mail envoyé.

Sur l'interface du site, plusieurs icônes offrent la possibilité de s'amuser à des jeux variés : point à relier pour faire un dessin, décoration de scènes...de quoi faire patienter vos chérubins jusqu'au réveillon.

Mais la véritable magie de ce site est qu'il permet d'envoyer à chaque enfant une lettre de remerciement signée par le père noël et accompagnée de décorations en papier. Tout cela grâce au formulaire d'adresse rempli en ligne. De quoi faire pétiller les yeux de surprise et de joie ! Pour visiter le secrétariat du père noël

… …

1.3. Ciao_Commentaire_Electromenager

(TITRE:SAV Darty Electroménager:TITRE)

Suite a un incendie de ma cuisine j'achette une table a induction et un réfrigérateur américain afin de remplacer un congélateur bahut et mon vieux réfrigérateur étant des modèle d' exposition la réduction parait alléchante mais correspond a la prolongation de garantie a 5 ans je trouve que celas ne coute pas très chère a dartypour ce qui est de la livraison j'ai attendu un semaine et demi pour un modèle d'exposition a un kilomètre de chez moi de plus l'or de l'enlèvement de mon ancien congélateur bahut par l'aide au livreur de Darty Asnières celui-ci ma voler un pèse personne qui se trouvais a proximité du congélateur en le mettant dedans le pire s'est qu'il a fait ça devants moi et que je m'en suis rendu compte le lendemain il vas surement perdre sa place de travail car je vien de déposer plainte au commissariat de police

(TITRE:Rowenta RO4723.11:TITRE)ce produit est niquel! vous pouvez passez l'apirateur chez vous (que ce soit le matin tres tot ou alors le soir tres tard!) alors qu'une personne dort dans la pièce d'a coté! sans la reveillé! tous le monde adore écouter de la musique en fesant le menage mais le bruit de l'aspirateur couvre toujours la musique, mais la c'est possble!! cet aspirateur a un tres bon raport qualté-prix! quand on goute au silence de cet aspirateur il est impossible d'en changer! vous adorerez sont disigne unique et des couleur qu' aucun autes ne vous apporteras! sa force d'aspiration est exceptionel et son bout triangulaire est tres pratique et peut aller dans tout les coins et recoin de votre maison!

Acheter le vous ne serez pas deçu!

(TITRE:DeLonghi BCO 260 CD:TITRE)

Grands amateurs de café a la maison, nous devions changer de cafetiere puisque je venais de casser la verseuse de notre derniere cafetiere, nous avons donc decidé d'acheter un combiné cafetiere expresso et notre choix s'est arreté sur le combiné delonghi BCO 260. parce que nous avions envie de pouvoir préparer des expressos, mais que lorsqu'il y a du monde, une cafetière c'est quand meme plus pratique.

Nous nous sommes donc rendu chez conforama roanne, et nous sommes repartis avec le combiné DELONGHI BCO 260 et en prime 2 paquets de dosettes souples type ESE de la marque LAVAZZA, ce café est genial....tout cela pour 129 euros.

La couleur du combiné semblait noire, mais la vendeuse nous a assuré qu'elle était bleue et que c'était un effet d'optique dut aux néons du magasin, mais finalement, a la maison, elle semble tout aussi noire, et c'est tant mieux.

Ce combiné est definitivement très pratique, il affiche 15 bars de pression, c'est ce qu'il faut compter si vous voulez un expresso bien mousseux, il est vendu avec deux tailles de percolateurs le premier pour un grand café ou deux petits et le deuxieme pour un petit café serré et ses deux grands réservoirs d'eau amovibles (un pour le café cafetiere et l'autre pour les expressos).Et là, j'insiste sur "grands", car a l'inverse d'une senseo, on est pas obligé de remplir le reservoir expresso plusieurs fois par jours .

Le porte filtre avec filtre permanent en fil d'or est aussi très facile d'utilisation. vite enlevé, vite remit en place. facile à nettoyer...

Mes amis et ma famille viennent souvent chez moi et on se retrouve devant un bon expresso ou un capuccino selon les gouts on en profite puisqu'on peut adapter les dosettes souples de type ESE alors maintenant il y en a pour tous les gouts, café amande, cappuccino, expresso,etc.....on peut meme utiliser la buse vapeur pour faire de la mousse de lait ou du thé elle aussi est très facile à nettoyer.

Franchement, je l'adore et maintenant que je l'ai j'aurais vraiment du mal a m'en passer.

Si vous deviez investir dans une machine a café, je vous conseille celle la, je l'ai acheté il y a six moi maintenant et si c'etait a refaire je le referais sans aucune hesitation.

(TITRE:DeLonghi BCO 260 CD:TITRE)

Je suis une fan d'expresso mais j'aime aussi le café long !! alors je me suis lancée, le compromi, une machine deux en un !!!j'ai cherché et comparé divers produits de différentes marques. Mon choix s'est arrêté sur la Delonghi BCO 260 /CD.

Pas mal, le café est bon.. C'est un bon début....

La couleur est ,soit disant, bleue. Je dis soit disant car en fait, elle fait plutôt noir ... le bleu n'apparaît que si la cafétière se trouve en pleine lumière du jour !!! Un bleuté assez noir en fin de compte , Pas terrible...bref, passons la couleur !!!

A l'usage, je la trouve très pratique, grâce à ses 2 réservoirs amovibles (un pour le café long et l'autre pour les expressos). on les prends, on les remplit au robinet et hop, on les remet en place. très facile , faut juste faire attention de ne pas en renverser au passage et aussi de ne pas soulever le capot des réservoirs avec force car les petites attaches qui le tiennent cassent assez facilement (j'ai bien-sûr cassé un coté !)le porte filtre est aussi très facile d'utilisation. vite enlevé, vite remit en place. facile à nettoyer... mais attention, il y a dedans une petite pièce qui sert pour le stop goutte... et si vous ne faîtes pas attention (comme moi) en jetant votre filtre à café, vous pouvez le mettre en même temps à la poubelle. Le voici donc avec un porte filtre qui n'a plus la fonction stop goutte !! je me suis rendue chez le magasin qui me l'avait venue et oh, surprise, pas de pièces détachées pour le porte filtre !!!!je dois donc faire avec !!!

Bien dommage tout cela ..qu'importe, j'aime ma cafetière, je suis fière d'inviter mes amies à boire un expresso et des bons capuccinos car elle fait même la mousse de lait (trop bon !!!!),, grâce à sa buse vapeur très efficace et aussi très facile à nettoyer.

Mais voilà, la poignée de ma verseuse se détache du bol (non, non, elle ne casse pas complètement, elle se détache juste de la partie haute de la verseuse...nutile de vous dire où je suis retournée ?? ingo, au service après vente !!!!t là, surprise totale : On ne peut pas remplacer la verseuse. raisons évoquées : Les pièces détachées ne sont pas fabriquées pour ce genre d'articles (à savoir ma cafetière) qui vaut moins de 150 euros !!! Cela revient trop cher au fabricant !!le vendeur me dit donc de racheter une autre cafetière, ou de faire avec et d'essayer de trouver une verseuse similaire !!!! vrai info ou belle intox, à vous de juger....

J'ai donc fait le choix de trouver sa jumelle mais impossible de trouver une verseuse qui corresponde à sa forme et à sa hauteur !!!e voici donc belle et bien coincée !!!'était sans compter sur ma fabuleuse colle extra forte !!!j'ai recollé la poignée tout simplement. Cela m'a couté un tube entier de colle extra forte en gel et le tour est joué !!! (moins cher au final qu'une verseuse neuve)depuis, tout va à merveille !!! et en plus, cela ne se voit presque pas !!!!

Bon, hormis ces petits défauts de pièces, il faut bien avouer que le café est bon, que les performances même de la machine sont pour moi de très bon rapport qualité-prix.

