Extraction d'information : Génération automatique d'une base de données d'offres d'emploi

Inaugural-Dissertation

zur Erlangung des Doktorgrades
der Philosophie an der Ludwig-Maximilians-Universität
München

vorgelegt von

Sandra Bsiri

am 27. April 2007 in München

Tag der mündlichen Prüfung : 06.07.2007
Betreuer : Herr Prof. Dr. Franz Guenthner
Zweitgutachter : Herr Prof. Dr. Klaus Schulz
Drittgutachter : Herr Prof. Dr. Christian Böhm

Extraction d'information : génération automatique d'une base de données d'offres d'emploi

Résumé

Les technologies internet ont fourni au marché du travail les moyens d'une large diffusion et de traitement de l'information en temps réel, cependant au lieu d'aboutir à une centralisation des offres d'emploi, on se trouve face à un marché morcelé et à une redondance des données. Dans la conjoncture instable actuelle du marché de l'emploi, l'objectif des travaux présentés ici est de mettre en place une plate-forme unique centralisée réunissant en temps réel toutes les offres d'emploi disponibles sur le réseau Internet et dispersées à ce jour sur une multitude de vecteurs de diffusion parallèles. Nous avons implanté un premier système de reconnaissance et de classification automatique de pages d'accueils d'entreprises qui nous permet de tenir à jour un annuaire d'entreprises, à partir desquels nous écupérons les annonces des postes vacants. Nous avons construit dans une seconde phase une quantité importante de grammaires locales, nous permettant d'extraire l'information nécessaire à la transformation de l'espace de représentation des offres d'emploi initialement rédigées en plein texte en des document sémantiquement structurées et ainsi remplir automatiquement la base de données des annonces de postes.

Mots clefs :

Extraction d'information, grammaires locales, dictionnaire électronique, classification non supervisée,

Informationsextraktion zur automatisierten Erzeugung einer Datenbank frankophoner Stellenanzeigen

Zusammenfassung :

Das Internet hat sich dank seiner hohen Verbreitungseffizienz zum zentralen Medium des Stellenmarktes entwickelt. Das Spektrum konkurrierender Stellenbörsen und anderer Diffusionswege im Netz hat allerdings zu einer hohen Redundanz der Daten und eingeschränkter Transparenz geführt. In dieser Arbeit wird am Fall des frankophonen Stellenmarktes illustriert wie mit linguistisch basierten Methoden alle im Netz verfügbaren Stellenangebote erstmals auf einer zentralisierten Plattform gesammelt und gefiltert werden können. Ein System zur automatischen Erkennung und Klassifikation von Firmen-Homepages wurde implementiert, um eine Firmen-Datenbank zu erstellen und auf aktuellem Stand zu halten. Basierend auf diesem Verzeichnis werden die HTML-Strukturen auf Anchortexte durchsucht, die zu den offenen Stellen führen. Lokale Grammatiken und elektronische Lexika wurden ausgearbeitet, um in einer zweiten Phase jene Informationen zu extrahieren, die für die Umwandlung der reinen Textform der Stellenangebote in ein semantisch strukturiertes Dokument notwendig sind. Über diese linguistische Analyse der einzelnen Einträge wird die Datenbank automatisch gefüllt und höchste Selektivität bei Suchanfragen gewährleistet.

Schlüsselwörter :

Automatische Informationsextraktion, lokale Grammatik, elektronische Lexika, nicht überwachter Klassifizierung

à Uli,
à mes parents

TABLE DES MATIÈRES

CHAPITRE 1

Description du Projet

Introduction

La conjoncture actuelle qui connaît une vraie instabilité du marché du travail et un taux de chômage assez élevé nous a encouragé à concilier progrès technologique et intérêt social. Notre objectif est de mettre en place une plateforme unique centralisée réunissant toutes les offres d'emploi disponibles sur le réseau Internet et dispersées à ce jour sur une multitude de vecteurs de diffusion parallèles.

D'après une étude élaborée par le groupe *FocusRH* il y aurait eu plus de 900 000 offres d'emploi et de stages publiées en 2006 dans les 500 sites emploi étudiés [1]. Ces chiffres ne reflètent pas la réalité, car une quantité importante d'offres est publiée simultanément sur plusieurs intermédiaires du marché de l'emploi sous des formats hétérogènes qui engendrent des pertes d'information considérables et d'autres sont publiées sur des vecteurs informationnels différents comme les sites carrières d'entreprises ou sur des forums de discussion spécialisés ou des annuaires de petites annonces surtout pour les métiers d'artisans dont les salaires ne dépassent pas les 20 K€ par an. Certaines entreprises refusent également de publier leurs annonces sur les sites emploi de grande notoriété pour des raisons de coûts d'une part et pour ne pas recevoir d'autre part de nombreuses candidatures ne correspondant en rien à leurs besoins. Ainsi il existe une grande quantité d'offres d'emploi éparpillées sur le Net et qui restent souvent non atteignables par les candidats intéressants.

De plus une enquête réalisée par l'entreprise de travail temporaire *Kelly Services*[2] au premier trimestre 2006 sur un échantillon de 19.000 personnes dans 12 pays d'Europe a montré qu'Internet était le premier outil utilisé dans la recherche d'emploi et que le taux d'utilisation de ce média par les

[1] Le guide des 500 meilleurs sites emploi, édition 2006

[2] http ://management.journaldunet.com/repere/outils-recherche-emploi.shtml

français à la recherche d'emploi s'élèverait à 70 % avec un taux de 73 % de femmes contre 65 % d'hommes.

Une multitude d'intermédiaires du marché du travail ont vu le jour avec l'évolution des techniques d'Internet. Nous parlerons dans le chapitre suivant des job boards généralistes et spécialisés, des agrégateurs d'offres d'emploi, des cabinet de communication RH ainsi que des intermédiaires classiques que représente la Presse écrite par exemple. Ceux-ci ont certes réussi à rendre le marché du travail plus transparent et plus facilement accessible aussi bien aux recruteurs qu'aux demandeurs d'emploi, néanmoins au lieu de réussir une convergence des offres électroniques vers une grande bourse de l'emploi unique, une plateforme de partage de l'information, ils ont conduit à une redondance des données et à l'augmentation du taux des données bruitées non pertinentes. Ceci est dû surtout à l'interaction directe des protagonistes de la scène de l'emploi entre eux et à l'abolition des filtres que représentaient les intermédiaires traditionnels.

En outre, les sites emploi généralistes soucieux de chiffres, en ont oublié la qualité. Leur soucis d'atteindre le nombre le plus élevé d'offres d'emploi publiées, le nombre le plus élevé de CVs déposés et le nombre le plus élevé de visiteurs les a fait négliger le rôle social qu'ils ont à jouer sur le marché du travail. Ils se sont désintéressés de la qualité des services offerts en se contentant d'utiliser des méthodes d'indexation et de recherche traditionnelles purement statistiques. Face à ces sites généralistes, les sites emploi spécialisés font surface pour combler ce manque informationnel et tentent de répondre aux besoins spécifiques de certaines catégories professionnelles, géographiques ou sectorielles en se concentrant non plus sur la quantité mais sur l'homogénéité des offres proposées. On se retrouve donc face à un marché morcelé en micro-marchés concurrents car pour ces entreprises la publication des offres représente plus un marché qu'un service d'intérêt général.

Au vu de ces investigations et face à un tel taux d'utilisation de ce média comme intermédiaire de l'emploi, nous nous sommes fixés comme objectif d'élaborer un outil plus adapté aux besoins des demandeurs d'emploi en proposant un système qui s'applique à rechercher et rassembler toutes les offres existantes sur le marché tout en assurant une analyse linguistique de chacune des offres pour mieux servir les besoins des utilisateurs. Nous avons donc réfléchi à des solutions automatiques nous permettant d'atteindre ces pages Web et de les transformer en des entrées sémantiquement structurées ou semi-structurées compréhensibles ultérieurement par un système de recherche d'information.

Ce travail de thèse s'inscrit dans le cadre d'un projet de valorisation de la recherche. Il s'agit d'un encouragement à la création d'entreprise à partir de technologies innovantes. L'entreprise cible est un *moteur de recherche d'emploi* qui profite dans toutes ses phases de traitement du résultat de nos recherches dans le domaine de l'extraction automatique d'informations et ce, en nous basant particulièrement sur des méthodes linguistiques d'analyse et

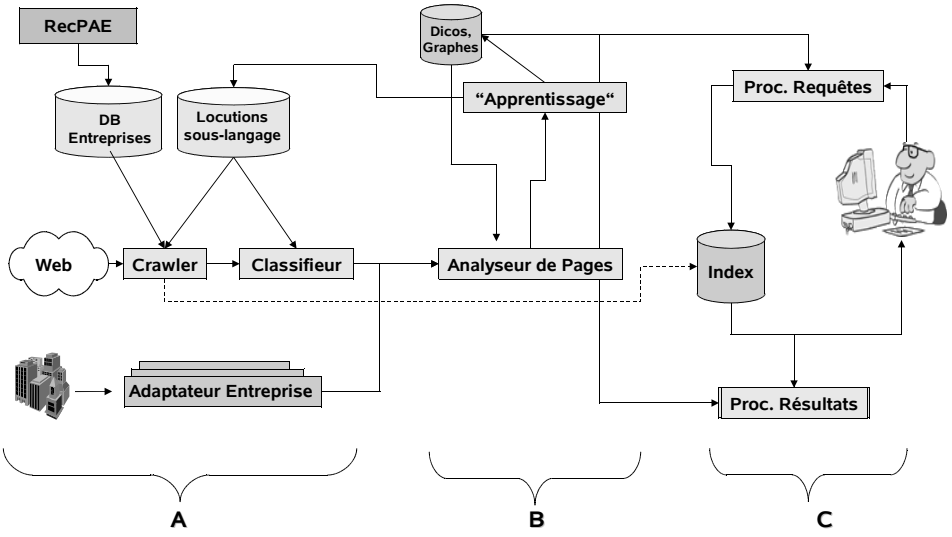


FIG. 1.1 – Une vue d'ensemble de notre système

de traitement de la langue naturelle.

Nous avons mis au point un système multi-niveaux, qui use de théories linguistiques et surtout de la notion de grammaires locales dans toutes ses phases de traitement et d'analyse. Les différents modules élaborés pour servir notre objectif sont détaillés tout au long de ce manuscrit. Nous tentons dans ce chapitre de présenter une vue d'ensemble du système et de montrer comment les modules interagissent entre eux pour répondre à nos besoins.

1.1 Notre système

Nous présentons au niveau de la figure (fig- 1.1) une illustration du système global que nous proposons en réponse aux limites observées dans les systèmes similaires. Elle rappelle à première vue et de par ses trois phases principales le comportement d'un moteur de recherche ordinaire, où l'on retrouve la phase :

- ▷ Recensement et découverte des documents
- ▷ Analyse des documents et extraction d'information
- ▷ Recherche et traitement des requêtes

Ces phases globales sont notées par les lettres « A, B, C » dans la figure (fig-1.1). Chacune d'entre elles représente un ensemble de sous-systèmes que nous avons conçu pour répondre au mieux à nos besoins de collecte des offres d'emploi disponibles sur le Web et de transformation de celles-ci dans

des représentations sémantiquement chargées afin de proposer de meilleurs services aux utilisateurs par rapport aux intermédiaires de l'emploi actuels.

1.1.1 Collecte des documents (offres d'emploi)

La première phase est celle de collecte des documents à stocker, qui sont dans notre cas des pages d'offres d'emploi. Le but est alors de mettre en oeuvre un système capable d'aller chercher et reconnaître les pages Web des offres dispersées sur Internet sur différents vecteurs de distribution.
Les offres d'emploi sont publiées sur plusieurs plateformes entre les sites d'emploi professionnels, les sites des entreprises appelés « les sites carrières » les forums de discussion et les annuaires de petites annonces spécialisés.
La plus grande partie des offres est néanmoins publiée sur les sites carrières des entreprises qui se trouvent parfois et surtout pour les grandes entreprises dupliquée sur les moteurs de recherche d'emploi mais beaucoup d'entreprises se contentent aussi de poster leurs offres uniquement sur leurs sites.

En vu de ces observations nous avons développé deux stratégies de découverte des offres publiées sur Internet. La première se charge de parcourir les sites carrières des entreprises et la seconde de développer un robot d'indexation focalisé qui reconnaît à travers la terminologie spécifique au langage spécialisé des offres d'emploi, les URLs en rapport avec les annonces de vacances de postes par rapport à toute autre URL.

Reconnaissance de pages d'accueil d'entreprises

Avant d'être en mesure de chercher les offres d'emploi publiées sur les sites carrières, il est nécessaire de disposer d'une base de données de sites d'entreprises. Notre objectif étant de centraliser l'information sur le marché de l'emploi, nous nous devons de tenir une base de données d'entreprises à l'image de la scène économique réelle et non pas une liste répertoriant uniquement les grandes entreprises connues en omettant toutes les petites et moyennes compagnies méconnues du grand public. Pour faire face à cette exigence nous avons développé un système de reconnaissance automatique de pages d'accueil d'entreprises que nous avons baptisé *RecPAE* et que nous présentons en détail au niveau du chapitre 6.

Il s'agit d'un système de classification automatique, qui décide pour chaque URL reçue en entrée si elle appartient à la classe *« Organisation »* ou à la classe *« Autres »*. Nous ne nous préoccupons pas à ce niveau d'une classification sémantique en fonction des secteurs d'activité car une entreprise de *développement logiciels* peut aussi bien proposer des postes de cuisinier pour sa cantine. Ceci pour insister sur le fait qu'une classification des entreprises ne garantie pas la classification automatique ultérieure de ses postes vacants.
Le système *RecPAE* est basé sur une analyse structurelle et sémantique pour la reconnaissance des pages d'accueil des organisations qui partagent un cer-

tain nombre de descripteurs que nous avons été en mesure de traduire dans une représentation permettant au système de décider automatiquement de la classe de l'URL.

Une fois les pages d'accueil des entreprises identifiées il faut alors repérer les postes vacants qui y sont proposés. Ceci est rendu possible par la structure HTML des pages d'accueil des entreprises qui disposent d'un hyperlien interne au site avec un texte d'ancre appartenant à la classe « Jobs ». La classe *« Jobs »* est une classe dans laquelle nous avons réuni des séquences extraites lors de la phase d'apprentissage. Ces séquences, que l'on aperçoit au niveau de la liste ci-dessous, représentent des textes d'ancres trouvés sur les pages d'accueil des organisations et dont la sémantique en rapport avec le recrutement permet de rediriger les visiteurs du site vers la page des postes vacants de cette entreprise.

consulter nos offres
consultez nos offres d'emploi
carrière chez Xxxx
emploi
jobs
postes vacants
rejoindre notre équipe
recrutement en ligne
travaillez chez Xxxx
...

Nous avons récolté plus de 80 séquences que l'on augmente au fur et à mesure de nos analyses par apprentissage semi-automatique. Semi-automatique car nous avons écrit certains patrons de pré-classification du genre : si un nouveau texte d'ancre est rencontré et qu'il contient les mots *« recrutement, emploi, job... »*, alors il est mis dans la classe « Jobs »/. Nous donnons plus de détails sur les différentes classes sémantiques que nous avons créé au niveau du chapitre 6.

En fonction de la taille des entreprises, les postes vacants sont soit listés les uns après les autres sur une même page Web, soit seuls les titres des postes liés par des hyperliens sont présents, soit un mini système de recherche d'offres internes est disponible. A ce niveau, différents modules du système sont appliqués en fonction des cas. Notre système rencontre des difficultés surtout avec le premier cas, où plusieurs offres se trouvent sur une même page Web. Le problème majeur rencontré en ce point est la difficulté de segmentation correcte des différentes annonces et la difficulté d'association correcte des informations extraites aux différents postes décrits.

Robot d'indexation focalisé

Pour la seconde source de récupération des offres d'emploi nous avons développé un robot d'indexation focalisé, qui à travers la terminologie des offres d'emploi, retient uniquement les URLs qui nous intéressent. Les URLs retenues sont considérées comme des annonces d'emploi potentielles qui devront toutefois passer par une seconde phase d'analyse et de classification avant d'intégrer la base de données des offres. Ce robot use d'une terminologie qui se compose essentiellement de syntagmes plus ou moins composés, récoltés manuellement à partir d'un large corpus d'apprentissage.

Ces syntagmes sont de deux types. Les premiers de type nominal pour la plupart permettent d'organiser sémantiquement l'offre selon les multiples catégories qui y sont décrites. Elles représentent les titres de sections dans les offres d'emploi. Nous avons pu classifier ces derniers dans 13 classes sémantiques dont nous montrons un échantillon au niveau du tableau 1.1.

TAB. 1.1: Échantillon des classes sémantiques des différentes sections d'une offre d'emploi

Classes	
Mission	Description de la mission : Le candidat aura pour mission : Mission du stagiaire : Votre mission : Votre mission consistera à : Vous accomplirez les missions suivantes : ...
Les tâches	Principales tâches : Dans ce cadre, vos tâches seront les suivantes : Description des tâches principales : Fonctions et tâches du poste : Vous aurez la charge de : Fonctions et tâches : ...
Compétences	Compétences en : Compétences techniques exigées Connaissances et compétences requises Il devra posséder les compétences et qualités suivantes Les compétences requises pour ce poste sont Profil / Compétences : Profil requis / Compétence : ...

La suite sur la page suivante

Tab. 1.1 – suite de la page précédente

Classes	
Qualités	Qualités recherchées Possédant une ou plusieurs des qualités : Qualités spécifiques demandées Les qualités requises sont les suivantes : Les qualités techniques requises pour cette mission sont Profil / Qualités : Profil requis / Qualités spécifiques : Vos qualités ...
Connaissances	Connaissances techniques Connaissances : Vous avez des connaissances particulières dans : Connaissances personnelles souhaitables Connaissances et qualités requises ...
Profil	Profil/compétences Profil du candidat Profil idéal Description du profil du candidat Votre profil : Le profil recherché
Expérience et Formation	Expérience : Niveau d'expérience et formation : Expérience professionnelle : Ayant l'expérience suivantes Expérience technique en : Vous devez avoir de l'expérience dans : L'expérience requise pour ce poste ...
Durée	Durée de la mission : Durée : Durée du CDD : Durée du contrat : ...
Date de début	Date de début de mission : Début de mission Début de la mission ...

La suite sur la page suivante

TAB. 1.1 – suite de la page précédente

Classes	
Lieu	Lieu de mission Poste basé à : Lieu du poste lieu de travail : ...
Description du poste	Description de l'offre : Description du poste Description de l'offre : Détails de l'offre Titre du poste : Secteur du poste : ...
Salaire	salaire : Nous vous offrons un salaire de : Rémunération : ...
Coordonnées du contact	société : Nom du contact : Email du contact : Fax du contact : Adresse : Tel : Fax : ...

A côté de ce premier type de syntagmes, nous avons retenu également des locutions s'articulant autour de verbes supports ou de mots spécifiques venant enrichir presque toute offre d'emploi. Il s'agit d'un ensemble de formulations propres aux offres d'emploi qui permettent à notre Crawler spécialisé de classifier une page Web dans la classe « Offres » ou alors dans la classe « Autres »des URLs à ignorer. Ces formulations (voir un échantillon dans le tableau ci-dessous) sont surtout intéressantes dans le cas des offres non structurées où l'on ne trouve pas les syntagmes du premier type décrit antérieurement et qui représentent des preuves de classification sûres.

le poste est à pourvoir très rapidement
merci d'adresser votre candidature par courrier
vous avez une 1ère expérience d'au moins X ans
vous justifiez d'une expérience dans un poste similaire
vous serez responsable d'une équipe
vous êtes de formation bac au minimum
vous êtes débutant ou justifiez d'une première expérience
vous êtes déjà impérativement confirmé
vous aurez la responsabilité de
vous avez de solides connaissances en
vous êtes rigoureux, créatif
merci de préciser vos prétentions salariales
nous recherchons pour notre siège basé
...

Nous comptons un total de 500 formules entre les deux types décrit ci-dessus. Cette première phase de notre système délivre donc un ensemble d'URLs d'offres d'emploi potentielles. Elle permet de par le filtre terminologique de réduire la quantité d'URLs à analyser dans la seconde phase d'analyse et de transformation de la représentation des annonces initialement en plein texte dans une représentation compréhensible par le système.

1.1.2 Transformation de l'espace de représentation des documents

La seconde phase de notre système et qui représente une grande partie de nos travaux est l'analyse et l'extraction d'information à partir des offres d'emploi. Dans cette étape nous tentons d'extraire les informations nécessaires au remplissage du formulaire correspondant à la structure de notre base de données. Le but étant de transformer les documents écrits en langage naturel en des documents semi-structurés, chargés d'une sémantique permettant d'améliorer les résultats de recherche d'emploi et ce en rapprochant les requêtes plein texte à des requêtes sur des champs d'une ou plusieurs tables d'une base de données.

Ainsi à partir d'un document provenant du collecteur d'URLs, nous essayons d'extraire automatiquement les informations suivantes :

Structure d'une offre d'emploi à extraire	
Date de publication	22. Jan 2007
Date limite de postulation	fin février
Date d'embauche	mi- Mars
Nom du poste	ingénieur d'étude en électromécanique
Type de contrat	intérim à temps partiel : 1 à 2 jours/semaine
Durée du poste	8 mois renouvelable
Lieu du poste	sud-est de Paris
Rémunération	selon profil
Référence du poste	MOR34544/ing-21
Expérience minimum	expérience de 2 à 3 ans dans un poste similaire
Formation du candidat	de formation Bac+5 de type école d'ingénieur
Nom de l'entreprise	CGF Sarl
Coordonnées de l'entreprise	**Adresse :** 34 bis rue Berthauds, 93110 Rosny **Téléphone :** 0 (+33) 1 12 34 45 67 **Fax :** 0 (+33) 1 12 34 45 68 **Email :** contact@cgf.fr **Home Page :** http :///www.cgf
Contact à qui adresser la candidature	Directeur des RH, Mr. Brice
Domaine d'activité de l'entreprise	Construction électromécanique

De plus nous sommes en train de construire des grammaires autour de verbes supports pour extraire les informations sur les tâches de la fonction à occuper ou sur les compétences à apporter. Ces deux sections d'une offre emploient des verbes d'action très particuliers (assurer, gérer, veiller,...) que nous essayons de localiser à partir de notre corpus d'offres. Elles permettront dans une étape ultérieure de raffiner les recherches des candidats sur les compétences exigées par exemple ; ce qui leur permettra de rechercher un poste non pas par le biais du titre qu'il pourrait avoir mais en fonction des compétences attendues chez les candidats idéaux ou de rechercher un poste en fonction des tâches à accomplir décrites. Ceci est très intéressant car les demandeurs d'emploi savent souvent ce qu'ils peuvent faire mais ne connaissent pas l'intitulé exacte de ce que peut être l'emploi correspondant à ces compétences.

Ainsi pour chaque URL reçue en entrée, le système recherche toutes les informations citées ci-dessus. Si au moins 50 % d'entres elles sont reconnues, le document est accepté comme étant une offre d'emploi et sera introduit dans la base de données. Le seuil de 50 % est ajustable en fonction des

informations trouvées et des patrons d'extraction utilisés. Nous ajoutons certaines règles de priorités et d'exceptions qui permettent de classifier un document dans la classe « Offres » même si moins de 50 % des informations recherchées sont reconnues. Certains patrons d'extraction représentent des preuves solides de classification. Si on arrive à trouver par exemple le titre du poste et que l'on reconnaît une locution du type *« si le poste vous intéresse veuillez nous envoyer votre candidature à l'adresse suivante »*, alors nous supposons que le reste de l'information n'a pas été explicitement exprimée ou bien que nos grammaires doivent être enrichies mais qu'il s'agit bien d'une offre d'emploi.

Toutes ces informations ne sont pas forcément présentes dans une annonce d'emploi. Certaines entreprises décrivent minutieusement le poste à pourvoir pendant que d'autres se limitent au juste minimum, surtout dans le cas des offres postées sur les sites carrières. Celles-ci sont souvent très concises car le candidat est en mesure de comprendre par raisonnement analogique que l'annonce qu'il trouve sur le site d'une entreprise doit concerner celle-ci et aucune autre, c'est pourquoi ce genre d'annonce se limite souvent au titre du poste et à une brève description en omettant par exemple les informations sur le recruteur.

1.2 Les offres d'emploi

Nous proposons dans cette section des exemples d'offres d'emploi publiées sur Internet pour montrer le type d'information avec lesquels notre système doit interagir. Une offre d'emploi peut être de trois types. Dans le premier type, les annonces rappellent des formulaires, on y reconnaît une structure et une séparation des informations en sections et sous sections. Nous montrons un exemple de ce type d'offre dans la figure 1.2.

Visuellement, les différentes sections sont distinguables de par la représentation en gras et en souligné de leurs titres : *« Missions proposées »*, *« Profil, Rémunération »*, *« Type de Contrat »*, *« Contact »*.

Le second type d'offres correspond à des textes compactes comme dans l'exemple de la figure 1.3, où la structure Html ne suffit pas à la reconnaissance des différentes catégories informationnelles décrites.

Et le troisième type correspond aux offres très concises comme celle de la figure 1.4, où le peu d'information précisé est présent hors contexte.

Ces trois exemples donnent une idée des documents et du langage spécialisé auquel nous nous intéressons. Nous montrons dans le reste de ce rapport les différentes analyses que nous appliquons à ces derniers pour obtenir une base de données homogène et rigoureuse d'offres d'emploi disponibles sur Internet.

Société de conseil en formation et recrutement spécialisé secteur informatique recherche pour l'un de ses clients, un(e) collaborateur(rice) pour réaliser les missions suivantes :

HOTE / HOTESSE D'ACCUEIL A TEMPS PARTIEL

Missions proposées :
Accueil téléphonique et physique.
Gestion courrier.
Assistance administrative et logistique.

Poste basé à Vélizy (78) : RER C + Bus, ou Train : Chaville + Bus.

Profil :
· Pour mener à bien cette mission un niveau BAC minimum est requis avec une expérience en accueil téléphonique et physique indispensable
· Qualités : bonne présentation, discrétion, autonomie, bonne communication
· Maîtrise des outils bureautiques et outils de messagerie (Outlook express)
· Anglais courant (pratiqué dans le cadre d'une expérience professionnelle)

Rémunération : 9,5 euros brut/heure, temps partiel

Type de contrat : CDD de 6 mois

Merci d'envoyer vos CV format Word

Contact :

Pauline CATTEAU
Société ILSIS
Service Recrutement
3 allée de Londres
91953 Les Ulis cedex B
01 64 86 47 07

recrut@ilsis.fr

FIG. 1.2 – Offre d'emploi : type 1

Rédacteurs spécialisés

Entrecom recherche en permanence des rédacteurs spécialisés dans tous les secteurs d'activité pour répondre aux demandes de ses clients.
Le rédacteur spécialisé est chargé de l'élaboration du contenu rédactionnel des supports édités par Entrecom. Sa connaissance approfondie d'un domaine (de l'imagerie médicale au marché de l'énergie) lui permet d'être un interlocuteur crédible et pertinent auprès de nos clients.

II est rigoureux dans sa recherche et vérification d'informations et maîtrise parfaitement l'écriture journalistique. Une connaissance du monde de l entreprise est indispensable ainsi qu un bon niveau d anglais.

FIG. 1.3 – Offre d'emploi : type 2

Profession: LA FRIANDISE
Cherche pâtissier à Seppois 2ans d'Expérience tél entre 7h et 12h
03 89 25 60 29

FIG. 1.4 – Offre d'emploi : type 3

1.3 Plan de la thèse

Le présent manuscrit est organisé en 8 chapitres. Après la description globale du projet et la délimitation du contexte d'application, le second chapitre fait l'inventaire des intermédiaires du marché du travail et présente une comparaison des 10 sites emploi les plus populaires sur le territoire français. Les trois chapitres suivants, se consacrent à un état de l'art sur le fonctionnement des moteurs de recherche, l'extraction automatique d'information et l'élaboration de dictionnaires électroniques de mots simples et composés, trois domaines dans lesquels évoluent les travaux présentés. Le chapitre 6 décrit le système de reconnaissance et de classification automatique des pages d'accueil d'entreprises, baptisé *RecPAE* et le chapitre 7 présente le dictionnaire électronique des noms de profession élaboré, ainsi que les techniques utilisées pour son augmentation. Le chapitre 8 se concentre sur la description de la phase d'extraction d'information et de transformation de l'espace de représentation des offres d'emploi initialement en plein texte. Ce manuscrit est alors clôturer par une conclusion générale et les perspectives envisagées.

CHAPITRE 2

Les intermédiaires sur le marché de l'emploi

> « *Trouvez un travail que vous aimez et vous ajoutez cinq jours à chaque semaine.* »
> H.Jackson Brown

Introduction

Le marché du travail a vu émerger depuis le début des années soixante, de nombreux intermédiaires ayant pour métier de rapprocher les recruteurs des demandeurs d'emploi. Deux acteurs institutionnels voient le jour dans les années soixante, à savoir l'ANPE, l'Agence Nationale Pour l'Emploi et l'Apec, l'Association pour l'emploi des cadres. Dans les années qui suivent, l'émergence d'une multitude d'acteurs privés est à observer ; on voit se créer les cabinets de communication Ressources Humaines et les agences de travail temporaire. Cependant depuis les années quatre-vingt-dix, ces acteurs traditionnels doivent faire face à l'arrivée massive des sites d'emploi, des job-boards et des agrégateurs d'offres d'emploi rendu possible par le développement des techniques Internet et la croissance de son utilisation dans le monde.