J'invite toutefois la Société Delonghi à travailler sur les accesoires et leur solidité !!!pour le reste, c'est plutôt pas mal. Je la possède depuis un moment désormais et elle tourne tous les jours !!!

… …

1.4. Doctissimo_Forum

(TITRE:Protection des mineurs. - Aide sur le fonctionnement des forums - FORUM Santé:TITRE)
Modératrice5rofil : équipe de Modération posté le 07-03-2014 à 13:57:15 bonjour à tous, je m'occupe désormais de la protection des mineurs sur le forum de doctissimo.fr .i vous rencontrez une quelconque difficulté, n'hésitez pas à m'envoyer un message privé
Voici quelques conseils qui pourraient vous être utiles:http://blog.doctissimo.fr/mode [...] rs-672359/
Tes parents doivent savoir que tu t'es inscris sur nos forums et te permettre d'y surfer.
Il est en outre nécessaire que tu donnes ton âge réel.
Ne communique jamais des informations qui te sont personnelles :- Ton nom de famille- Ton prénom- Ton numéro de téléphone- Ton adresse- Ton école- Ta villeorsque tu communiques ton adresse mail par message privé, fait bien attention à ce que cette adresse ne comporte aucun renseignement personnel te concernant (comme ton nom de famille et ton prénom).
Bien qu'une personne puisse te sembler sympathique au premier abord, ne la rencontre pas ! En effet, une personne mal intentionnée peut se cacher derrière un pseudo et s'inventer une histoire, un âge, une vie ! Aussi, si une personne te propose un rendez vous parles en d'abord à tes parents et ne t'y rends jamais seul, fais toi accompagner.
Fais attention à bien choisir ton interlocuteur lorsque tu envoies une photo de toi.
Il est facile de copier-coller cette photo et la diffuser un peu partout sur internet sans ton accord.
Ton corps t'appartient, aussi, personne n'a le droit de te forcer à faire quelque chose qui te met mal à l'aise. un internaute te fait une proposition gênante, embarrassante (comme un proposition pour parler sexe sur Skype), contactes moi . Pour cela, il suffit de m'envoyer un message privé, où de cliquer sur l'alerte mineur. Il faudrait aussi dans ce genre de cas, parler à un adulte, ou à une personne en qui tu as confiance.
Si tu tombes par inadvertance sur une image ou une vidéo qui te choque, ferme ton ordinateur et parles-en à un adulte. Il pourra nous contacter à l'aide de l'alerte mineure ou encore à l'aide de message

privé pour qu'on puisse effacer le plus rapidement possible cette image ou vidéo. tu dois savoir en outre que tu es responsable pénalement du comportement que tu pourrais avoir sur nos forums

N'hésite jamais à m'envoyer un message privé, même en cas de doute !on forum,

(TITRE:Docti se connecte tout seul - Aide sur le fonctionnement des forums - FORUM Santé:TITRE)

Profil supprimé posté le 03-06-2014 à 07:38:09 bonjour iphone, l'appli docti se connecte toute seule! Je me déconnecte et 1 h après c'est connecté sans avoir rien fait (ni taper mot de passe ni rien!) C'est très embêtant et ça le fait à plusieurs personnes de ma connaissances! Une idée pour que ça ne le fasse plus? bonjour, je l'avais déjà signalé, il y a plusieurs mois... mais effectivement, rien n'a changé ! 'est donc l'inverse du docti sur ordi qui se deconnecte sans qu'on le fasse, ni qu'on le souhaite, surtout quand on est en train de poster. De plus en plus curieux ce qui se passe ici

Docti est gouverné par la 4ème dimension... https://www.youtube.com/watch?v=55KH1ei-WD0 j'aime autant cette 4° dimension ci :http://www.youtube.com/watch?v=fl98PNQm2O4

Notre jeunesse! ça en rappelle des souvenirs n'est ce pas ?

Oui!

(TITRE:Pub qui déconne et lecture impossible - Aide sur le fonctionnement des forums - FORUM Santé:TITRE) avrilette2010

Inna Allah m3a sabirinrofil : octinaute Hors Compétitionosté le 04-04-2014 à 10:57:46 jrur ramadanettes y'a la pub garnier pr cheveux crépu qui se fout en pleine page TOUTE BLANCHE (donc ça déconne) rien ne s affiche et on peut meme plus lire les titres des sujets. Ca rend la navigation sur le site impossible, merci de bien vouloir rectifier! 'ai eu ça aussi c'était des voitures. Oh la crise. et ça se renouvelle aussi

(TITRE:endométriose, Lutéran, Gattilier et Achilée - Endométriose - FORUM Santé:TITRE)

DJ6574osté le 26-04-2014 à 16:56:47 bonjour, je souffre de douleurs pelviennes depuis un moment qui se sont intensifiées depuis environ 1 an. J'ai été mise sous Lutéran 10 avant de passer une IRM qui a permis de poser le diagnostic : endométriose à "un stade débutant" (mais douleurs bien intenses tout de même), avec adhérences dans l'utérus, nodule dans le col de l'utérus et surtout adhérence sur le ligament utéro sacré. epuis, je continue le Lutéran mais j'ai parfois tout de même des douleurs (surtout au moment des rapports sexuels où la douleur est vraiment très très intense et le soir, avant l'heure de la prise du Lutéran). J'ai découvert depuis peu l'Achilée Millefeuille que jachète en magasin bio et qui, je crois, me soulage (du moins + que l'antadys et compagnie). aujourd'hui une amie m'a parlé du

Gattilier. Cette solution me tente pas mal car l'idée de prendre encore longtemps un traitement chimique tel que le Lutéran me dérange un peu, surtout que vue la galère pour diagnostiquer mon endo, les messages que je vois de filles opérées X fois qui ont toujours mal etc, je n'ai plus vraiment confiance en les médecins et les méthodes traditionnelles. certaines d'entre vous ont-elles essayé ces méthodes? J'ai lu qu'il était conseillé d'arrêter tout traitement hormonal avant de commencer vraiment le Gattilier. Pensez vous que ce soit une mauvaise idée d'arrêter le Lutéran pour passer à Gattilier + Achilée Millefeuille ? je n'ose pas poser la questions à ma gynécologue car je connais d'avance sa réaction. merci à celles qui sauront m'éclairer huile d onagre aussi m aide bcp le reste pas encore essayee

(TITRE:Informations sur le Don d'ovocytes en France - Entraide : mamans en difficulté - FORUM Grossesse bébé:TITRE) enfantmagique1

Profil : doctinaute d'argent posté le 06-01-2012 à 13:07:14 bonjour à toutes, pour celles qui sont déjà maman, vous pouvez permettre à une femme de connaitre un jour ce bonheur....pour savoir comment ça se passe en France, vous pouvez visiter le site d'information que j'ai créé, dans ma signature....l faut :- être maman- être majeure et avoir moins de 37 ans- être en bonne santé- avoir l'accord de son conjointe suis disponible pour toute question, ici ou en MP. bientôt up uel est l'age limite ? Quels sont les effets des traitements ? ondulee a écrit :uel est l'age limite ? Quels sont les effets des traitements ? Bonjour Ondulée, pour donner ses ovocytes il faut :- avoir déjà au moins un enfant- avoir entre 18 ans et 36 ans et 364j maximum- avoir l'accord de son conjoint, je copie-colle un paragraphe qui provient du site du GEDO : Les risques et inconvénients

Les traitements pour le don d'ovocytes sont pratiqués à large échelle et après de nombreuses années dans le cadre de la FIVintra-conjugale.

Cependant, comme pour tout acte médico-chirurgical, d'exceptionnelles complications demeurent possibles (pesanteurpelvienne, légers saignements, réponse excessive des ovaires à la stimulation ; de façon rarissime, problèmes anesthésiques, hémorragiques ou infectieux).a prévention de leur apparition ou leur survenue nécessite un encadrement spécialisé rigoureux de chaque tentative.

… …

1.5. Internaute_Forum_Bricolage

(TITRE:Quel est votre avis sur la marque Beko ? [Résolu]:TITRE)

Bonjour, j'envisage de refaire ma cuisine et par la meme de changer mon électroménager. J'aurais voulu votre avis sur la marque Beko. J'envisage d'acheter un lave vaisselle ainsi qu'un frigo (congelo en bas). J'ai pu voir que leurs prix étaient trés (trop ???) compétitif et je voudrais notamment savoir si la qualité était au rendez vous (merci aux pros qui fréquentent le forum :-)) et si on pouvait les faire réparer en cas de panne (ou si trouver la carte / moteur en panne était mission impossible).

Merci à tous pour vos conseils et expériences autour de cette marque.

Stephanebonjour,

si tu cherches un prix, tu prend ce que tu veux si tu cherches un bon rapport qualité/ prix, tu prends une autre marque .... quand les pieces ne sont pas disponibles on peut les attendre longtemps ..... très très longtemps !!!!

Bonjour La Sudiste,

merci pour ta réponse... Ce que je dois en comprendre entre les lignes, c'est que la qualité n'est pas au rendez vous avec Beko ?

Merci,

Stephaneje ne suis pas la pour critiquer telle ou telle marque, je conseille juste de prendre une autre marque, style BOSCH

Bonjour,

Je vous trouve un peu dur avec la marque BEKO,

Personnellement, j'ai un lave vaisselle de cette marque depuis 1 an, et je n'ai encore pas rencontrer de problème. Il est vrai que cette marque n'est pas cher et permet aux modestes ménages de s'équiper pas trop cher. D'ailleurs j'ai rarement acheter de la marque (trop cher parfois).

Maintenant j'ai acheter un fer à repasser Delonghi (bonne marque à priori) il a fait 6 mois !!! Je pense qu'on peut tomber très bien comme on peut tomber très mal, même avec une bonne marque.

Je trouve que ce n'est pas facile de donner son avis, à moins d'avoir eu de réel expérience. Bonsoir davkar,

Je ne peut dire que la sudiste est la spécialiste électroménager de ce site. Niveau expérience et je ne la connait pas personnellement, seulement au travers des réponses qu'elle donne sur ce forum, je pense

qu'il n'y a pas photo. Si vous relisez ses réponses, vous verrez qu'elle essaie d'être objective. Vous avez raison, les appareils de cette marque sont trés compétitifs au niveau achat et, c'est trés bien quand on a peu de moyen. Pour autant, le SAV n'est pas au RdV, c'est juste ce qu'indique la sudiste à steph, qui demandait un avis avant achat. On ne peut que vous souhaiter que votre lave vaisselle fonctionne trés longtemps, BEKO ou pas.

Bonne soirée.

Bonjour, je possède quant à moi un congélateur Beko depuis 4 ans, je n'ai jamais eu de souci particulier avec, sauf sur un point: c'est un congélateur à grands tiroirs (ce que je voulais) mais bien chargés j'ai fendu un tiroir et cassé une porte... Un petit peu de colle et c'est reparti pour un tour. Pour le SAV, je ne peux rien en dire.

Cordialement

beko a eviter ,2frigo en panne a coté de chez moi et pas réparable selon les vendeurs , et ils n'ont pas encore 2 ans

Bonjour Daniel, ce sont des vendeurs qui vous ont dis qu'il n'était pas réparable, mais est ce que vous l'avez montré a un dépanneur ... parce qu'un vendeur reste un vendeur ... ce n'est pas un dépanneur ... chacun son métier !!!la nunuche

bonjour

j'ai acheté le 12 février dernier un frigo congélo, c'est une cata.

le magasin me mène en bateau en prétendant que la société Beko va me contacter: ce qui est complètement faux. Actuellement il fait entre 20 à 22 degrés dans le frigo; au motif que le gaz est parti. très drôle. pas moyen de conserver des aliments avec une température pareille. je déconseille très fortement. le prix ce n'est pas tout. certes la marque est compétitive mais le produit est "nul".

a éviter absolument

Pouvez vous contacter le service après vente au 03/890.86.90 afin de résoudre au plus vite votre problème. Cordialement

Ca c'est ce que j'ai comme numéro client: 01.58.34.46.46 . Si sav beko pouvait nous dire pourquoi un numéro en Belgique a priori.

(TITRE:Probleme de frigo:TITRE) indesit type df01x mod R34 j'ai changé la sonde car le frigo gelait tous les produis et le moteur ne s'arrêtait pas, il ne s'arrête toujours pas et il descend à-3 voir -5 .Merci

Bjr, peu tet que la sonde est mal mise? donne la tempé du congélot?? et la bonne référence du frigot??ref 93231630000 S/N 110152 89 je l ais brancher sur un minuteur sinon le compresseur sarette jamer +je tourne la molette sur 0 l ampoule seclaire et setain pas meme la porte fermer

bonjour, vous n'avez certainement pas remplacé la sonde mais le thermostat (son capillaire ne peut être séparé du thermostat) le capillaire est enfiché dans le fourreau prévu a cet effet un petit adhésif rouge indique la profondeur qu'il doit atteindre si vous avez respecté le positionnement des fils sur le dit thermostat c'est que dans ce cas le frigo a un décollement de paroi et de fait,

l'appareil serait irréparable (le conditionnel étant le respect du schéma de câblage) l'utilisation d'un minuteur n'étant qu'une solution provisoire qui ne pourra perdurer hélasle thermostat que jais achetter na pas d adhesif de profondeur je laimie en buttee dans son fourreaure,

avez vous dans ce cas effectué le câblage a l'identique des numéros gravés sur le thermostat??

pour info un thermostat qui ne coupe pas et fait fonctionner en continu un compresseur est jamais en cause même si le frigo gèle sauf s'il ne coupe pas sur arrêt la cause étant une température de détente(il faut -30°C) insuffisante les causes sont toutefois diverses , bouchage partiel , décollement de paroi ,compresseur qui faibli clapets hs

bref , cela est très technique et de la compétence uniquement d'un frigoriste

(TITRE:Frigo ariston:TITRE)

Bonjour,

Le moteur de mon frigo fonctionne, mais ne refroidi plus y a t il un réparateur en belgique ?bonjour,

vous devez avoir plein de dépanneur en Belgique, mais j'ai bien peur qu'ils vous annoncent la mort de votre appareil ... car si celui ci fonctionne en continu (je parle du moteur) mais qu'il ne fait pas de froid, c'est certainement un problème sur le circuit frigorifique ....

(TITRE:Comment changer thermostat frigidaire Electrolux ?:TITRE)

Mon frigidaire Electolux tourne tout le temps . Il y a du givre à l'intérieur ;j'en ai déduit que le thermostat était HS mais je ne sais pas le changer. Il est situé sur le bandeau en façade du frigo . Il y a un grand fil en cuivre qui en part et je ne vois pas où il aboutit ...

Merci à tous bonjour, le fil que vous appelez est en fait la sonde du thermostat, qu il faut remettre dans son emplacement qui va dans la paroi_ du fond du frigo bonjour à tous j'ai un soucis avec mon combiné electrolux : le congelatuer fonctionne toujours alors que l frigidaire ne fait plus de froid ; il a deux ans seulement. alors je suis un peu dans la panade !

merci de me' aider avnt une éventuelle canicule !bonjour,

donnez la référence ou dites nous si il y a un seul ou 2 moteurs !! si le compresseur fonctionne en continu, si il est chaud ... etc ..

Bonjour,

Est-ce un froid ventilé ? L'avez-vous dégivré suffisamment ?