Nous décrivons dans ce chapitre les différents acteurs du marché de l'emploi et la contribution d'Internet dans l'élargissement du marché du travail. Nous montrerons également les changements apportés par Internet sur les métiers d'intermédiaire de l'emploi. Nous commençons par donner un aperçu des intermédiaires traditionnels comme la Presse et les agences de communication RH, afin de mieux situer les sites emploi dans leurs contextes. Nous nous attarderons à définir les notions nouvelles de job-boards, agrégateurs d'offres d'emploi et de moteur de recherche d'emploi pour conclure par une description des sites emploi les plus populaires en France accompagnée d'une

comparaison technique sur le plan de la recherche pour montrer ultérieurement en référence à celle-ci nos contributions dans ce domaine.

2.1 Les intermédiaires traditionnels

Le rôle des acteurs traditionnels comme la presse, les agences RH et les agences Intérim, agissant depuis de longue date sur le marché du travail, se trouvent menacé depuis l'arrivée d'Internet et le développement des techniques de diffusion d'offres d'emploi par le Web. Bien que la Presse et les agences RH ne soient pas des intermédiaires directs du marché du travail, car elles ne mettent pas en relation demandeurs et employeurs, elles mettent cependant à disposition des employeurs des espaces de recrutement susceptibles d'atteindre les candidats potentiels.

L'effet économique le plus déplorable semblerait atteindre l'acteur Presse, qui a observé une baisse de trafic de 20% en trois ans entre 2002 où le taux de consultation des annonces de la presse nationale s'élevait à 57 % et 2004 où le taux de consultation a chuté pour atteindre 37% [1] [24]. Cette baisse est cependant moins évidente côté employeur. Il semblerait que 80 à 90% des chiffres d'affaires en ce domaine reviennent toujours à la Presse. Ceci se laisse expliquer par les prix beaucoup plus élevés des publications des annonces dans la presse écrite que sur Internet, le prix de la diffusion d'une annonce sur le Web coûterait jusqu'à 10 fois moins cher que le prix de la publication dans la presse écrite. Même si la presse prise encore d'une position pionnière en terme de revenus dans ce domaine, elle est loin derrière en terme de quantité d'offres d'emploi diffusées. Les offres publiées dans la presse écrite le sont le plus souvent doublement : à la fois dans les sites d'emplois ainsi que dans les sites carrières des entreprises.

D'après un rapport publié par *les Echos*, le quotidien aurait perdu entre l'année 2000 et 2004 plus de 80% de son chiffre d'affaires dans cette activité. Pareillement pour le quotidien français *Le Figaro* qui aurait vu le nombre d'espaces publicitaires d'offres d'emploi vendus, chuter de 17 000 espaces en Septembre 2001 contre une moyenne de 3 500 en 2004.

La Presse s'est donc vu confrontée à un choix entre 2 solutions aussi peu attractives l'une que l'autre pour contrer ce manque considérable de chiffre d'affaires :

- renoncer à la part de marché de la médiation des offres d'emploi et perdre sa notoriété d'intermédiaire sur le marché du travail. Solution favorable à court terme dans une phase de crise dans les médias.
- basculer sur le nouveau marché et se lancer dans la médiation des offres d'emploi sur Internet. Solution défavorable à court terme car accentuerait ses difficultés financières présentes et à court termes.

[1] Ces statistiques ont été réalisées par TNS-Sofres à la demande de Regionsjobs en 2004

Les principaux groupes de presse ont tout de même opté pour la solution de migration, ils ont fait leur entrée sur le marché de la publication des offres d'emploi en ligne. Cependant, et comme nous le détaillerons ultérieurement, le type d'activité est largement différent. Ils doivent passer d'une activité de simple vente d'espaces publicitaires à des activités de services de suivi des candidats de par la gestion d'une Cvthèque, de l'envoi automatique de mails et d'intermédiation physique entre les deux parties concernées. Ils ont donc opté pour la plupart pour la solution de rachat de sites emploi déjà établis. Le site emploi *Cadremploi* créé en 1990 sur Minitel et présent depuis 1996 sur le net a été racheté par la régie publicitaire du Figaro. Le Monde quant à lui a repris la technologie de Keljob en proposant en 2004 le site emplois *Talent.fr*.

2.2 Le marché du travail et Internet

Les technologies internet ont fourni au marché du travail les moyens d'une large diffusion et de traitement de l'information en temps réel ainsi qu'une réduction considérable des coûts pour les différents acteurs. Au milieu des années quatre-vingt-dix on observe l'arrivée d'une multitude de sites emploi ayant pour objectif de centraliser l'information et de fournir aux acteurs du marché de l'emploi une plate-forme commune de communication moins formelle et plus flexible. Employeurs et candidats peuvent publier ou consulter les offres et les CV depuis leur poste de travail. Côté employeur, les avantages sont la flexibilité de diffuser, corriger et retirer les offres en temps réel à des prix très faibles par rapport aux publications à travers les intermédiaires traditionnels et côté candidats, les avantages sont l'accès rapide et facile aux offres disponibles sur le marché.

Une offre d'emploi diffusée dans la presse écrite a une durée de vie très longue elle ne peut être mise à jour ni être effacée si le besoin n'est plus actuel. Les entreprises doivent compter sur la réception de candidatures pendant des mois après la parution des offres, or leur traitement est très coûteux bien qu'il n'existe pas de loi obligeant les entreprises à répondre positivement ou négativement aux candidatures reçues. La plupart d'entre elles prennent le temps d'étudier les dossiers (car toujours à la recherche du mouton à cinq pattes) c'est d'ailleurs pourquoi certaines entreprises préfèrent diffuser leurs annonces dans des intermédiaires peu connus.

Force est de constater que la contribution d'Internet à l'extension du marché du travail est considérable, cette technique a permis la fluidification de l'information de par la multiplication des vecteurs de diffusions et en a ainsi facilité et accéléré l'accès. Selon Fondeur « Internet permettrait la constitution de sorte de *bourses à l'emploi* dans lesquelles employeurs et candidats entreraient directement en contact » [24]. On observe une croissance de plus de 230% en 3 ans dans le nombre d'offres d'emploi diffusées,

FIG. 2.1 – Évolution du nombre des offres d'emploi diffusées sur Internet

d'après le baromètre semestriel publié par *Keljob* [2] il y avait plus de 400 000 offres sur Internet en France en 2006 contre 177000 en août 2003.

Cette évolution a également permis de gagner de nouveaux acteurs sur le marché du travail, il s'agit d'*individus en emploi* qui viennent augmenter le public des candidats. Ce sont des personnes occupant déjà des postes mais s'informant en temps réel de l'évolution du marché pour d'éventuels changements de stratégies de carrières. Dans la conjoncture actuelle où l'on constate une préférence des entreprises pour les candidats expérimentés, spécialistes, ces nouveaux candidats dit passifs représentent une évolution positive leur permettant d'atteindre les prétendants correspondant le mieux à leurs besoins.

On observe sur le net une multitude de bourses à l'emploi que l'on peut classifier sous quatre catégories principales :

1. les *Job-boards*
2. les *agrégateurs d'offres d'emploi ou les moteurs de recherche d'emploi*
3. les *sites carrières*
4. les *blogs emploi*

2.2.1 Les Job-boards

Les job-boards sont les sites emploi développés aux États Unis dans les années quatre-vingt-dix et par extension tous les services actuellement proposés. Une traduction littérale de job board serait « panneau emploi », qui

[2] source de l'étude à voir sous http ://edito.keljob.com/recruteurs/articles/1206/barometre-octobre.html,http ://edito.keljob.com/index.php ?id=27#1595

d'ailleurs exprime bien l'idée de base véhiculée par ces derniers qui est de fournir un espace d'affichage d'offres d'emploi en temps réel. Les job-boards jouent un rôle d'informateur et non pas directement de médiateur, ils rapprochent certes les employeurs des demandeurs mais uniquement par groupement des annonces d'emploi dans un espace réservé accessible par les deux parties concernées, ils n'ont pas pour métier de les mettre en contact.

Certains d'entre eux considèrent même que leur métier d'infomédiaire pourrait être complètement automatisé du fait que les recruteurs remplissent eux-même les formulaires préétablis pour la publication des offres et que les candidats se chargent de remplir les formulaires de dépôt de CV. Aucun traitement d'informations supplémentaires n'est nécessaire. Des algorithmes de concordance de chaînes de caractères classiques sont largement suffisants pour chercher dans l'index des offres d'emploi celles correspondantes aux termes de la requête, du fait que celles-ci aient été introduite dans la base de données à travers le remplissage manuel de formulaires. Il est donc possible d'affiner les requêtes en précisant ou en limitant la recherche à certains champs du formulaire sans qu'il y ait besoin de faire appel aux techniques de Traitement Automatique des Langues Naturelles(TALN). Ils prétendent par conséquent que le modèle technique utilisé est indépendant des cultures, des langues et des pays. C'est le cas du géant *Monster* implanté dans plus de 37 pays dans le monde et bien que leader dans la plupart de ces pays d'implantation au niveau du marché de l'e-recrutement, il se limite d'employer un maximum de trois spécialistes techniciens dans chacun de ces pays, la plus grande partie du personnel se compose de spécialistes Marketing, de commerciaux et de spécialistes de communication.

Les job-boards rencontrés sur le Web proposent l'information continue de façon presque à temps réel et de façon neutre et automatique. Les offres sont classées selon leurs dates de publication et non selon le prix payé pour la publication, ce qui donne la possibilité aux petites entreprises de se faire connaître et de gagner de la notoriété à des coûts très bas.

La nature de l'activité d'un job board implique une tendance à la centralisation et à la généralité. Il s'agit srouper le plus grand nombre d'offres disponible sur Internet pour ainsi gagner une audience maximale. *Monster* a su par exemple répondre à ce besoin en restant le plus généraliste possible, en centralisant les offres quelque soient les régions, les secteurs d'activité et les segments professionnels. Or la généralisation n'est pas le seul moyen de gagner de l'audience, un certain nombre de job-boards ont adopté une stratégie de spécialisation et de ciblage spécifique de l'audience pour faire face aux géants comme *Monster*. La spécialisation peut être régionale, professionnelle ou sectorielle, une telle focalisation permet d'offrir aux candidats ciblés des services plus adéquats à leurs besoins. La plupart des job-boards spécialisés disponibles sur le marché français ont choisi une hybridation de ces critères, ils ont tout d'abord opté pour une audience territoriale, qui leur permettent de mieux répondre aux besoins nationaux spécifiques aux

français, accompagné de cette première restriction ils choisissent encore une restriction au niveau professionnel, comme c'est le cas de *lesjeudis* pour les métiers de l'ingénierie informatique ou *super-secrétaire* spécialisé dans les métiers du secrétariat et de l'administration, au niveau sectoriel également comme le job board *Jobtech* spécialisé dans le secteur des BTP et industrie ou encore *Jobtranport* pour les secteurs de transport et logistique mais la majorité des job-boards actuels spécialisés dans les secteurs d'activités sont des job-boards pour les secteurs informatique, télécommunications et multimédia (*e-recrut, 01net ...*). Une des spécialisations particulière au territoire français concerne le segment cadre, cette spécialité est très particulière à la France d'où émerge ce terme. Ceux-ci comme *Cadremploi* ou *CadresOnline* se focalisant sur la publication des offres destinées aux cadres.

Classiquement un job board propose deux services principaux qu'il fait payer à l'employeur :

1. La diffusion des offres d'emploi
2. La consultation de la CVthèque

La norme étant que les entreprises paient pour publier leurs annonces de postes vacants et que les candidats puissent les consulter gratuitement. L'envoi des mails aux candidats est également un service gratuit, il fût le service qui permit aux job-boards de faire leur entrée en force dans le domaine des intermédiaires sur le marché du travail.

Cependant certains acteurs ont adopté un modèle économique différent, ils font payer les candidats pour la consultation des offres alors que la publication est gratuite pour les entreprises, c'est ce qu'on appelle un modèle B2C -Commerce grand Public- [3]. Ce modèle n'a cependant pas pu survivre longtemps sur le territoire français surtout avec la sortie en Janvier 2005 de la loi de cohésion sociale qui interdit de faire payer la consultation des offres d'emploi et ce quelque soit leur type.

Cadremplois fondé en 1990 sur le Minitel suivait un modèle B2C et a continué même en basculant sur internet en 1996, seulement face à la concurrence des plus grands job-boards comme *Monster* disposant déjà d'une quantité très importante d'offres dont la consultation était gratuite, ils ont dû revoir leur modèle et ont migré vers un modèle économique B2B.

Sur le territoire américain en revanche on observe le succès de *The Ladders* qui s'est spécialisé dans les jobs à plus de 100K*Dollars* par an et qui continue à gérer un modèle où les candidats payent un abonnement mensuel ou semestriel pour consulter les offres qui, elles, sont gratuitement postées par les recruteurs.

[3] B2C *Business to Customer - commerce grand public* est un modèle économique impliquant d'une part une entreprise et de l'autre des clients privés il se différencie du modèle B2B *Business to Business -commerce inter-entreprises* qui met en jeux deux formes juridiques indépendantes

La seconde source de revenu des job-boards après la publication des offres est la consultation des CV. L'avantage de la consultation de la CVthèque est à observer chez les cabinets de communication RH qui sont à la recherche de candidats particuliers pour leurs clients tout en assurant un anonymat sur le véritable client en besoin de personnel. Les recruteurs finaux se contentent en général d'une approche passive dans laquelle ils diffusent les offres et attendent la manifestation des candidats. Ceux-ci n'ont souvent pas le temps de chercher eux-mêmes leurs candidats dans une banque de données, de plus le coût de la consultation de la CVthèque est largement plus élevé que le coût de la diffusion d'une offre.

Nous incluons les deux sites emploi institutionnels *Anpe.fr* et *Apec.fr* dans la catégorie des job-boards bien qu'ils suivent des modèles économiques différents. Les deux intermédiaires du marché du travail *ANPE* et *APEC* sont des sites financés directement par l'état et ayant pour but commun de rendre l'information sur le marché de l'emploi la plus transparente possible. Elles ont comme objectif premier de réduire le taux de chômage et jouent par conséquent un rôle médiateur entre recruteurs et demandeurs d'emploi. Leurs services sont gratuits aussi bien pour les employeurs lors du dépôt des offres que pour les demandeurs d'emploi lors de la consultation : Ces institutions sont des intermédiaires directs entre entreprises et candidats. Les entreprises recrutant des chômeurs inscrits à l'*ANPE* bénéficient entre autre d'une aide financière étatique pendant la première année d'embauche du candidat.

2.2.2 Les agrégateurs d'offres d'emploi

Les agrégateurs d'offres d'emploi sont les espaces Web permettant une recherche simultanée sur plusieurs sites emplois. L'idée de base d'un agrégateur est de centraliser l'accès aux données et permettre à un utilisateur un aperçu global de l'offre générale disponible sur les différents sites emploi à partir d'un point d'attache unique. Ils se caractérisent par le fait qu'ils n'indexent pas les offres sur leurs sites mais redirigent les utilisateurs directement sur les sites émetteurs des offres d'emploi. Les sites référencés sont le plus souvent les job-boards et les sites carrières.

Les agrégateurs d'offres d'emploi se différencient des job-boards en deux points essentiels, l'un d'ordre économique et l'autre d'ordre technologique. Au niveau du modèle économique on distingue une différence entre la France et les États-Unis. En France, le service est facturé directement aux émetteurs des offres, en fonction des clics ou alors en fonction du nombre d'offres référencées. Le payeur n'est plus l'annonceur mais le référenceur, ce qui dans le cas des sites carrières représente une seule structure. *Keljob* et *Optioncarriere* sont les deux agrégateurs les plus populaires sur le territoire français. Aux États-Unis en revanche le service de référencement est gratuit, la source essentielle de revenus des agrégateurs est la publicité contextuelle et

évènementielle. Le modèle pionnier dans ce domaine est le cas d'*Indeed et SimplyHired*.

La distinction technologique des agrégateurs par rapport aux job-boards est qu'ils utilisent à la manière des moteurs de recherche, des robots d'indexation qui parcourent les sites d'emploi pre-sélectionnés à la recherche des pages emploi.

Il est force de constater que la plupart des agrégateurs ont recourt actuellement aux deux technologies conjointement : la recherche à travers les robots d'indexation et la collecte d'offres directement à travers les employeurs. C'est pourquoi nous préférons pour ce type de sites emploi l'appellation « moteurs de recherche d'emploi ».

Un tel système doit disposer d'un taux élevé d'offres référencées pour être concurrentiel face aux géants du marché et si les sites emploi spécialisés à faible audience voient en une telle collaboration des avantages précieux, les job-boards de grande renommée et de grand trafic ne voient pas l'intérêt d'être référencés sur les agrégateurs et sont souvent contre cette idée. *Monster*, qui publie à lui seul un nombre très élevé d'offres, a toujours refusé d'être référencé alors que l'*Apec* qui est une association initialement syndicale a toujours accepté et ce car elle a comme objectif la baisse du chômage sur le territoire et pense que plus le marché du travail est transparent et accessible mieux les offres atteignent leurs cibles.

Au début de l'apparition des agrégateurs, ceux-ci se sont permis comme c'est le cas des moteurs de recherche généralistes de référencer toute sorte de job-boards ou de sites emploi institutionnels sans pour autant en avoir reçue une autorisation explicite. Cela a d'ailleurs valu un procès judiciaire à *Keljob*[4] avec les trois sites *Anpe.fr*, *Cadremploi* et *CadreOnline*. L'argument majeur de *Keljob* fût que les offres référencées n'étaient pas directement stockées sur leur propre serveur mais ils redirigeaient automatiquement utilisateurs vers les sites émetteurs, néanmoins les plaignants, *Cadremploi* en l'occurrence voyait en ce phénomène une intrusion dans la propriété intellectuelle car les liens référencés étaient des liens profonds accessibles par requêtes sur leurs bases de données. Le procès finit en faveur des plaignants et par la promulgation d'une loi d'interdiction de référencement des offres d'emploi provenant d'un site emploi tiers sans son autorisation. Aux États Unis une telle loi n'existe pas et les agrégateurs ne sont pas limités.

2.2.3 Les sites carrières

Les sites carrières sont des espaces réservés au recrutement sur les sites des entreprises. Suivant la taille et l'importance de l'entreprise le site carrière varie entre la simple liste des postes vacants disponibles et le vrai moteur de recherche d'emploi interne fournissant parfois même des fonctions avancées

[4]premier agrégateur d'offres d'emploi en France apparu en 2000

de recherches et de sélections, des fonctions de dépôt de candidatures ou de dépôt de CV.

On observe une croissance importante des espaces d'offres d'emploi disponibles directement sur les sites Corporate des entreprises. Une étude réalisée par l'ANPE en 2003 dans le cadre du *Recrutement par Internet* indique que 45% des établissements, toutes tailles confondues, disposaient de sites Web et que 54% y publiaient leurs offres d'emploi et d'après les résultats dévoilés sur l'étude faite par Keljob en 2005 sur les *« pratiques de recrutement en ligne des 1000 premières entreprises françaises »* le nombre d'offres d'emploi publiées directement sur les sites carrières des entreprises augmentait d'un total de 13.400 offres en 2004 contre 25.000 offres en 2005. Avec la naissance des agrégateurs verticaux définis dans la section précédente, il devient possible à des entreprises d'audiences limitées d'augmenter leur visibilité à des coûts très réduits sinon inexistants.

Pour répondre à cette tendance, L'ICANN "Internet Corporation for Assigned Names and Numbers" a introduit en 2005 une nouvelle extension " .jobs". Cette extension serait réservée aux sites carrières et chaque entreprise a la possibilité de déposer ses offres d'emploi sur son site sous le domaine *NomEnt.jobs* pour ainsi faciliter aux agrégateurs d'offres d'emploi, que nous appelons dorénavant Moteur de recherche d'emploi, de trouver aisément les offres sur les sites carrières.

2.3 Les Blogs emploi

Le e-recrutement 2.0 fait son entrée, un nombre important de bloggeurs RH lance des Start-up de recrutement en ligne dit e-recrutement2.0, il s'agit des *jobblogs*.

La nouvelle mode des blogs emploi est en pleine expansion en France, on voit se créer de plus en plus de « Start-up » fonctionnant techniquement à la manière des blogs et ayant pour thématique celle des job-boards spécialisés. Qu'est ce qu'un blog et en quoi les jobblogs se différencient-ils des job-boards traditionnels ?

Un *blog*[5] est la contraction de *Web* et *log*, il s'agit pour un individu de tenir une sorte de journal personnel sous forme de site Web. Un blog a une thématique spécifique sur laquelle l'auteur publie régulièrement des bulletins pouvant être commentés par les lecteurs.

La publication est facilitée par l'emploi d'un logiciel spécialisé permettant de mettre en forme les textes et les illustrations, de construire des archives automatiquement, de faire des recherches au sein de l'ensemble des billets

[5]Le mot *blog* fût francisé par l'Office québécois de la langue française (OQLF) en *blogue* pour permettre d'induction des différentes formes comme bloguer, blogeur etc. En France en revanche, la *Commission générale de terminologie et de néologie* a opté pour le mot *bloc-notes*

mais aussi de gérer les commentaires. Ces logiciels permettent à des non spécialistes de tenir une présence sur Internet sans pour autant avoir des connaissances préalables sur la conception de sites Web.

Les blogs emploi peuvent être classés en deux types. Le premier type est publié par des individus à la recherche d'emploi ou ayant été à la recherche d'un emploi et profitant des blogs pour partager leurs expériences dans ce domaine avec la communauté intéressée, ce qui a valu à certains bloggeurs de par leurs réflexions intéressantes de se faire remarquer par des recruteurs. On trouve aussi dans cette catégorie des spécialistes sur la question de l'emploi et des acteurs de RH qui suivent l'actualité sur le marché du travail et du recrutement et contribuent ainsi à l'ouverture du marché par la publication régulière sur leurs blogs emploi de notes sur la question [6].

Le second type de blogs emploi est de type entreprise, il s'agit de Start-up de recrutement en ligne à la manière du Web2.0 qui exploitent les avantages technologiques des blogs pour afficher et distribuer les offres et les profils des candidats. Ils mettent à disposition des candidats des agents de recherche paramétrables qui distribuent par flux RSS les offres filtrées à temps réel, ils restent ainsi en phase avec les job-boards par le service de l'« alerte-email ». L'avantage majeur de tels systèmes est la capacité à générer des candidatures ciblées et ce car ils ont la certitude d'être lu par un public en phase avec leurs annonces dans un marché ou seul le chiffre est important au dépend de la qualité.

Reprenons l'exemple concret de *houra.fr*[7] qui a fait l'expérience de publier son annonce dans le site *cadremploi.fr* et lui a valu le traitement de 500 candidatures pour un poste de chef de projet Marketing et « Aucun ne correspondait au profil voulu » -d'après le directeur commercial et Marketing-. La même proposition publiée sur le blog *media-job.net*, spécialiste des jobs dans le domaine des médias n'a recueilli que 50 candidatures dont 3 ont débouché sur des entretiens. « L'annonce est bien plus détaillée sur un blog, les profils des postulants n'en sont que mieux ciblés. C'est un sacré filtre de sélection pour les recruteurs », se félicite-t-il.

Les jobblogs ne sont certainement pas concurrentiels aux 100 job-boards français les plus connus, mais j'estime qu'ils méritent d'être cités dans un aperçu sur les intermédiaires du marché du travail sur Internet. Nous ne parlerons d'ailleurs pas des blogs emploi dans le comparatif proposé ultérieurement entre les sites emploi les plus populaires en France.

L'argument majeur des jobblogs est que les sites emploi traditionnels font une course aux chiffres les plus élevés en terme de trafic, de nombre d'offres publiées de nombre de CV disponibles et ne s'intéressent pas à la qualité des services d'intermédiation qu'ils proposent. Les recruteurs reçoivent à travers

[6] Blogs emploi intéressants : http ://erecrutement.wordpress.com/tag/analyses-globales-divers/ ou encore http ://altaide.typepad.com/jacques$_f roissant_a ltade$/2006

[7] `http://www.strategies.fr/archives/1387/page_26477/management_blog_mode_d_emplois.html`

ce genre d'intermédiaire, comme mentionné dans l'exemple précédent, une grande quantité de candidatures dont pas plus de 3 % ne répondraient à leurs besoins. Les jobblogs insistent sur leur ciblage plus restreint des candidats de par la thématique spécifique du blog. Or je pense qu'il suffit que ces jobblogs gagnent un peu plus de notoriété pour retomber dans le même problème. En effet le nombre de candidats est proportionnel au trafic du site. De plus ils insistent sur le fait que les offres d'emploi publiées sur les jobblogs sont plus détaillées mais qu'est ce qui empêche les recruteurs de détailler leurs offres avant de les poster sur les autres sites emploi ? La diffusion par RSS a déjà été adoptée par plusieurs sites emploi comme *Cadremploi ou Keljob*, mais si le but est de préserver d'un grand nombre de candidatures inadéquates, en quoi est-ce que leur distribution sur plus de vecteurs informationnels diminuerait cette tendance, au contraire, si les offres peuvent être distribuées par RSS, elles sont alors atteignables plus facilement car disponibles sur différentes plateformes.

2.4 Étude comparative des sites emploi les plus populaires en France

Dans la description et la comparaison présente, nous nous intéressons aux sites emploi côté utilisateur, nous montrons à travers des requêtes lancées sur les sites choisis les avantages et les faiblesses de ces derniers au niveau du traitement d'information, de l'indexation et de la recherche des offres. La description ne s'étend pas sur le côté ergonomique des sites, bien que souvent les résultats de recherche sont satisfaisants mais l'ergonomie laisse à désirer de par les pages trop pleines d'information, ou la multitude de réponses ou de clics nécessaires avant d'atteindre la ou les pages souhaitées.

FocusRH publiait en 2006 le *« guide des 500 meilleurs sites emploi »* en France dans lequel les auteurs présentent les informations clé des 500 sites choisis pour être les meilleurs selon 4 critères essentiels : l'originalité du design, la richesse du contenu, la fraîcheur des offres et la qualité de l'ergonomie. Les sites sont classifiés par type (job board, agence Intérim, société de communication RH) et chaque description apporte des informations sur plus de 20 critères comme le type de candidat cible (étudiants, jeunes diplômés, cadres, dirigeants, professions intermédiaires, artisans, ...), la façon de postuler, le nombre d'offres d'emploi ou le nombre de stage disponibles, etc.

Nous ne nous intéressons pas à un recensement des sites mais uniquement à comparer les sites emploi les plus usités en France pour montrer leurs faiblesses et ultérieurement nos contributions et nos améliorations dans ce domaine. Nous avons donc choisi 9 des sites emploi les plus populaires et ceux contenant le plus d'offres. Les offres sont souvent comptabilisées plusieurs fois par chaque site, c'est pourquoi le nombre de 900 000 offres d'emploi disponibles en 2006 déclaré par *FocusRH* reste invérifiable et incertain. Nous

pensons qu'ils ont simplement additionné les nombres d'offres disponibles dans les 500 sites choisis sans s'intéresser à leur exclusivité.
Dans ce premier tableau, nous récapitulons les informations générales des 9 sites choisis, que nous avons trié par ordre décroissant du nombre d'offres d'emploi proposées.

	type	secteur	cible	Nb. offres	CVthèque	alerte email
Anpe	job board	généraliste	tous	**230 050**	oui	oui
Optioncarriere	agrégateur	généraliste	tous	**187 719**	oui	oui
Keljob	agrégateur	généraliste	tous	**56 500**	oui	oui
Monster	job board	généraliste	tous	**28 500**	oui	oui
Cadremploi	job board	généraliste	cadres	**20 500**	oui	oui
Apec	job board	généraliste	cadres	**16 906**	oui	oui
Regionsjob	job board	régional	tous	**16 780**	oui	oui
Directemploi	job board	généraliste	diplômés	**16 600**	oui	oui
Emploiregions	job board	régional	tous	**12 210**	oui	oui

On remarque à travers ce tableau que l'ANPE, l'Agence Nationale pour l'Emploi en France, est celle qui dispose du plus grand nombre d'offres d'emploi et c'est d'ailleurs celle qui renferme le nombre d'offres le plus varié en terme de domaines d'activité et de métiers proposés. On y trouve la majorité des offres de main d'oeuvre et des métiers d'artisans disponibles sur la toile. Dans ce tableau on remarque aussi que les agrégateurs d'offres d'emploi dépassent de loin les job-boards en nombre d'offres d'emploi, pour les raisons cumulatives dont nous avons parlé plus haut dans ce chapitre. L'alerte Email et la Cvthèque sont également des services incontournables pour les intermédiaires sur le marché de l'emploi actuel, ce sont des services proposés par tous les sites que nous comparons.

Une recherche d'emploi dans les sites proposés est dirigée par un certain nombre de champs à remplir ou à cocher qui permettent d'affiner les requêtes des utilisateurs. Ces critères de raffinement de la requête sont en rapport avec la classification interne du site emploi et de sa structure de données interne. Nous listons dans le tableau suivant les différents critères disponibles dans les 9 sites choisis.