… …

1.6. MeilleurDuChef_recettes_de_cuisine

(TITRE:Aspic à la mousse d'écrevisses - Recette de cuisine - MeilleurduChef.com:TITRE)

Phases techniques de la recette :

Napper le fond des assiettes de service d'une couche de gelée de poisson. Laisser prendre au froid. Décortiquer les écrevisses préalablement cuisinés "à la Bordelaise". Observer les queues et quelques têtes avec les pinces.es couper en deux dans le sens de la longueur et retirer le boyau central. Réserver 1 queue d'écrevisse par aspic réalisé. Réunir dans un mixer les queues d'écrevisses et la brunoise de légumes ayant servie à faire la sauce bordelaise. Mixer finement. Débarrasser dans un cul de poule et ajouter la gelée froide. Remuer. Ajouter la crème fleurette légèrement fouettée. Remplir des moules à darioles de mousse aux écrevisses. Laisser prendre au frais. Dresser sur les assiettes couvertes de gelée. Décorer avec quelques pluches de cerfeuil.

(TITRE:Choux fourrés à la mousse de foie gras - Fiche recette avec photos - MeilleurduChef.com:TITRE)

Phases techniques de la recette :1

Pour réaliser cette recette de choux fourrés à la mousse de foie gras, commencer par préparer tous les ingrédients.2

Préparer la pâte à choux, pour cela confectionner la panade34

Coucher les choux à la poche à douille munie d'une douille unie, sur une plaque à pâtisserie légèrement graissée. Passer la dorure sur chacun d'eux, à l'aide d'un pinceau...567

Sauce béchamel :

Faire fondre le beurre dans une sauteuse8

Ajouter la farine en une seule fois.9

Bien mélanger au fouet et cuire quelques minutes à feu modéré.10

Il faut que la farine se mélange parfaitement au beurre. Attention à ne pas laisser de grumeaux de farine et à ne pas faire brunir le roux blanc, qui comme son nom l'indique, il doit rester blanc. Laisser tiédir.11

Porter le lait à ébullition.12

Verser le lait bouillant sur le roux blanc froid...1314

Mousse au foie gras :

Préparer tous les ingrédients.15

Tailler le foie gras en morceaux et les placer dans un hachoir.16

Ajouter le beurre pommade puis la sauce béchamel froide. Compter entre 100 et 200 g de sauce béchamel pour les quantités indiquées ci-dessus. La quantité dépendra de la texture finale de la mousse que l'on voudra obtenir.17

Assaisonner de sel fin et de poivre du moulin. Mixer finement.18

Débarrasser et réserver au frais jusqu'au moment de l'utilisation.19

Ici j'étale ma préparation sur un papier film en une fine couche de façon à ce que mon mélange refroidisse rapidement, reprenne de la texture et soit utilisable rapidement.20

Dressage :

Couper les choux en deux avec un couteau à dents.21

Garnir les choux à l'aide d'une poche à douille munie d'une douille cannelée. Décorer en plantant les parures de foie gras découpées en bâtonnets.

(TITRE:Mini-quiches au saumon fumé - Recette de cuisine avec photos - MeilleurduChef.com:TITRE)

Phases techniques de la recette :1

Préparer tous les ingrédients.2

Faire une pâte brisée3Éplucher les poireaux et ne conserver que le blanc. Les laver et les émincer finement.4

Les faire suer au beurre, à feu doux, pendant 8 à 10 minutes.5

Abaisser la pate brisée...6foncer un moule à mini-tartes en silicone.7

Dans un cul de poule...89

Ajouter les poireaux...1011

Garnir les fonds de tartes avec cette préparation.12

Cuire à four chaud, 180°C...1314

Au terme de la cuisson...15

(TITRE:Courgette farcie - Recette de cuisine avec photos - MeilleurduChef.com:TITRE)

Phases techniques de la recette :1

Préparer tous les ingrédients.2

Couper les deux extrémités des courgettes.3

Les cuire 10 minutes dans une eau bouillante salée.4

Les rafraîchir et les égoutter.5Évider l'intérieur de chaque courgette. Réserver la pulpe.6

Monder, peler, épépiner et concasser les tomates.7

Farce : faire revenir les oignons hachés à l'huile d'olive.8

Ajouter la tomate et la chair à saucisse.9

Cuire quelques minutes...10

Pendant ce temps, hacher les champignons de Paris et les faire suer à part dans une sauteuse.11

Rassembler la chair à saucisse cuite avec les champignons cuits...1213

Bien mélanger à la spatule en bois.14

Garnir l'intérieur des courgettes de la farce. Disposer dans un plat à gratin.15

… …

2. Extraits des corpus de noms de métiers

2.1. BlogCarriere_Actualites

(TITRE:Recherche d'emploi et développer son réseau professionnel id-carrieres Le Blog:TITRE)

Vous vous interrogez sur la façon d'animer votre présence sur les réseaux sociaux pour rechercher un emploi ou une opportunité d'évolution ou encore pour développer votre réseau professionnel ?

Nous avons retenu cette infographie initialement destinée aux marketers car à minima 20 conseils sur les 21 proposés peuvent être repris par chacun d'entre nous pour développer une démarche Réseau Social efficace mais aussi équilibrée et motivante.

Le conseil n°20 relatif à la tenue d'un blog doit être évalué par chacun en raison de l'engagement personnel qu'il constitue (aptitudes, disponibilité, plaisir...).

A relever que ces conseils sont aussi en grande partie les bons réflexes d'une démarche Réseau off line. Bonne lecture en images ! source Infographie

Partager la publication "Recherche d'emploi Démarche réseau 20 conseils simples de marketing social..."

FacebookwitteroogIe+interestiadeoinkedIn-mail

(TITRE:Recherche d'emploi et réseaux sociaux : le pouvoir de l'annonce id-carrieres Le Blog:TITRE)

Les candidats ont toujours exprimé leur insatisfaction par rapport aux méthodes de recrutement par annonce : trop subjectives quand l'artisan recruteur sélectionnait après une lecture de tous les CV papiers, trop aléatoires quand les outils automatiques de gestion des candidatures sélectionnent les CV à partir de mots-clés... La démarche Réseau et aujourd'hui les réseaux sociaux numériques offrent la possibilité de rechercher un emploi autrement, et pourtant les candidats ne l'utilisent pas. Ils nous disent même parfois ne pas y penser spontanément...

Ce que nous disent les candidats

Les candidats restent très attachés à la candidature suite à une annonce d'offre d'emploi.

Ce que nous disent les candidats :« l'annonce, c'est la garantie d'un poste à pourvoir, d'une réelle opportunité d'emploi. » « l'annonce permet d'orienter ma candidature, de mieux l'argumenter à partir des attentes exprimées. » « le recrutement par annonce, c'est plus rapide. » « la procédure pour candidater est précisée dans l'annonce, en général une lettre de motivation et un CV. »

L'annonce répond aux besoins des candidats ?

En dépit des insatisfactions exprimées, les attentes du candidat, les annonces semblent répondre au besoin de sécurité des candidats : des informations précises, des règles à suivre, des normes connues. Un besoin de repères dans une période où d'autres repères s'effacent par contrainte (perte d'emploi) ou par choix (souhait d'évolution).

Les repères que fournissent les annonces répondent aussi au besoin de se sentir compétent dans sa recherche d'emploi, de bien conduire sa recherche d'emploi en utilisant des méthodes qui, bien que questionnées, perdurent. Une pérennité perçue comme une preuve de leur efficacité. L'annonce, le CV, la lettre de motivation perdurent aussi parce qu'ils apparaissent comme une norme partagée par un collectif composé des autres candidats, et des recruteurs !

L'annonce, les réseaux sociaux et les besoins des candidats ?

La démarche Réseau dans laquelle s'intègre les réseaux sociaux numériques est généralement perçue comme moins rassurante.

Le candidat projète la démarche comme plus incertaine en termes de résultats. Socialement plus risquée. Socialement moins valorisante aussi (position de demandeur).