	Anpe.fr	**optioncarriere.com**	**keljob.com**	**monster.fr**	**Apec.fr**	**regionsjob.com**	**cadremploi.com**	**directemploi.com**	**emploiregions.com**
Intitulé du poste	×		×						
Code de l'emploi	×								
Domaine professionnel	×	×	×	×		×	×		×
Lieu de travail	×		×	×	×	×	×	×	×
Secteur d'activité de l'entreprise	×				×	×	×	×	
Société			×						
Type de contrat	×		×	×		×	×	×	
Salaire	×				×		×		
Qualification	×					×			
Temps de travail	×			×					
Date d'émission	×				×				
Domaine de formation	×								
Type de formation	×							×	
Expérience	×		×		×	×			
Agents de recherche	×	×	×	×	×	×	×	×	
Mots clés	×	×	×	×	×	×	×	×	×
Exclusion de la recherche									

Dans ce tableau, nous remarquons que les critères de recherches avancées des offres dans les sites emploi sont plus ou moins les mêmes. L'ANPE est de nouveau en tête de par sa classification des offres sur plus que 14 critères comme « le code de l'emploi, le secteur d'activité de l'entreprise, le salaire, le type de formation, etc. ». Les 9 sites présents dans cette comparaison ne permettent pas d'exclure un terme ou une expression de la requête, ce qui pourrait lever l'ambiguïté dans plusieurs situations. Ils ne permettent pas non plus de choisir la section dans laquelle la requête devrait être lancée. Ce détail est très important pour améliorer les résultats de recherche. En effet si l'utilisateur pouvait chercher les termes correspondant à ses compétences et limiter ensuite la recherche sur les champs appropriés dans la base de données, il obtiendrait uniquement les offres dont les compétences exigées se rapprochent des siennes et obtiendrait moins de réponses bruitées non adaptées à ses besoins. Il serait par ailleurs faisable aux sites emploi actuels de mettre en place une telle amélioration car les offres d'emploi y sont introduites manuellement par un remplissage d'un formulaire où les différentes catégories informationnelles sont dores et déjà séparées et identifiables.

2.4.1 Les limites des sites emploi actuels

Les limites majeures observées sur les sites emploi actuels sont dues au manque de traitement linguistique des documents stockés. La langue française étant riche en synonymie et en polysémie, il est nécessaire d'apporter des analyses linguistiques, morphologiques, sémantiques et lexicales à la phase d'indexation des offres pour aspirer à des résultats de recherche très performants. Les sites d'emploi actuels se contentent de méthodes de recherche «plein texte», où les séquences de caractères de la requête sont directement comparées et mises en correspondance avec les documents de la base de données. Or ceci mène à une dépendance accrue entre la performance du système et la capacité de l'utilisateur à bien formuler sa requête et à bien choisir les termes adéquats. Ainsi pour une requête avec le terme «pharmacienne» et une autre avec le terme «pharmacien», les utilisateurs ne devraient pas obtenir des résultats complètement différents comme c'est le cas présentement, car il s'agit simplement d'un même besoin informationnel qui se voit exprimé différemment en fonction du sexe de l'utilisateur . De même pour une recherche sur le terme «ingénieur» qui ne délivre pas les offres proposant des postes d' «ingénieurs».

Les personnes à la recherche d'emploi ne sont pas en mesure de savoir si elles n'ont pas trouvé le poste désiré parce qu'il n'existe pas ou bien simplement parce qu'elles ont usé de requêtes non appropriées. Mis à part ce problème de morphologie, les candidats ne savent souvent pas l'intitulé exact du poste qu'il recherche et utilisent des paraphrases ou des périphrases à la place. Les sites de recherche d'emploi actuels ne sont pas satisfaisants sur ce point non plus.

Pour parer le problème de la polysémie, l'ANPE, propose à l'utilisateur de raffiner sa requête en choisissant la catégorie exacte qu'il sous-entend dans la classification ROME des métiers et des professions. Ainsi si on entre la requête «Glacier» dans le champs correspondant à l'intitulé du poste au niveau de l'interface de l'ANPE, on obtient une liste de classes et de métiers qui correspondent tous au mot glacier et dans laquelle on doit choisir celle qui correspond à nos besoins, comme dans la liste suivante :

- ◇ 11212 Laveur de vitres spécialisé/Laveuse de vitres spécialisée
  - Glacier (vitres)
- ◇ 13221 Employé polyvalent/Employée polyvalente de restauration
  - Glacier
- ◇ 13321 Exploitant/Exploitante de café, bar-brasserie
  - Cafetier Glacier
- ◇ 45122 Opérateur/Opératrice sur machines et appareils de fabrication des industries agroalimentaires
  - Glacier (IAA)
- ◇ 47112 Préparateur/Préparatrice en produits de pâtisserie-confiserie

– Glacier
– Maître glacier

Dans les autres sites emploi, les intitulés des postes ne sont pas normalisés, bien que la nomenclature officielle ROME soit à la disposition de chaque entreprise et que son utilisation peut améliorer de beaucoup les performances de tels systèmes. Ainsi si une personne à la recherche d'un poste de serveur se contente d'une requête avec le mot «serveur», celle-ci s'avère ambigüe et l'utilisateur se voit retourner des offres d' «informaticien spécialiste de serveurs web » par exemple. Si par contre elle venait à raffiner la requête par un terme supplémentaire comme « restauration », ce dernier diminue alors les chances de trouver les offres de serveurs dans l'hôtellerie ou dans les pâtisseries. Notre site d'emploi préféré est celui de l'ANPE, bien qu'il faille *cocher, sélectionner, cliquer* maintes fois avant d'obtenir un résultat : ces étapes permettent néanmoins de lever les ambigüités possibles.

Dans le tableau suivant nous résumons les possibilités de recherche des différents sites en comparaison.

	Anpe.fr	**optioncarriere.com**	**keljob.com**	**monster.fr**	**Apec.fr**	**regionsjob.com**	**cadremploi.com**	**directemploi.com**	**emploiregions.com**
intitulé exacte	+	+	+	+	+	+	+	+	+
Résultat malgré fautes d'orthographes dans la recherche rapide	-	-	-	-	-	-	-	-	-
Résultats malgré fautes d'orthographe dans la recherche avancée	-	-	-	-	-	-	-	-	-
recherche de sous-chaîne	-	-	+		-	+	-	+	-
transformation morphologique	+	-	-	-	+	-	+	-	-

On remarque que les sites emploi les plus populaires en France offrent tous à l'exception de l'ANPE et Cadremploi, une recherche plate des termes de la requête dans l'indexe de la base de données. Ils ne tolèrent pas non plus les fautes d'orthographes et ne proposent pas un module de reconnaissance des fautes.

Conclusion

Internet a contribué, par la mise en place de moyens techniques, à l'élargissement du marché du travail et à la diffusion en grande masse de l'information. Elle a permis également de faciliter l'accès à ces données et de les rendre accessibles aux différents acteurs du marché de l'emploi. Néanmoins cette évolution a échoué à mettre en place un dispositif de centralisation. L'information est certes abondante mais également distribuée sur une multitude de supports de diffusions, elle est même très souvent redondante.

Le *HR-XML* est un ensemble de spécifications basées sur le modèle XML définissant un standard pour la diffusion des offres d'emploi. L'objectif est de faciliter l'échange et le traitement automatique des annonces d'emploi. Cependant ce standard met beaucoup trop de temps à se propager. L'adoption d'un tel format faciliterait la reconnaissance automatique des offres d'emploi. Cependant et comme en témoigne l'expérience faite avec XML, le Web se développe à une vitesse beaucoup plus rapide que la progression d'un tel standard.

C'est pourquoi nous tentons durant nos travaux et dans le cadre du fonctionnement technique d'un moteur de recherche d'emploi de faire appel à des analyses linguistiques du contenu des sites d'entreprises pour reconnaître d'une part les sites carrières automatiquement et y extraire les offres d'emploi qui devront tout d'abord passer par une première étape d'identification avant de passer à la phase d'extraction automatique des informations nécessaires au remplissage de la base de données des offres d'emploi.

CHAPITRE 3

Les moteurs de recherche

> « *Quiconque a essayé un jour d'entrer dans Internet sait qu'il ne faudrait pas parler d'autoroutes de l'information mais plutôt de labyrinthes.* »
> Jacques Attali

Introduction

Face à la croissance exponentielle des données textuelles disponibles sur le Web et au développement considérable du nombre d'utilisateurs du réseau, la recherche d'information nécessite rapidement des outils plus efficaces pour l'extraction des données recherchées par les internautes.

D'après une étude publiée par *Nielsen-NetRatings*, spécialiste dans l'analyse d'audiences Internet, *Google* aurait dépassé en novembre 2006 la barre des 3,1 milliards de requêtes, soit un total de 103 millions de requêtes par jour et de 1193 requêtes par seconde et ce uniquement aux États Unis où le moteur de recherche *Google* bien que leader du domaine, ne couvre que 49,5% de la part du marché. Ainsi rien qu'aux États Unis, on compte à partir des 10 moteurs de recherche les plus utilisés un nombre total de 5,8 milliards de requêtes au mois de Novembre 2006 toujours selon le rapport publié par *Nielsen-NetRatings*. Sachant de plus que le nombre de sites Web par habitant en 2002 se situait aux État unis à la 4ème position avec seulement 64 pour mille derrière l'Allemagne qui dépassait déjà le taux de 82 pour mille de sites Web par habitant sur la toile. On peut imaginer l'ampleur des recherches sur Internet effectuées dans le monde et l'importance de ce média comme source d'information. Il est par conséquent primordial de s'intéresser de près aux moteurs de recherche actuels et de détecter leurs limites pour ainsi agir à temps dans un domaine en pleine expansion.

Chaque individu ayant déjà cherché une information sur Internet à travers un moteur de recherche générique a forcement ressenti une frustration face aux réponses obtenues. Frustration par rapport à la qualité des réponses -trop ou peu nombreuses - mais frustration aussi par rapport à la présentation de celles-ci. Il arrive souvent que l'ensemble des documents retournés résultats contiennent certes tous les mots de la requête mais ne satisfont en rien les besoins informationnels réels de l'utilisateur. Ce phénomène est connu pour être le *bruit* en recherche d'information. Le cas contraire du *silence* est aussi fréquent, il s'agit du cas où le nombre de documents retournés par le moteur de recherche est très faible ou nul.

Cette frustration est due au fait qu'un utilisateur n'est pas forcement spécialiste du domaine pour lequel il recherche de l'information, il use donc dans sa requête de mots clés non représentatifs du domaine à explorer. Ce problème a aussi une dimension, qui n'a rien à avoir avec la requête mais avec les algorithmes de recherche utilisés. Les données textuelles sont traitées dans les moteurs de recherche actuels de la même manière que n'importe quel ensemble d'objets non porteurs de sémantiques ni de relations signifiantes entre les différents composants.
Si on va dans un entrepôt de voitures et que l'on y cherche *une voiture de couleur rouge ET* ayant *3 portes*, il est très facile de sélectionner initialement toutes les voitures ayant 3 portes puis d'extraire à partir de ce sous-ensemble toutes celles qui sont peintes en rouge. Alors qu'il est moins évident de retrouver des informations sur *l'effet de serre* si l'on cherche dans un entrepôt de données, des documents contenant le terme *effet ET* le terme *serre* sans tenir compte de la relation particulière et de la sémantique qu'ils portent quand ils apparaissent dans ce composé.

Ainsi face à la croissance du volume des informations disponibles sur Internet, un système de recherche doit être capable de récolter l'information éparpillée sur tout le réseau mais aussi d'extraire à partir de cette masse gigantesque l'information pertinente répondant le mieux aux besoins des utilisateurs. Dans cette section nous nous intéressons tout d'abord au fonctionnement des moteurs de recherche génériques pour nous concentrer ensuite sur leurs limites, et montrer l'intérêt des moteurs de recherche spécifiques ainsi que l'intérêt d'intégrer les méthodes linguistiques d'analyse du langage naturel dans de tels systèmes.

3.1 Fonctionnement général

Un moteur de recherche est un service de recherche d'information dans lequel les fonctions principales de collecte, d'analyse et d'évaluation des documents ainsi que les fonctions de gestion de l'indexation et de recherche sont complètement automatisés [28]. Un moteur de recherche est composé de 3 sous-systèmes indispensables à son fonctionnement :

Collecteur de documents ou Robot d'indexation
Un moteur de recherche dispose d'une composante logicielle appelée souvent *crawler, spider, robot d'indexation* qui parcourt le Web activement à la recherche de nouveaux documents. Cette composante n'existe pas par exemple dans les catalogues Web où les données sont récoltées manuellement à travers le remplissage de formulaires.

Analyseur et Évaluateur de documents
La seconde composante d'un moteur de recherche est le système d'analyse, de traitement et d'évaluation des documents recueillis par le Crawler dans la phase précédente. Il s'agit d'un ensemble de composantes logicielles capables d'évaluer les documents sur certains critères et de décider de leur admission ou de leur rejet, ainsi que de la manière de les indexer. Les méthodes des Systèmes de Recherche d'Information (SRI) basées sur les mots clés sont en général utilisées.

Processeur de requêtes
Cette composante est chargée de la gestion des requêtes utilisateurs, de leurs transformations et mise en correspondance avec la collection documentaire. Le processeur de requêtes recherche dans l'index des documents ceux compatibles avec les termes de la requête et use de méthodes de Recherche d'Information (RI) pour calculer la pertinence des différents documents à rendre à l'utilisateur.

Nous décrivons dans la suite de cette section les trois phases de traitement d'un moteur de recherche. Nous commençons par l'analyse des méthodes de découverte et collecte d'information à partir du Web puis nous présentons certaines méthodes utilisées par les moteurs de recherche pour la compréhension des documents écrits en langue naturelle ainsi que des méthodes de transformation dans des formats capables d'être explorer par un langage de requête partageant un même formalisme interne.