La démarche Réseau semble ne pas répondre au besoin de sécurité de la plupart des candidats, l'objectif d'un rendez-vous n'est pas aussi précis que la réponse à une annonce pour un poste. Les modalités sont à réinventer à chaque contact... Les candidats nous disent très souvent ne pas se sentir capables de développer une démarche réseau. Les arguments présentés le plus souvent étant qu'ils n'ont pas de réseau, qu'ils n'ont jamais fait de réseautage, qu'ils n'ont pas appris à le faire, qu'ils n'ont pas de plaisir à l'idée de le faire...

La perception d'une absence de contrôle sur le résultat et d'une absence de contrôle sur le déroulement de la démarche Réseau génèrent une réticence fréquente à l'adopter, à l'expérimenter.

Si on compare la démarche Réseau à la démarche Annonce, le point commun serait objectivement l'absence de contrôle sur le résultat !

La candidature à une annonce est largement soumise à l'évaluation du sélectionneur ou d'un outil, à la décision du recruteur. Les candidats semblent néanmoins davantage envisager la candidature à une annonce et son argumentation (CV, lettre de motivation, entretien) comme un élément d'influence sur le résultat de la sélection. Ce qui est aussi valable pour la démarche Réseau : l'échange permet un ajustement plus pertinent et plus ouvert de l'argumentation, donc une influence plus grande sur le résultat...

L'absence de contrôle sur le déroulement de la démarche Réseau en comparaison de la candidature à une annonce est objectivement valable, l'annonce étant davantage normée que la première. La démarche Réseau relève plutôt de bons réflexes et de pratiques éprouvées que de modes opératoires et règles valables en tout temps et en tout lieu...

Les trois études conduites en 2012 par decco

RegionsJobou

IPSOS démontrent la difficulté des réseaux sociaux à s'imposer comme une tendance lourde des pratiques de recherche d'emploi alors qu'ils semblaient répondre aux attentes intrinsèques des candidats pour un recrutement davantage personnalisé et ouvert.

L'annonce semble bénéficier par ses caractéristiques d'un certain pouvoir sur la motivation des individus à l'utiliser parce qu'elle répond à certains de leurs besoins : sécurité, sentiment de compétences, sentiment d'influence... et parce qu'elle est aujourd'hui une norme partagée collectivement.

Elle peut être aussi le signe d'une certaine résignation des candidats...

Nous aborderons dans les prochains jours, les conditions pour une adoption plus large des usages sociaux dans la recherche d'emploi, convaincus de leur efficacité pour les chercheurs d'emploi et pour les recruteurs et de leur meilleure adéquation aux enjeux et aux besoins des salariés et des employeurs. Convaincus qu'ils constituent à ce jour la piste d'innovation la plus contributive à une transformation des pratiques de recrutement y compris par annonce...

A lire aussi sur le sujet :

Le recrutement aurait-il raté l'opportunité des réseaux sociaux ?

Le dossier : le recrutement se transforme-t-il ?

Partager la publication "Recherche d'emploi Réseaux sociaux Le pouvoir de l'annonce ?"

FacebookwitteroogleinterestiadeoinkedIn-mail

(TITRE:Les 30 sites d'annonces d'emplois id-carrieres Le Blog:TITRE)

Le site L'Etudiant.fr a publié une analyse comparative d'une trentaine de sites d'emplois en France. L'étude mentionne le nombre d'offres d'emplois proposées et évalue la qualité du site à partir de 2 critères clés pour les personnes en recherche d'emploi ou en veille : la « fraicheur » des annonces et la pertinence du moteur de recherche à partir des mots clés.

Vous trouverez sur le site de l'Etudiant.fr associé à chaque site d'emploi analysé des informations complémentaires sur le positionnement des offres et son fonctionnement.

La liste n'est pas exhaustive. Pensez aussi à utiliser les moteurs d'offres d'emploi comme obiJoba ou imply Hired

Compte tenu de la diversité des libellés d'emplois et des contenus de missions, il faut analyser régulièrement les résultats obtenus et tester d'autres mots-clés pour vos alertes Emplois.

Partager la publication "30 sites d'emplois évalués. Pensez aussi aux mots-clés de vos alertes"

FacebookwitteroogleinterestiadeoinkedIn-mail

… …

2.2. BlogEmploi_Annonce

(TITRE:Mode(s) d'emploi - Virage Conseil : Merchandiseurs H/F:TITRE)

Merchandiseurs H/F

Entreprise :

Virage Conseil

Contrat :

CDD

Localisation :

Corrèze - 19

Virage Conseil recrute des Merchandiseurs H/F en CIDD.

Virage Conseil est une société spécialisée dans l'externalisation des services du développement des ventes : Force de vente, animation commerciale, merchandising, télémarketing. Dans le cadre de sa croissance, nous recherchons des Merchandiseurs H/F en vue de mettre en avant des produits de grande marque.es 400 collaborateurs de Virage Conseil sont aujourd'hui au service de plus de 250 marques nationales et internationales. Descriptif du poste :es postes en CIDD sont à pourvoir de manière permanente sur toute la France. Vos missions seront les suivantes :- Gestion de fond de rayon.- Théâtralisation.- Montage de tête de gondole.- Mise en avant des produits, lutte anti-rupture.- Gestion des mises en place des produits dans le respect des préconisations merchandising du client. Vous êtes ponctuel(le), sérieux(se) et motivé(e) pour travailler en binôme avec les commerciaux de nos différents clients ou en autonomie totale.

Profil recherché et compétences clés :

Issu(e) d'un Bac ou d'une bonne expérience en merchandising.es circuits GMS/GSS/GSB, et le merchandising vous sont familiers. Sensible aux produits, à la valeur des marques, à l'innovation, vous êtes capable de prendre des initiatives et de vous adapter aux demandes spécifiques de nos clients.

Salaire

Non précisé.

Adresse postulez ! Vous souhaitez intégrer un groupe dynamique au service des plus grandes marques, rejoignez-nous ! Contacts : s.benlarbi@virageconseil.com - w.juan@virageconseil.com

Référence : MERCHANDISEUR. http://www.virageconseil.com publié le 25/04/2014

Réf : M/19

(TITRE:Mode(s) d'emploi - Assystem : Ingénieur Assurance Qualité Biotechnologies H/F:TITRE)

Ingénieur Assurance Qualité Biotechnologies H/F

Entreprise :

Assystem

Contrat :

CDI

Localisation :

Ile de France

Assystem emploie près de 11 000 collaborateurs dans le monde et a réalisé un chiffre d'affaires de 871,4 MEuros en 2013.

Dans le cadre de notre développement, nous recherchons un(e) Ingénieur Assurance Qualité Biotechnologies H/F. Pour la réalisation de l'un de nos projets dans les secteurs des Biotechnologies, vous êtes le(la) référent(e) assurance qualité sur un projet d'extension de bâtiment. ce titre, vous êtes en charge des lots EPU, EPPI, AC, autres fluides, procédés de fabrication, systèmes NEP, système Ultrafiltration, automatisme, MAL et autoclave. Vous effectuez :- la revue des cahiers des charges de fournisseurs,- la revue des plans qualité,- la revue de la design qualification,- la revue des protocoles FAT/SAT fournisseur, - la réalisation des FAT/ SAT fournisseur,- l'identification et la revue des procédures impactées par la modification,- la rédaction des documents de stratégie interconnexion bâtiments et l'analyse d'impact,- la revue de Master Plan, - le plan directeur qualification/validation, - les études d'optimisation de validation de nettoyage et de procédé.

Une expertise confirmée :

Issu(e) d'une formation supérieure Ingénieur ou Université de niveau Bac +5, vous disposez d'une expérience minimale de 5 ans sur un projet similaire dans le secteur pharmaceutique. Vous avez l'habitude des relations client, et vous savez vous adapter à des contextes techniques et relationnels variés. Doté(e) d'un excellent relationnel, vous possédez idéalement une expérience dans le domaine des Biotechnologies. Anglais apprécié. Nous vous offrons :ne rémunération annuelle liée à votre expérience. La possibilité de participer à un vrai projet d'entreprise.ne possibilité d'évoluer dans un environnement de travail à la pointe de la technologie.ne entreprise qui saura reconnaître votre implication et vous faire évoluer.

Salaire

Non précisé.

publié le 16/05/2014

Réf : EPIMBL4939_RJ

(TITRE:Mode(s) d'emploi - Ergos Interim : Chauffeur SPL H/F:TITRE)

Chauffeur SPL H/F

Entreprise :

Ergos Interim

Contrat :

Travail temporaire

Localisation :

Miramas - 13

Ergos Intérim recrute pour l'un de ses clients, acteur référent dans le transport, des Chauffeurs SPL H/F.

Livraison régionale de boissons en camions bâchés au départ de Miramas. Vous devrez pour cela :- Décharger et livrer les commandes sur le lieu de réception.- Prendre soin de son outil de travail et assurer la propreté de son camion.- Respecter les délais de livraisons.- Remonter les disfonctionnements issus des livraisons.- Prévenir le service livraison de tout retard ou tout autre problème rencontré dans ses missions.- S'assurer de la satisfaction du client au moment de livraison. Notre permis EC ainsi que votre FIMO, FCO sont à jour. Vous êtes titulaire de la carte conductrice.

- Présentation soignée.- Sens du service, aptitudes relationnelles.- Ponctuel(le)et réactif(ve).

Salaire

Non précisé.

Merci d'envoyer votre cv à marignane@ergosinterim.fr publié le 15/05/2014

Réf : PJ/CS/13M/1573807

(TITRE:Mode(s) d'emploi - Abalone Mantes la Jolie : Monteur Réseaux Electriques Aéro-Souterrain H/F:TITRE)

Monteur Réseaux Electriques Aéro-Souterrain H/F

Entreprise :

Abalone Mantes la Jolie

Contrat :

CDI

Localisation :

Poissy - 78

Abalone, agence d'emploi, vous propose aujourd'hui une très belle opportunité.

Nous recherchons, pour l'un de nos clients, un Monteur Réseaux Electriques Aéro-Souterrain H/F sur le département des Yvelines (78). Poste en CDI à pourvoir immédiatement. Vous intégrez une société créée depuis 1961 et spécialisée dans l'électricité de bâtiment et l'éclairage publique au capital de 500 000 € , société familiale en plein essor. Missions : vous devrez travailler sur le réseau de distribution d'énergie électrique en haute et basse tension, dans des zones rurales ou urbaines. Vous réaliserez l'implantation et la prolongation d'une ligne existante, vous serez en charge de l'augmentation de la capacité de transport de la ligne. Vous interviendrez principalement sur les lignes aériennes de transport de courant haute tension en mettant en place des poteaux ou pylônes, dérouler et fixer les câbles au sommet. Vous serez également en charge de différentes missions de terrassements, manutention er de déroulement de conducteur de tranchée. Vous effectuez également les connexions aériennes et le raccordement entre le réseau aérien et souterrain. Vous mettrez en place le réseau basse tension, et l'installation de l'éclairage public.

Expérience :

Vous avez idéalement une expérience de 2 ans minimum sur un poste similaire. Vous possédez des qualités de technicien(ne). Vous avez une capacité à travailler en équipe sous la coordination d'un chef d'équipe, vous aimez le travail en plein air, vous êtes insensible au vertige, vous êtes mobile sur le département du 78. Alors n'hésitez plus, contactez-nous ! Ce poste est fait pour vous !

Salaire

Fixe entre 1800 et 2400 brut sur 12 mois (à négocier selon profil) + primes de paniers et déplacements

Adresse merci d'envoyer votre candidature à votre agence Abalone à l'adresse mail suivante : mantes@abalone-interim.com publié le 07/05/2014 Réf : PRJ/MREAS/78P/1490835

… …

2.3. RejoinsJob_Actualites

(TITRE:Etudiants étrangers : la circulaire Guéant abrogée - RegionsJob:TITRE)

Un soulagement pour de nombreux étudiants étrangers. Manuel Valls, le nouveau ministre de l'Intérieur a confirmé aujourd'hui l'abrogation de la circulaire Guéant. Un nouveau texte devrait prochainement voir le jour et ainsi éviter aux étudiants de "ne plus être dans cette insécurité à quelques mois de la rentrée universitaire" a déclaré Manuel Valls. Et de poursuivre : "c'est une chance pour eux et c'est aussi une chance pour notre pays".

Quels changements pour les étudiants ?

Pour l'heure, le contenu précis du nouveau texte n'est pas encore connu. Mais selon le président de la Fédération des associations générales étudiantes (Fage), il devrait notamment interdire l'expulsion des étudiants dont le titre de séjour a expiré ainsi que la réduction des délais d'instructions des dossiers par les préfectures.

Une circulaire critiquée

Lors de sa publication, la circulaire Guéant avait provoqué un tollé au sein du monde enseignant et des professionnels. Quelques mois plus tard, pour faire taire les critiques, le gouvernement de l'époque avait alors décidé de l'assouplir. Les étudiants diplômés au moins d'un master et pourvu d'un emploi ou titulaire d'une promesse d'embauche pouvait alors bénéficier "d'un titre de séjour autorisant l'exercice d'une première activité professionnelle", détaillait la circulaire de l'époque. Un assouplissement qui n'avait pas permis de calmer la colère des universitaires.

(TITRE:OPEN JOB DAY : soirée recrutement et baptême d'hélicoptère à Lyon - RegionsJob:TITRE)

Les équipes d'OPEN se mobilisent pour accueillir les candidats lors d'une soirée recrutement qui aura lieu à Lyon le 13 juin 2012. Cet évènement s'adresse aux Ingénieurs d'Études, Concepteurs, Chefs de Projet, Experts techniques en Nouvelles Technologies (JEE, .Net), Ingénieurs Systèmes (Unix / Windows), Intégrateurs, Pilotes d'exploitation, Ingénieurs de production...

(TITRE:Quels métiers recrutent le plus de saisonniers dans le Centre ? - RegionsJob:TITRE)

En France, 39,2 % des projets de recrutement 2014 portent sur des emplois à caractère saisonnier. Parmi les secteurs qui prévoient le plus d'embauches saisonnières dans les régions du Centre, l'agriculture reste en première position suivi du commerce et des services. Une tendance qui, d'une

année sur l'autre, reste la même. Voici les principaux métiers qui recrutent des saisonniers dans le Centre, le Limousin et l'Auvergne.

Centre

(TITRE:Soirée recrutement à Rennes avec les experts de Sogeti - RegionsJob:TITRE)

Sogeti, filiale à 100% du groupe Capgemini, est l'un des leaders des services informatiques et d'ingénierie de proximité, spécialisé dans la gestion des applicatifs et des infrastructures ainsi que le testing. Son expertise se positionne dans les domaines de la Cybersécurité, la Mobilité, la Business Intelligence et les solutions liées à la Transformation des Infrastructures.

(TITRE:La France pourrait manquer de diplômés en 2020 - RegionsJob:TITRE)"40% des jeunes français ont un emploi contre 69% des jeunes néerlandais. 42,5% des seniors français travaillent contre 74,6% en Suède..." "L'emploi en France en 2020", le cabinet McKinsey dresse les faiblesses hexagonales pour relancer la création d'emplois, maintenir les seniors au travail et adapter les compétences des étudiants aux attentes des entreprises.

Réduire les écarts de compétences

Pour McKinsey, alors que les jeunes fortement diplômés sont formés à des métiers sans avenir et déconnectés du monde du travail, les moins diplômés en sont, eux, exclus. Une inadéquation des niveaux d'études qu'il est alors nécessaire de "réduire pour traiter en profondeur la question de l'emploi", constate le cabinet. Et d'estimer urgent d'élever le niveau de compétences en les adaptant à l'économie du savoir tout en améliorant l'orientation des formations initiales et continues vers des secteurs porteurs. En clair, il serait urgent de préparer les futurs diplômés à l'économie du 21ème siècle afin de mieux les orienter vers des métiers porteurs.