3.1.1 Le robot d'indexation

Le but ultime des grands moteurs de recherche est de repérer le nombre le plus important de documents à partir du WWW, ceci implique une maintenance permanente face à un Web qui évolue en exponentiel et face aux changements constants des documents déjà répertoriés. Afin de satisfaire cet objectif un tel système doit disposer de méthodes permettant d'une part d'identifier de nouveaux documents sur la toile et d'autre part d'identifier les changements survenus dans ceux déjà référencés. Une solution s'est imposée dans ce domaine, il s'agit des robots d'indexation, dits aussi « Crawlers, Spiders ou agents d'indexation ». Les agents d'indexation sont des composantes en fonctionnement constant, ayant pour mission de parcourir les hyperliens qui se référencient les uns les autres à la recherche de nouvelles *URLs*[1] à

[1] une URL - Uniform Resource Locator- est l'adresse unique d'une page Web sur Internet

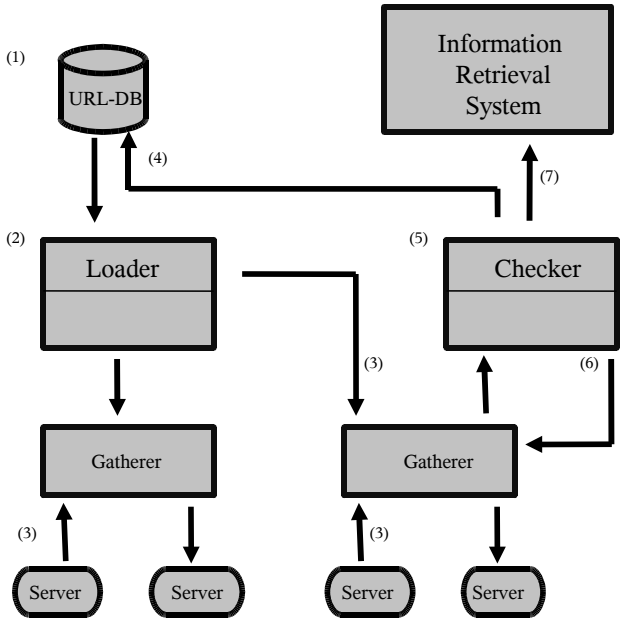


FIG. 3.1 – Fonctionnement d'un robot d'indexation

partir d'une liste de liens hypertextes initiaux et de vérifier le contenu de celles déjà identifiées. Ces robots sont également paramétrables pour limiter le traitement à un certain type de documents ou de protocoles et en exclure d'autres. L'illustration 3.1[28] montre le fonctionnement d'un tel robot, on y distingue :

1. Assembleur (Gathrerer)
2. Chargeur (Loader)
3. Base de Données des URLs
4. Vérifieur (Checker)

L'assembleur Le rôle principal de l'assembleur est d'actualiser la collection documentaire et de l'augmenter continuellement. Cette mise à jour compte aussi bien la vérification de la persistance des documents que la reconnaissance des modifications. Les moteurs de recherche tentent de visiter le plus souvent que possible les documents déjà indexés pour reconnaître les évolutions au plus vite et éviter aux utilisateurs de mauvaises surprises. L'initialisation du processus de *Crawling* se fait grâce aux URLs disponibles dans la base de données et des modules sont chargés d'extraire tous les liens hypertextes qui y sont référencés dans le but d'en détecter de nouveaux. L'hypermédia étant un réseau mondial de liens hypertextes, il est théoriquement possible à un moteur de recherche d'atteindre toutes les URLs du réseau

en poursuivant au fur et à mesure les nouveaux trouvés dans les documents connus.

Le chargeur La fonction du chargeur est principalement la gestion des robots lancés sur les différents serveurs. Il est responsable de la distribution des requêtes aux assembleurs et à l'optimisation des ressources systèmes. Étant donné que les documents n'ont pas tous la même fréquence d'évolution, ce serait de la dilapidation de ressource de les mettre à jour tous à la même fréquence. C'est pourquoi les URLs sont classées dans diverses catégories pour lesquelles les fréquences de mise à jour sont définies en fonction de critères tel que le type de programmation des documents, le type des documents ou la profondeur des documents dans la hiérarchie.

Le vérifieur La fonction principale du vérifieur est d'éviter que des ressources indésirables soient transférées à l'étape d'analyse et de traitement. Des filtres de différentes natures peuvent être employés en fonction des besoins des moteurs de recherches, nous en présentons ici les trois les plus usités :

- ◇ Filtre de document : vérifie le type de données du document à travers le message *Mime-Type* passé dans l'entête HTTP. Certains moteurs de recherche n'indexent pas les documents de format *.doc* ou *.ppt*, dans ce cas ils sont dérivés par ce filtre.
- ◇ Reconnaissance de doublets : vérifie pour chaque document si le contenu a déjà été indexé sous le même nom de domaine ou un autre. Des fonctions de hachage cryptographiques qui permettent de calculer une clé unique de 16Bytes par document sont appliquées. Si le document est le même l'algorithme produit la même clé.
- ◇ Filtre de l'URL : vérifie initialement l'existence de l'URL par le biais de l'entête HTTP, puis vérifie l'existence de caractères de ponctuations comme $, & qui assurent qu'il s'agit d'un lien dynamique et les liens dynamiques ne sont pas indexés par les moteurs de recherche. Un filtre très répandu est la comparaison de l'URL avec une liste d'interdits (Black List). Ceci permet d'écarter les documents non conformes aux règles nationales ou en désaccord avec la licence d'utilisation du moteur de recherche.

Si le moteur de recherche veut mettre à jour les URLs de sa base de données, le robot d'indexation commence par extraire toutes les URLs de la catégorie à vérifier et les transferts au chargeur qui les distribue à son tour sur les différents assembleurs disponibles. Les assembleurs adressent alors des requêtes HTTP au serveur Web et retournent les données reçues au vérifieur qui se charge de supprimer les URLs non valides de la base de données et de transmettre les autres à la phase de traitement et d'indexation. Les nouvelles URLs découvertes sont alors ajoutées à la liste des URLs à traiter.

3.1.2 L'indexation

La phase d'indexation consiste au prétraitement et à l'analyse des documents mais aussi à la construction d'un index de stockage. Nous décrivons dans cette section les techniques d'indexation les plus employées ainsi que les modèles permettant d'extraire des informations à partir de cet index. L'un des principaux problèmes rencontrés par les systèmes de recherche d'information est de mettre en correspondance les documents stockés avec l'information recherchée. C'est pourquoi il est indispensable dans un tel système de s'intéresser de près à l'optimisation de la phase d'indexation.

Les documents de la collection sont initialement convertis dans un formalisme représentatif interne. Ce formalisme est choisi tel qu'il permet une représentation fidèle des milliards de documents retrouvés par le moteur de recherche et tel qu'il supporte facilement les mises à jour. Il est très important de trouver une représentation efficace qui soit compatible aussi bien avec les documents de nature textuelle qu'avec les requêtes également sous forme textuelles. Une fois le formalisme fixé, il s'agit de définir une fonction de correspondance, qui permette la comparaison entre requêtes et documents stockés afin d'y extraire les plus pertinents.

Les documents reçus à partir du robot d'indexation sont tout d'abord délivrés des balises des langages de programmation et de script. La langue naturelle du texte restant est alors dans une seconde phase, identifiée par application d'algorithmes à base de modèles de Markov cachés ou de dictionnaires électroniques. La reconnaissance de la langue des documents permet une meilleure catégorisation de ces derniers et diminue ainsi les cas d'ambiguïté des termes polysémiques dans plusieurs langues.
Une fois la langue identifiée vient l'étape de segmentation du texte et de reconnaissance des unités lexicales correspondantes. Les moteurs de recherche usent à cet effet des signes de ponctuation et des espaces pour reconnaître les limites des termes. Cette méthode qui montre ses limites surtout dans les langues comme le français ou l'anglais où les mots composés sont des suites de graphies séparées par des espaces ou des tirets. La liste des unités lexicales trouvée est ensuite délivrée à un lemmatiseur, qui se charge de remplacer chaque terme par sa racine lexicale. La logique de l'usage des techniques de lemmatisation est que les mots fléchis ne perdent pas leurs significations s'ils sont rapportés à leurs formes de base, il est ainsi équivalent en terme sémantique de retenir les formes standards d'un mot à la place de ses formes fléchies et représente en même temps un gain énorme en terme de ressources de stockage. Ainsi les documents comportant les termes « *mangera, mangeront, mangé* » seront tous indexés par la même unité lexicale « *manger* » ce qui permet d'augmenter le nombre de documents pertinents répondant à toute requête comportant une forme fléchie du verbe « *manger* ».

Une fois le prétraitement terminé, vient le choix de la représentation de stockage des documents, formalisme qui servira également à la transcription

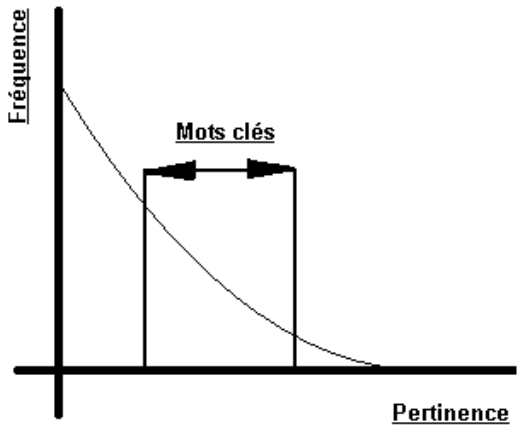


FIG. 3.2 – la distribution de la fréquence des termes d'un corpus

des requêtes dans la phase de recherche. Il existe différents algorithmes de transformation des documents et des requêtes. Cependant la méthode la plus utilisée reste celle dite en *sac de mot* qui revient à représenter chaque document par l'ensemble des mots qu'il contient. Cette méthode montre certaines limites, du fait que tous les mots d'un document ne sont pas porteurs d'informations significatives (voir la loi de Zipf à la figure 3.2). Il est dès lors préférable d'épargner les ressources matérielles et le temps des ressources logicielles en localisant uniquement les mots clés représentatifs du document. Les techniques de choix des mots clés font appel à des algorithmes de sélection d'attributs connu en anglais sous « *Feature selection* ».

Dans ce domaine, il est très fréquent de se baser sur des méthodes statistiques et particulièrement sur la loi de la fréquence des termes. Cette loi, bien étudiée dans la littérature sous le nom de Zipf law [77, 67], est une loi de puissance (le nombre d'occurrences du i-ème terme variant en puissance inverse de i), elle stipule le fait qu'un auteur a tendance dans la rédaction d'un document de reprendre les mots importants à maintes reprises, au lieu de rechercher constamment de nouveaux concepts synonymes, ce qui engendre une fréquence élevée de ces termes dans le document en question mais une fréquence moyenne au niveau de la collection entière. Expérimentalement, l'histogramme des fréquences (Fig. 3.2) montre bien que la collection contient un nombre important de mots rares (termes trop spécifiques), et un nombre réduit de mots très fréquents (les mots vides).

Or face à l'accroissement des documents à stocker, la fréquence ne suffit plus pour retrouver les meilleures similarités entre documents et requêtes, il est important de prendre d'autres critères en considération comme la proxi-

mité des termes, leurs positions dans le document ou encore leur ordre d'apparition. Ces critères commencent à intégrer lentement les moteurs de recherche existants sur le marché.

La liste des mots les plus représentatifs est ainsi sélectionnée, ces derniers sont en général pondérés par la valeur *TF-IDF*, où *TF* représente la fréquence du terme dans le document, et le *IDF* représente l'inverse du nombre de document dans lesquels le terme est présent. Cette pondération est très souvent utilisée en recherche d'information (RI) car elle permet de capturer l'importance d'un mot dans un document et en même temps son pouvoir discriminant par rapport au dit document.

À la fin de cette étape d'analyse et de traitement, on obtient pour chaque document et pour chaque requête un ensemble de mots pondérés censé représenter au mieux leurs contenus respectifs. Il est indispensable à présent de choisir une structure de données permettant de stocker ces représentations et permettant parallèlement de faire efficacement des fouilles dans la collection documentaire afin de trouver des similarités entre documents et requêtes. La structure de données adoptée par la majorité des moteurs de recherche est une structure dite de « fichier inversé » où les documents sont indexés par leurs mots clés identifiés auxquels sont attachés des données inversées, organisées de tel sorte qu'une requête sur le mot clé livre la totalité des documents le contenant. Les mots de l'index sont classés par ordre alphabétique, accompagnés de la position de leurs différentes occurrences dans les documents. Cette méthode a le mérite d'un temps d'accès très rapide. On peut voir une illustration d'une telle structure dans la figure 3.3[28].

3.1.3 Les modèles d'indexation

La création de l'index n'est pas une fin en soit, l'index permet grâce au formalisme qu'il adopte de faciliter la communication entre les requêtes utilisateurs et la collection documentaire. Il est nécessaire dans un premier temps de traduire les requêtes dans le même vecteur représentatif que les documents pour qu'une concordance soit réalisable. La requête subit alors les manipulations de la phase de traitement et d'analyse identique à celles appliquées aux documents avant leur indexation pour obtenir en sortie un vecteur de mots clés lemmatisés à la manière des documents de la collection.

Si l'on observe un système de base de données comme SQL basé sur des tables structurées et non pas des index, on remarque qu'ils suivent une logique de prise de décision binaire, une requête n'obtient de résultats positifs que dans le cas où elle est similaire à 100 % à un enregistrement de la base de données. Les résultats obtenus sont alors des enregistrements qui répondent exactement à tous les critères décrits dans la requête SQL. Effectivement, cette prise de décision résulte d'une comparaison des sous-chaînes de la requête avec le contenu des champs des tables de la base de données.

En fouille de texte un raisonnement flou est prisé, une réponse n'est pas

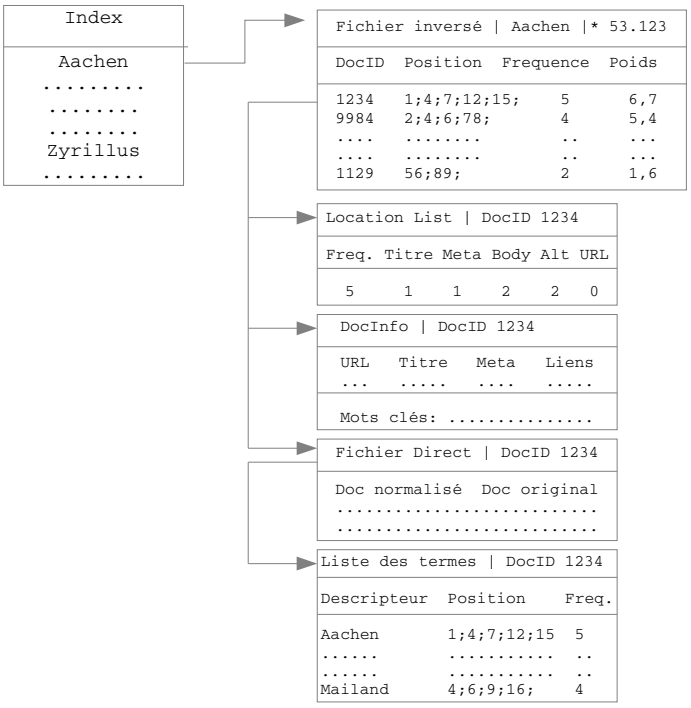


FIG. 3.3 – index

constituée uniquement de documents correspondant à 100% à la requête. Un calcul de pertinence entre un document et une requête est nécessaire. Cette pertinence est une valeur comprise entre 0 et 1, plus le degré de pertinence se rapproche de 1 plus le document est similaire à la requête. Cette différence de raisonnement est due à la nature textuelle des documents et à la phase de sélection des attributs qui suit une approche statistique automatique.
De par la richesse de la langue naturelle, pour exprimer un même phénomène, deux auteurs auront tendance à user d'une terminologie et de procédés stylistiques différents en fonction de la nature officielle ou pas de l'écrit ou du penchant plutôt technique ou littéraire de l'auteur. C'est pourquoi il est inconcevable d'exclure des documents qui ne sont pas à 100% adéquats avec la requête et qu'on calcule en RI des taux de pertinence entre les documents et les requêtes. Ainsi face à une requête, un système de recherche d'information délivre un ensemble de documents pondérés par des degrés de pertinence qui seront classés par ordre décroissant, cette organisation est nécessaire face à la quantité de documents disponibles dans l'index. Nous présentons ici les principales techniques d'indexation des systèmes de recherche d'information et montrons dans la section suivante les principales approches de calcul de pertinence employées par les moteurs de recherche actuels.

Il existe dans ce domaine de la recherche d'information trois modèles classiques pour la classification des documents résultat d'une requête :

1. les modèles booléens ;
2. les modèles vectoriels ;
3. les modèles probabilistes.

On voit naître ces dernières années les modèles linguistiques qui demeurent néanmoins en arrière plan dans les technologies utilisées par les moteurs de recherche que l'on trouve sur le marché. Nous reviendrons ultérieurement sur les traitements linguistiques en RI.

Nous utilisons dans la suite de cette description la représentation formelle suivante :
Soit la collection documentaire D et la collection de mots clés T définis tel que :

$D = \{d_1, d_2, ..., d_i, ..., d_n\}$, l'ensemble des documents de la collection.

$T = \{t_1, t_2, ..., t_j, ..., t_m\}$, l'ensemble des termes retenus comme représentatifs de D.

3.1.3.1 Le modèle booléen

Dans le modèle booléen, décrit par Salton dans [58] : Chaque document d_i, est représenté par la conjonction des termes qu'il contient, sans pondération de ces derniers par des poids d'importance.
La représentation du document numéro 32 par exemple serait : $d_{32} = t_1 \wedge t_6 \wedge t_{84} \wedge t_{132} \wedge t_{798}$
La requête q est aussi formalisée par une expression logique, représentée par la conjonction, la disjonction ou la négation de ses termes entre eux.

$$q = t_1 \wedge t_2 \vee (\neg t_3 \wedge t_4)$$

Pour qu'un document d_i soit candidat pour une requête q, il faut que tous les termes $t_1, .., t_j$ de la requête soient présents dans le document.

Un tel modèle a l'avantage d'être très simple à mettre en oeuvre, et de permettre un calcul très rapide du degré de similarité, c'est ce qu'il lui vaut une grande notoriété et son utilisation dans les moteurs de recherche actuels. Il possède cependant plusieurs inconvénients qui dégradent la performance de ces systèmes. Ce modèle ignore l'ordre d'apparition des termes, ainsi que leur éloignement les uns des autres ; il ne donne pas non plus la possibilité d'exprimer l'importance des termes dans les document par rapport au corpus, seul la présence ou l'absence des mots entre en considération dans le calcul de la similarité. De plus seuls les documents contenant tous les termes de la requête sont considérés comme similaires à celle-ci. C'est à dire si 3 termes sur 4 existent dans un document et qu'il n'y a aucun document qui les

contiennent tous, la requête ne trouvera aucun résultat, ce qui fausse la performance du système, car il se peut que l'un des termes utilisé dans la requête et qui n'a pas été retrouvé, soit peu significatif pour l'information recherchée.

3.1.3.2 Le modèle vectoriel

Dans un modèle vectoriel [59], les documents et les requêtes sont représentés dans un espace à n dimensions, où n est le nombre de termes sélectionnés représentant le document et q le nombre de termes représentant la requête. Ainsi soit le document d_i et la requête q_q suivants :

$$d_i = (w_{1i}, ..., w_{ji}, ..., w_{ni})$$
$$q_q = (w_{1q}, ..., w_{jq}, ..., w_{nq})$$

où w_{ij} est le poids du terme t_j dans le document d_i et w_{jq} le poids du terme t_j dans la requête q. Chaque document et chaque requête sont ramenés à la forme vectorielle dans le repère décrit ci-dessus. Cette représentation commune permet une comparaison et un calcul de similarité plus aisé.

Les approches courantes utilisent pour le calcul de la similarité entre les deux vecteurs ($Vecteur_{document}$ et $Vecteur_{requete}$) la formule de *Dice*, la formule de *Jacquard* ou encore le cosinus de l'angle formé par les deux vecteurs. Le calcul du cosinus reste néanmoins la pratique la plus fréquente.

$$sim(\vec{d_j}, \vec{q}) = \frac{\sum_{i=1}^{k} w_{i,j}.w_{i,q}}{\sqrt{\sum_{i=1}^{k} w_{i,j}^2}.\sqrt{\sum_{i=1}^{k} w_{i,q}^2}}$$

Il revient à faire le produit scalaire entre les deux vecteurs : celui de la requête d'une part et celui d'un document de l'autre. Pour une requête donnée, on calcule son produit scalaire avec chaque document de la collection. Plus l'angle entre les vecteurs (*document, requête*) est étroit, plus le coefficient de similarité est grand. La figure 3.4 [42] illustre la représentation spatiale des documents d_1, d_2 et de la requête q dans un espace à 3 dimensions représenté par les trois termes t_1, t_2, t_3. La valeur du cosinus entre le vecteur de la requête $\vec{q}$ et celui du document $\vec{d_1}$ est supérieure au cosinus entre $\vec{q}$ et $\vec{d_2}$ car l'angle formé entre les deux premiers est plus étroit.

L'avantage de ce modèle par rapport au précédent est que la liste des documents résultats retournée à l'utilisateur est une liste triée par ordre décroissant de pertinence, alors que dans le premier, les documents de la collection sont regroupés dans deux ensembles distincts, celui des documents non pertinents et celui des documents pertinents, sans y introduire aucun degré de similitude par rapport à la requête. Dans le modèle vectoriel, et grâce à l'appariement partiel, on obtient des réponses même si les documents ne contiennent pas strictement tous

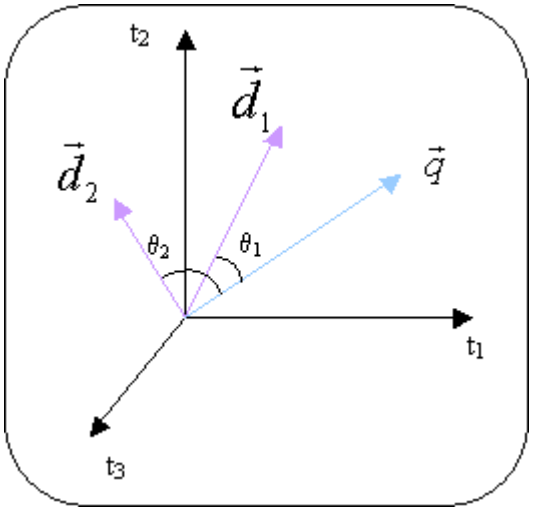


FIG. 3.4 – modèle vectoriel pour 3 documents et 1 requête

les termes de la requête. L'inconvénient majeur de ce modèle ainsi que du précédent est la représentation des documents dite en *sac de mots* qui ne tient pas compte de la dépendance des mots entre eux, ni de leur ordre d'apparition.

3.1.3.3 Le modèle probabiliste

Dans le cadre du modèle probabiliste il s'agit de calculer la probabilité de pertinence d'un document pour une requête. L'idée est de retrouver les documents les plus probablement pertinents et ceux ayant la probabilité la plus faible d'être non pertinents. La fonction de similarité évalue la pertinence du document d_i pour la requête q en usant du théorème de Bayes dans la majorité des cas pour estimer les probabilités tel que :

$$Sim(d_i, q) = \frac{Pr(d_i).P(rel)}{Pr(d_i).P(rel)+Pn(d_i).P(nrel)}$$

où :

- Pr(d_i) représente la probabilité pour d_i d'être un document pertinent
- Pn(d_i) la probabilité pour d_i d'être un document non pertinent pour la requête q.
- p(rel) est la probabilité de pertinence et p(nrel) est la probabilité de non pertinence d'un document quelconque du corpus supposé fixe pour un corpus donné.

Un tel système est difficilement mis en place dans un moteur de recherche en raison du calcul des probabilités initiales qui suppose qu'on doit connaître la probabilité de pertinence sur un échantillon de document qui sera par la suite généralisée sur la totalité de la collection. Or la nature hétérogène des données récupérées par les robots d'indexation à partir du Web rend cette méthode peu applicable sur les moteurs de recherche.

A travers cet aperçu des modèles les plus connus de la recherche d'information, on peut très vite conclure que le manque d'information linguistique au niveau de la représentation des données diminue considérablement la performance de ce type de systèmes.

3.1.4 Les critères de pertinence : tf-idf versus popularité des hyperliens

La phase d'indexation délivre un ensemble de documents pertinents pour la requête utilisateur, il faut encore pouvoir classifier ces derniers par ordre de pertinence afin de présenter à l'utilisateur les plus pertinents en haut de liste. Les moteurs de recherche présents sur le marché utilisent essentiellement deux mesures de pertinence.

La première est celle que nous avons déjà évoqué antérieurement, qui applique l'algorithme de TF-IDF à la phase d'indexation des documents afin de pondérer les termes représentatifs par leur degré discriminant et par leur importance dans le document et dans la collection en entier. Cette méthode purement statistique est basée essentiellement sur la fréquence des termes. Leur pouvoir discriminant est calculé en fonction de leur fréquence dans le document par rapport à leur fréquence inversée dans la totalité du corpus. Cette pondération permet ainsi d'appliquer des mesures de similarité telle que celles de Brice ou Jackard pour produire une réponse numérique entre la requête et les divers documents pertinents.

La seconde méthode utilisée est plus innovante, il s'agit de la *popularité des liens*. Elle part du même principe que pour les publications scientifiques, où l'on considère que les articles les plus pertinents autour d'un thème sont ceux qui sont le plus souvent cités par d'autres auteurs dans la littérature. Similairement, afin d'ordonner la liste des documents résultats, les moteurs de recherche mesurent la popularité d'un site en se fiant à la nature des hyperliens et ce en considérant le nombre de référencement de ce dernier par d'autres sites et le nombre de visite du lien par les internautes comme garantie de pertinence. Plus un lien est référencé plus il sera important et plus un site est visité plus il sera également important.

On distingue donc dans ce concept deux méthodes employées par les moteurs de recherche actuels.

- ◇ **popularité des hyperliens** : Algorithme élaboré initialement par les développeurs de *Google*, connu sous le nom *PageRank*, il suit le principe que plus un document est référencé par les autres sites disponibles dans la collection documentaire indexée, mieux le document correspondant sera placé dans la liste des résultats. Cet algorithme produit une valeur numérique en fonction du nombre de référencement de chaque document de la liste des documents pertinents dans la totalité de la collection. Il est évalué en quatre étapes de calcul de pertinence, de calcul d'un classement initial par pondération des termes avec des méthodes statistiques en fonction de la fréquence du calcul du nombre de référencement et enfin de calcul du rang pour chaque document. Glöggler [28] explique cet algorithme en détail dans son livre sur les moteurs de recherche [28].
- ◇ **fréquence des clics** : Technologie mise en place en 1988 par le moteur de recherche *DirectHit.com*, elle est basée sur l'idée que les sites Web les plus visités par les internautes à partir de la liste des résultats correspondant à une requête sont forcément pertinents et répondent au mieux aux besoins informationnels des internautes face à cette requête. Ainsi, plus la fréquence des clics sur un lien est grande mieux il sera placé dans la liste des résultats d'une requête similaire ultérieure. Une amélioration de la position dans la liste des résultats d'une requête se fait à travers l'augmentation de la fréquence de clics sur les URLs du référant. Tous les clics effectués sur des références dans une liste de résultats seront comptabilisés, cette fréquence est alors sauvegardée dans une base de données accompagnée de l'URL correspondante. Les adresses IP des clients effectuant les clics sont également sauvegardées pour éviter les robots de clics automatiques ainsi que les clics trop répétés de propriétaires de sites Web dans une perspective d'améliorer leur rang. Cette fréquence est en outre rapportée à la durée d'existence du document dans la base de données, ce traitement évite que les nouveaux documents indexés soient lésés par rapport à ceux beaucoup plus anciens.

3.1.5 L'évaluation des systèmes de recherche d'information

Un système de recherche d'information est efficient s'il présente exactement à un utilisateur l'information qu'il recherche. Dans cette perspective deux mesures sont définies pour évaluer la performance et

l'efficacité d'extraction des système de RI. Il s'agit du *Rappel* et de la *Précision*. Le Rappel mesure le pourcentage d'informations correctement extraites, il est le rapport entre le nombre de documents pertinents trouvé et le nombre total de documents pertinents. Un rappel de 100 % signifie que le système a extrait tous les documents pertinents existants. La Précision mesure le pourcentage d'information extraite qui est correcte, elle est le rapport entre le nombre de documents pertinents trouvés et le nombre total de documents trouvés. Une précision de 100 % signifie que tous les documents extraits par le système sont pertinents.

Propriété 1 *Le rappel et la précision sont deux mesures d'efficacité introduites afin de comparer des systèmes différents.*

La précision évalue la proportion de documents pertinents trouvés par un système :

$$Précision = \frac{nombre\ de\ documents\ pertinents\ trouvés}{nombre\ de\ documents\ trouvés}$$

Le rappel évalue le nombre de documents pertinents trouvés par rapport au nombre de documents pertinents disponibles dans le système.

$$Rappel = \frac{nombre\ de\ documents\ pertinents\ trouvés}{nombre\ de\ documents\ corrects}$$

3.2 Limites

En observant le mécanisme par lequel les documents et les requêtes sont mis en correspondance (une comparaison stricte des chaînes de caractères) on peut en conclure que les systèmes de recherche d'information se trouvent très vite confrontés à plusieurs problèmes.

Le premier problème qui n'est pas soulevé par les systèmes classiques que nous venons de décrire est qu'une idée peut être exprimée de différentes manières, ce que certains diront en un mot, sera exprimé en prose par d'autres. Les systèmes classiques seront incapables de rendre un document pertinent qui contiendrait non pas les termes de la requête mais des termes sémantiquement proches. Un système classique traite les mots comme de simples chaînes de caractères, il est donc incapable de traiter la synonymie et encore moins la paraphrase. Si un utilisateur recherche des documents en rapport avec les « *voitures* », il ne pourra pas voir les documents existants dans la base de données

renfermant les termes « *automobiles* » ou « *d'autos* », alors que les trois termes répondraient à un même besoin informationnel.

Le deuxième problème rencontré causé par la richesse de la langue naturelle est la polysémie. Quand un utilisateur entre le mot 'avocat' dans sa requête, il se verra présenter des documents parlant aussi bien du fruit, l'avocat, que de documents en rapport avec le métier de *juriste*.

Ces deux problèmes mènent à des baisses énormes de la performance des systèmes de recherche d'information. La richesse de la langue naturelle en procédés stylistiques, en images, en synonymie, en paraphrase ou encore en polymorphie, fait qu'une recherche textuelle ne peut se limiter à une comparaison simple de chaînes de caractères et fait la limite majeure des systèmes de recherche d'information textuelle présents sur le marché.