Si le système éducatif français ne s'adapte pas au nouveau contexte économique, "il manquerait 2,2 millions de diplômés en 2020, tandis que 2,3 millions d'actifs sans diplôme ne trouveraient pas d'emploi", estime McKinsey. 5 priorités pour relancer la création d'emplois

Pessimiste, le cabinet juge également que la France devra dans les années à venir "plus que doubler sa capacité historique à créer des emplois pour enrayer l'érosion de la prospérité en assurant une croissance de son PIB de 2,1%""la France devrait créer pas moins de 370 000 emplois par an" dans les 10 ans à venir.

Classique dans ses recommandations, le cabinet estime que pour créer plus d'activités et résorber l'écart des compétences, il faut dans un premier temps réformer les conditions de l'emploi des seniors.

Les entreprises doivent alors "concevoir de nouveaux modes de travail adaptés" et les pouvoirs publics développer la formation tout au long de la vie. Seconde priorité à envisager pour la France : "assurer la montée en compétences de la main d'oeuvre et mieux adapter ces compétences aux besoins d'une économie du savoir d'améliorer la compétitivité-coût du travail et lui donner davantage de souplesse", "adopter des stratégies de croissance ciblées sur les gisements de création d'emplois pour tous les actifs, diplômés ou non" et enfin "améliorer les mécanismes d'ajustement entre offre et demande de main d'oeuvre"

Bref, pour "gagner la bataille des compétences dans une économie du savoir", la France devra mobiliser les acteurs du monde enseignant, les pouvoirs publics et les administrations compétentes ainsi que les entreprises. Un projet ambitieux !

(TITRE:Le groupe STURNO propose des stages pour sa filiale SPHERE - RegionsJob:TITRE)

Présentez-nous la société

Le groupe STURNO (650 collaborateurs) est une entreprise familiale à taille humaine (4 générations depuis 1920) et bénéfice d'une solide image de marque auprès des marchés publics et privés. Implanté sur le Grand Ouest (Normandie, Bretagne Pays de Loire), le Groupe STURNO comprend aujourd'hui trois grandes activités : - Les Travaux publics (Réseaux humides : eau potable, assainissement), (Réseaux secs : électricité HTA et BTA, gaz, Telecom et fibres optiques)- La Gestion des réseaux d'eau et d'assainissement, avec la société STGS- Les Déchets - Environnement, avec la société SPHERE (Collecte sélective, Collecte et traitement de Déchets Industriels, Exploitation de centre de tri, Collecte de déchets ménagers, Exploitation de centres d'enfouissements, Exploitation de déchetteries)

Aujourd'hui le groupe recrute, est-ce lié à une hausse de l'activité ?

Concernant la société SPHERE, nous recrutons plus généralement des chauffeurs, des équipiers de collecte ou des agents de tri. Mais actuellement, comme nous cherchons à accompagner la croissance de la société SPHERE et à développer son image de marque, nous souhaitons intégrer des stagiaires sur des projets transversaux tels : le développement informatique, l'optimisation de la production, le perfectionnement de notre logistique et l'émergence de notre politique QSE.

Quelles missions leur proposez-vous ?

Dans le cadre de sa dynamique de changement, la société SPHERE crée notamment de nouveaux outils informatiques pour gérer ses métiers : gestion des clients, gestion des tournées, reporting,

échanges de document via une plate-forme collaborative, informatique embarqué... Ces évolutions considérables nécessitent des spécialistes de ce domaine pour atteindre les objectifs de la Direction. A ce titre, nous cherchons un Technicien logistique environnement en contrat de professionnalisation sur le secteur Centre Manche (Avranches) dont la mission principale sera d'optimiser nos tournées de collecte ; ainsi qu'un stagiaire informatique qui aura pour objectif d'assurer l'intégration de terminaux embarqués dans les véhicules de collecte, de développer de nouveaux modules, et qui aura en charge de développer le site Web de la société SPHERE.

Nous proposons également un stage d'ingénieur production au sein d'un des plus performants centres de tri de la région Basse-Normandie. Ce dernier devra optimiser la collecte et la production du centre de tri de déchets papiers/cartons/plastiques dans le cadre des nouvelles consignes de tri Eco emballage (nouvelles générations de plastiques).

Enfin, nous souhaitons intégrer à la rentrée prochaine un stagiaire Qualité Sécurité Environnement qui aura pour mission d'accompagner le développement du système management intégré sur le centre de tri de Donville-les-bains (évaluation des risques, audit des actions mises en place, formalisation des modes opératoires, construction des documents pédagogiques, collecte et l'analyse les données, amélioration continue des procédures en place).

Recherchez-vous des profils particuliers pour ces stages ?

Outre des candidats diplômés de Bac + 2 à Bac +5 pour le poste de Technicien Logistique et de Bac + 4 à Bac + 5 pour les 3 autres missions de stages, nous souhaitons travailler avec des personnes rigoureuses capables de s'adapter aisément à l'ambiance de travail et être force de propositions. Le logisticien devra également être un bon communicant. En effet, il travaillera en lien étroit avec les chauffeurs de la société SPHERE.

Quelles sont les opportunités de carrière pour les stagiaires ?

Si nous parvenons à valider ces projets transversaux, il est envisageable que certains postes soient plus tard pérennisés grâce à la dynamique du Groupe STURNO. Cependant cela n'est pas encore à l'ordre du jour. Mais dans tous les cas, ces stages de 6 mois minimum et d'un an pour le poste de Technicien Logistique sont l'occasion de s'investir dans des projets stimulants. C'est aussi la chance de travailler en toute autonomie sur des projets intéressants.

… …

2.4. Forum_Emploi

(TITRE:Quels sont vos attentes pour ce poste? : Bistro : La Pause Café:TITRE)

Bonjour, Malgré toutes mes recherches sur le Web, je n'ai pas trouver d'éléments de réponse à cette question qui est pourtant systématiquement posée en entretien. Merci de m'éclairer à ce sujet car je ne vois pas vraiment comment y répondre?

EN FAIT le recruteur vous pose la question pour connaitre votre façon à vous progeter sur le postece que vous voulez faire? ce n'est pas une question piège il faut juste se mettre dans les basquettes du gars qui bosse. Bonjour, meci n'est en effet pas une question piège. Le recruteur cherche à savoir ce que vous recherchez dans le travail (au dela du salaire). Travaillez vous plutôt en autonomie? Est un poste très terrain et c'est ce que voua appréciez? Est ce un travail administratif qui demande de la rigueur? Donc vous pouvez répondre quelque chose du genre:"Je recherche en fait un poste dans lequel je puisse travailler en autonomie.."ou alors"Moi ce qui m'a plus dans ce poste, c'est la forte attente enterme de service client. C'est quelque chose qui est important pour moi..."ou bien "Je recherche un poste qui puisse me permettre d'élargir mes compétences dans le domaine de.... C'est pour franchir cette étape que j'ai postulé."

Gestion Encadrement.

(TITRE:femme btp : Bistro : La Pause Café:TITRE)

Bonjour, je suis électricienne bâtiment et je m'éclate quand je refait une maison de A à Z en élec. Je touche aussi un peu à tout dans le bâti, j'aime le travail manuel. ce jour je peine à trouver du boulot, les employeurs ne me laisse aucune chance de faire mes preuves et j'en ai ras le bol. a-t-il des femmes dans mon cas sur mon secteur ? J'envisage de monter une boite de btp de femmes (carreleuse, plombière, maçonne, menuisière, etc) et de mettre une grosse carotte à tous ces machos !bonjour moi je suis dessinatrice en maison individuelle depuis 23ans et j'ai la chance d'avoir changé plusieurs fois de sociétés (chance si on peut dire) moi j'en ai assez des patrons qui confondent portefeuilles sociétés avec le leurs une société que de femme pas mal comme idée mais dans ce milieu d'homme je que cela serai très diffcile bonne chance à vous

(TITRE:Aide pour ma recherche sur les relations au travail : Bistro : La Pause Café:TITRE)

Bonjour, Actuellement étudiante de master 1 en Psychologie du travail, je vous propose de participer à une recherche universitaire sur les relations au et dans le travail. Votre participation m'est

indispensable pour la bonne démarche de ma recherche. Aussi, vous trouverez dans les pages suivantes divers questionnaires auxquels vous devrez répondre le plus spontanément possible. Il n'y a pas de bonnes ou de mauvaises réponses, seules vos opinions personnelles comptent. Conformément au code de déontologie des Psychologies vos réponses resteront anonymes et strictement confidentielles. Votre aide me sera précieuse pour mener à bien cette recherche et je vous remercie du temps et de l'attention que vous y accorderezsuivez le lien:http://letrocdopinions.vixia.fr/reponse ... mquest=242

(TITRE:Enquête pour ceux et celles qui cherchent un emploi : Bistro : La Pause Café:TITRE)

Bonjour, dans le cadre de la création d'un cabinet de recrutement spécialisé, je mène actuellement une étude sur les outils Web utilisés dans le cadre de votre recherche d'emploi. Le nombre de questions à été volontairement limité afin que cette étude soit rapide à compléter. Bien entendu, vous pouvez diffuser le lien de cette enquête autour de vous en vue d'obtenir un maximum de réponses qui rendra cette étude réaliste. oici le lien : http://www.mon-enquete-enligne.fr/index.php?sid=17589 lang=fr

Les résultats de cette étude seront disponibles pour ceux qui le souhaitent. Merci d'avance de votre participation ébastien

(TITRE:chomage et maladie : Bistro : La Pause Café:TITRE)

Bonjour,

J ai 54 ans a nouveau au chomage pour la 3è fois. Je n'en peu plus . ET ma vie privée s'est fortement dégradée. Je deviens aigris et je ne communique plus ou peu. Je m'éloigne de mes amis et eux aussi s'éloignent de moi. Je me suis enfermé dans une bulle qui a gaché 40 années de ma vie. J'essaie de m'en sortir mais j'ai l'impression de faire des pas en arrière au lieu d'aller vers l'avant. .as de projets 'je n'ai jamais su en faire)

E m'étais même interdits de rire.

Opéré à 14 ans d'une opération à coeur ouvert j'ai été mis dans une bulle depuis ma naissance. Aucan suivi psychologique ni avant ni après l'opération. Solitaire j'ai été, solitaire je suis encore. Et je n'arrive pas à aider les personnes de mon entourages qui en auraient besoin en ce moment.

Et il faut encore que je trouve les force pour rebondir car j'ai encore besoin plusieurs années comme beaucoup et en plus j'ai 2 enfants qui n'ont pas encore faits leurs études. "est la première fois que je m'exprime par le biais d'un forum. merci d'avoir lu ce message

Bonjour,e l'ai lu votre message.t pour sortir de votre bulle : je vous offre un sourire.

(TITRE:Soutenez les etudiants : répondez au questionnaire (20 minut : Bistro : La Pause Café:TITRE)

Bonjour à tous,

Dans le cadre d'une étude sur la construction d'un questionnaire mesurant les styles de résolution dans un contexte de travail, nous souhaitons recueillir le plus grand nombre de réponses de personnes travaillant dans une entreprise (publique ou privée). Pour pouvoir contribuer à cette étude, il vous suffit de répondre à un questionnaire accessible grâce au lien joint ci-dessous. Toutes vos réponses seront évidemment récoltés de manière confidentielle et anonyme. De plus, nous ne manquerons pas de vous transmettre un compte-rendu de vos résultats si vous le souhaitez. Nous vous remercions par avance de l'aide que vous apporterez à cette recherche grâce à votre participation et vous invitons à diffuser ce message à vos contacts qui seraient susceptibles de participer à l'étude. Pour accéder au questionnaire de cette étude et à sa présentation veuillez suivre ce lien:https://docs.google.com/spreadsheet/vie ... c6MQ#gid=0i vous avez des questions à propos de cette recherche n'hésitez pas à nous contacter, nous restons à votre disposition pour y répondre.ien cordialement

(TITRE:Demarche pour un emploi pour un etranger? : Bistro : La Pause Café:TITRE)

Bonjour, je cherche actuellement un emploi pour mon ami bulgare. Sachant qu'elle maîtrise le slovaque, le bulgare, l'anglais, quelques bases d'allemand, connaît le minimum en français et à deja travailler 2 ans dans une boite de telephonie mobile. Pourriez vous m'aider à trouver la meilleur solution pour lui trouver un job, a qui doit je m'adresser, quelle institution serait la plus apte à m'aider...etc? Cordialement

(TITRE:Questionnaire "stress et conditions de travail" (thèse) : Bistro : La Pause Café:TITRE)

Bonjour à tous, j'effectue une thèse, je l'ai d'ailleurs bientôt finie, et j'ai créé un questionnaire à l'aide de Google Drive afin de recueillir des informations sur les conditions de travail des salariés. J'est un questionnaire très simple qui ne dure que 4 à 5 minutes. Le but de ce questionnaire est de déterminer à la fois si les salariés ont conscience de leurs conditions de travail et s'ils savent se prémunir de potentiels risques liés à leur activité professionnelle. Tous les salariés sont concernés. Le questionnaire est anonyme. Si vous êtes intéressé(e) pour partager votre expérience, je vous en remercie par avance. Vous trouverez mon questionnaire à cette adresse : http://tiny.cc/2o94dx

J'espère avoir le droit de poster ma demande d'aide ici

N'hésitez pas également à donner vos avis dans les commentaires, c'est toujours utile de partager ses expériences ! bientôt bey

(TITRE:Senior83 se présente à vous : Bistro : La Pause Café:TITRE)

Toc - Toc - Toc

Faute d'avoir trouvé la rubrique des présentations, je pousse la porte du bistro avec modération. Bonjour à toutes et à tous, non, non ce n'est pas le Père Noël avec un peu d'avance. J'aimerais bien l'être pour vous tous et pour moi-même mais la réalité en est autrement. Donc, à la veille de cette soirée en famille, je m'inscris sur Forum Emploi. Je me présenterais de façon plus complète dans quelques jours afin que vous puissiez vous faire une meilleure idée de ma personne et où j'en suis. D'ici là, bon noël à vous tous.enior83

(TITRE:Remboursement pôle emploi : Bistro : La Pause Café:TITRE)

Bonjour, je dois payer la somme de 441€ à pôle emploi, ne pouvant payer en une fois, j'ai demander à pôle emploi si je pouvais les reglés en mensualités de 50 € mais ils refusent et veulent des accomptes de 107€.our mon budget cela n'est pas simple. Penser vous que si je leur verse quand même 50€ jusqu'a rembousement complet, ils peuvent tout de même me chercher des ennuies ? Lorsque je leurs ai demander, ils m'ont écrit en me disant que si je ne versai pas 107€ ils s'adresserai aux organismes compétent pour me faire payer la somme restant dû en une fois.si vous avez déja vécu cette situation,merci de me faire partager votre experience, vos conseils sont les bienvenues merci à tous

(TITRE:Que faire si l'on reçoit plusieurs réponses positives? : Bistro : La Pause Café:TITRE)

Bonjour, je m'apprête à envoyer plusieurs candidatures spontanées (par courrier et par mail) et je me demandais ce qu'il fallait que je fasse au cas où j'aurais plusieurs réponses positives. merci d'avance pour vos réponses.

Planifie tous tes rendez vous et fonce aux entretiens !! ensuite tu pourra décortiquer tes entretiens et choisir ...ar les temps qui court le travail est rare !Bon courage k, merci ^^

Bon, j'espère avoir quand même au moins une réponse positive, ce serait bien ^^merci ^^

… …