Une solution proposée depuis peu d'années est d'introduire les méthodes de traitement automatique de la langue naturelle (TALN), qui permettent l'analyse des termes comme entités linguistiques à part entière et non plus simplement comme de vulgaires chaînes de caractères. Les trois phases classiques d'analyse : lexicale, syntaxique et sémantique peuvent alors être appliquées au niveau du prétraitement des documents et des requêtes pour ainsi découper le texte en unités lexicales de sens complexe et retenir les relations de dépendance entre les mots véhiculés dans les textes. Les documents ne seront plus représentés en sac de mots mais chaque unité lexicale simple ou complexe sera accompagnée de son analyse linguistique, qui entraîne une meilleure compréhension des textes et améliore par conséquent la performance du système.

CHAPITRE 4

L'extraction d'information

> *« La connaissance s'acquiert par l'expérience, tout le reste n'est que de l'information. »*
> Albert Einstein

Introduction

Face à la masse d'information en croissance continue disponible sur le net, une multitude de disciplines aspirant à une meilleure gestion et une meilleure compréhension de ces données sont apparues. Nous nous sommes penchés dans le précédent chapitre sur les méthodes de recherche d'information dans les bases de documents. Nous y avons montré l'intérêt du prétraitement et surtout la nécessité de changer l'espace de représentation des documents initialement écrits en langue naturelle pour un espace représentatif plus structuré exploitable par les machines et les systèmes de recherche en particulier.

Les annuaires électroniques dont les documents sont collectés et classifiés manuellement permettent des recherches mieux ciblées et rendent des résultats plus adéquats aux besoins informationnels des utilisateurs, or le grand handicap de ce type de systèmes est la quantité réduite de documents stockés face à la quantité réellement disponible. Les résultats sont certes pertinents mais se limitent à des domaines de connaissance très restreints. La représentation interne d'un annuaire électronique spécialisé est une base de données classique basée sur un modèle d'attributs valeurs. Ce modèle est rendu possible par le remplissage manuel de formulaires. En effet, prenons le cas d'un annuaire d'organisations, les employés disposent d'un formulaire où ils doivent

préciser, le nom de l'entreprise, son adresse, son domaine d'activité, son siège social, ses coordonnées téléphoniques et fax ainsi que son adresse mail et son URL. Ainsi une recherche affinée par lieu et/ou domaine d'activité permet d'atteindre plus rapidement le besoin de l'utilisateur, car il s'agit ensuite d'une simple requête sur les colonnes des tables de la base de données répondant à ces critères.

Il est certes impensable de réduire le problème de recherche d'information dans les documents textuels à de simples requêtes SQL mais il est possible de rapprocher les deux phénomènes pour améliorer les performances des systèmes de recherche d'information et ce en se focalisant sur la phase de transformation de l'espace de représentation des documents bruts en documents semi-structurés et en améliorant les techniques de choix des termes clés représentatifs dans la phase de sélection des attributs. C'est dans un objectif de conciliation de performance et de complétude que la discipline d'extraction d'information s'est développée.

Les deux notions de recherche d'information et d'extraction d'information sont souvent confondues dans l'usage commun, il est néanmoins très important de distinguer ces dernières qui sont plus complémentaires que comparables. Elles représentent deux disciplines aussi complexes l'une que l'autre.

L'extraction d'information (EI) se focalise sur l'organisation interne des documents. Elle consiste à remplir automatiquement des formulaires à partir de documents écrits en langue naturelle définissant ainsi la structure de stockage interne. L'EI met en oeuvre des méthodes d'analyses et d'interprétations de textes bruts pour la construction de représentations formelles permettant d'apporter automatiquement des réponses précises aux besoins informationnels des utilisateurs.

La recherche d'information (RI) en revanche permet d'identifier parmi un ensemble de documents structurés ceux correspondant au mieux aux requêtes postées pas les utilisateurs. La RI met en oeuvre des mesures de similarité permettant de comparer les documents à la requête et des méthodes d'ordonnancement permettant de classifier les documents du plus au moins pertinents.

L'extraction d'information est donc la fonction d'analyse et de compréhension de documents en langue naturelle (dit aussi textes libres ou documents non structurés). Elle permet la compréhension du contenu par l'extraction d'information pertinente remplissant les blancs dans un formulaire préétabli. Ainsi dans un texte d'une offre d'emploi, domaine d'intérêt de ces travaux, il s'agit de comprendre l'offre en identifiant, le nom du poste à pourvoir, le nom de l'entreprise en besoin de ressources humaines, la date de l'embauche prévue, le lieu de travail, les

compétences du candidats souhaitées et l'adresse à laquelle il doit postuler. Nous reviendrons ultérieurement en détail sur les informations sélectionnées pour être pertinentes à extraire pour décrire au mieux le contenu d'une offre. Il s'agit de comprendre la thématique du texte pour mieux le faire correspondre avec les requêtes utilisateurs.

L'extraction d'information peut être comparée à un filtre d'information pour un grand volume de textes. D'après la définition donnée dans [29], il s'agit de l'identification de classes d'évènements particulières ou de relations dans un document en langage naturel. Elle implique la création de représentations structurées de l'information pertinente capturant la sémantique dépeinte dans le texte. L'extraction d'information n'a pas la prétention de comprendre les textes dans leur globalité, elle se limite au contraire à la compréhension des passages de textes renfermant les informations nécessaires au remplissage des champs d'un formulaire pré-défini. Les utilisateurs de systèmes de recherche d'information ont des besoins informationnels qu'ils expriment très souvent en langage naturel ainsi si quelqu'un veut savoir les lieux des attentats survenus en Irak le 25 Janvier 2007 ou les entreprises ayant fusionné pendant le mois de février 2007, les systèmes à base de modules d'extraction d'information s'avèrent plus efficaces pour retourner des réponses précises et non pas un ensemble de documents pouvant contenir la réponse.

Cette idée de réduction de l'information d'un texte à une structure de table n'est pas nouvelle, sa faisabilité fût suggérée par Zellig Harris à partir des années 50 en définissant la notion de sous-langage. La notion spécifique d'EI a bénéficié d'une grande attention surtout avec le lancement de la conférence MUC[1] en 1987 qui se déroule sous forme de concours dans lequel différents systèmes d'extraction d'information sont évalués et comparés sur une même tâche pré-définie d'extraction d'entités nommées. Elle fût organisée sept fois entre 1987 et 1998.
La participation au concours se déroule alors en trois phases. Une fois la tâche à effectuer spécifiée, les participants reçoivent un programme d'évaluation automatique et un corpus pour lequel les informations à extraire sont spécifiées. Ces derniers ont alors entre 1 et 6 mois pour apprendre un nouveau système ou pour développer des patrons d'extraction dans le cas de méthodes à base de règles. La période d'apprentissage écoulée, les candidats reçoivent un nouveau corpus de test sur lequel ils appliquent leurs méthodes. Les résultats obtenus sont ensuite comparés par le comité avec les données manuellement extraites sur ce même corpus test et les systèmes sont comparés en fonction des valeurs de *Précision, Rappel* et *F-mesure* obtenues. Les informations à extraire

[1]Message Understanding Conference

19 March- A bomb went off this morning near a power tour in San Salvador leaving a large part of the city without energy, but no casualties have been reported. According to an unofficial source, the bomb - allegedly denoted by urban guerrilla commandos - blew up a power tour in the northwestern part of San Salvador at 0650 (1250 GMT).

Incident Type	bombing
Date	March 19
Location	El Salvador : San Salvador (city)
Perpetrator	urban guerrilla commandos
Physical Target	power tour
Human Target	-
Effect on Physical Target	destroyed
Effect on Human Target	no injury or death
Instrument	bomb

FIG. 4.1 – Échantillon d'un texte de la MUC-3

vont de la reconnaissance de noms propres, de dates, de mesures monétaires (Entités Nommées) à la reconnaissance de relations entre entités et reconnaissance de faits et évènements. La figure 4.1 [29] présente un exemple simplifié de la tâche d'extraction de la Conférence MUC-3. Il s'agit à partir de l'échantillon du texte présent en haut de la figure 4.1 de reconnaître automatiquement les entités nécessaires pour remplir le formulaire en bas de la même figure 4.1.

Chaque système d'extraction d'information est composé de deux phases d'analyses principales comme le montre la figure 4.2 qui en illustre l'architecture globale [29]. La première étape, d'analyse locale, permet d'identifier les faits simples à partir du document et la seconde, d'analyse du discours, permet d'inférer des faits plus complexes à partir de ceux reconnus à la phase précédente. On peut distinguer trois types de systèmes d'extraction d'information : Les systèmes à apprentissage, les systèmes à base de règle et les systèmes hybrides.

La succession de la MUC est assurée par la conférence CoNLL qui s'est néanmoins focalisée sur les systèmes d'extraction d'information à base statistique. L'enjeu est par exemple en 2002 et 2003 d'identifier les entités nommées dans des textes de langues différentes. La conférence met alors 4 corpus de 4 langues différentes à disposition des candi-

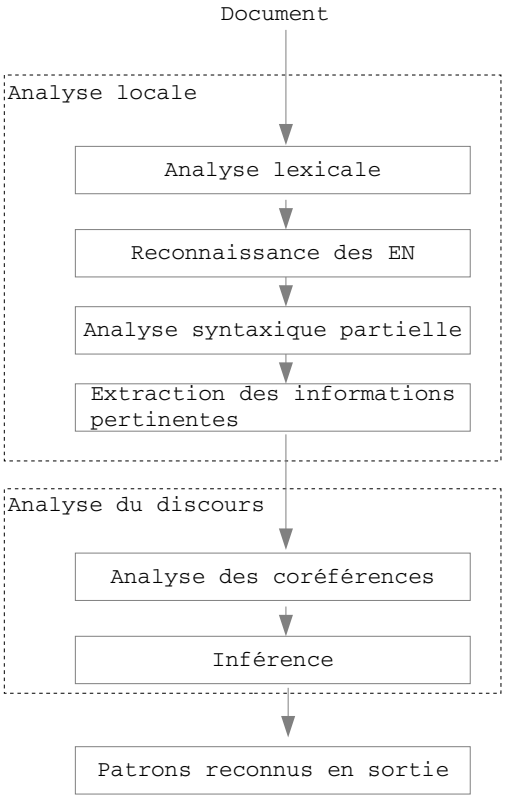


FIG. 4.2 – Architecture globale d'un système d'extraction d'information

dats qui doivent apprendre un système d'extraction d'entités nommées applicable aussi bien sur le hollandais que sur l'allemand, l'anglais et l'espagnol.

4.1 Systèmes d'extraction d'Entités Nommées

La reconnaissance et la classification automatique des Entités Nommées (EN) est une sous-tâche du domaine de l'*extraction d'information* qui à son tour est une sous-tâche de la *compréhension automatique de documents*. L'extraction des entités nommées se focalise sur la classification des termes d'un texte sous 7 catégories sémantiques : *Personne, Situation géographique, Organisation, Date, Mesure temporelle, Expression monétaire,* et la classe *Non-EN*. L'extraction d'information en général nécessite la reconnaissance d'informations plus détaillées que les simples entités nommées tel que la reconnaissance de relations, la reconnaissance d'acteurs et d'évènements dans un texte pour remplir des patrons évènementiels pré-définis. Ainsi dans une phrase comme *« Microsoft president Bill Gates... »* [7], on attend d'un système d'extraction qu'il rapporte qu'il s'agit d'une entreprise du nom de *Microsoft* et que la personne *Bill Gates* est un employé de l'entreprise *Microsoft*. Or avant de pouvoir retrouver les relations entre les noms propres *Bill Gates* et *Microsoft* il faut déjà être capable d'identifier ces termes comme étant des noms propres et surtout être capable de les classifier dans les catégories sémantiques adéquates, à savoir *Bill Gates* appartient à la classe *Personne* et *Microsoft* à la classe *Organisation*. Cette reconnaissance et classification des noms propres est connue pour être l'extraction d'entités nommées. Cette tâche est cependant moins triviale qu'elle ne le paraît, de simples dictionnaires contenant des noms de personnes, de lieux ou d'organisations ne suffisent pas pour atteindre des qualités comparables à des étiquetages humains. En effet, un grand nombre d'ambiguïtés sémantiques vient très vite dégrader les performances d'un tel système, les noms propres ont aussi en comparaison avec les noms communs des homonymes comme on peut le voir dans les exemples suivants.

(4.1) **Mr. Washington**, directeur général de **Washington Finance Inc**.

→ *est ce que* ***Washington*** *est un Lieu ?*
→ *est ce que* ***Washington*** *est un nom de Personne ?*
→ *est ce que* ***Washington*** *est un nom ou une partie de nom d'une Organisation ?*

(4.2) **Philip Morris** annonça mercredi 21.02.06...

→ *est ce que* ***Philip Morris*** *est une Entreprise ?*
→ *est ce que* ***Philip Morris*** *est une Personne ?*

(4.3) Dans la Une du **Paris Match**

→ *est ce que dans ce cas* ***Paris*** *est un Lieu ?*
→ *est ce que* ***Paris*** *est un nom ou partie d'un nom d'une Organisation ?*

(4.4) ... des ventes en France. **Comme Véolia Vinci** affirme que...

→ *est ce que* ***Comme Véolia Vinci*** *est un nom d'organisation ?*
→ *comment identifier que* ***Comme*** *n'appartient pas au nom de l'organisation ?*
→ *comment identifier que* ***Véolia*** *est une Entreprise et que* ***Vinci*** *en est une autre ?*

Comment est-ce qu'un système automatique peut-il reconnaître et classifier les noms propres en particulier et les entités nommées en général dans un texte écrit en langue naturelle ?

Les noms propres ont deux spécificités qui permettent aux systèmes automatiques de les repérer, la première est la typologie interne et la seconde est le contexte d'apparition externe. Ces deux notions de contexte interne et de contexte externe ont été introduites par [14] sous le nom de « preuve interne » et « preuve externe ». Il présentait alors un système d'extraction d'entités nommées basé sur ces deux notions. Un troisième outil utilisé par la plupart des systèmes d'extraction d'entités nommées est ce qu'on appelle les « Gazeteers » . Il s'agit d'un ensemble de listes homogènes renfermant des exemples de noms d'une catégorie à chercher. On aura alors une liste pour les noms des organisations, une pour les abréviations des noms des organisations, une autre pour les prénoms de personnes, etc. La taille des listes est variable en fonction des systèmes.

4.1.0.1 Les preuves internes

Les preuves internes sont des indices indiquant à coup sûr qu'il s'agit d'un nom propre et qui font partie intégrante de ce dernier. Elles peuvent être de types différents : mots, abréviations, suffixes, préfixes, mais aussi de type typographique. Les preuves internes se placent en général au début ou à la fin du nom propre, ce qui facilite la délimitation des bordures aux systèmes automatiques. Les noms des organisations par exemple sont souvent composés de l'abréviation de leurs statuts juridiques (SARL, SAS, GmbH, AG, Inc. ...), ou alors par un mot

comme Banque, Organisation, Société, Groupe... qui peut s'avérer être une preuve externe selon le cas, pareillement pour les lieux avec les mots tel que Area, Aire, mer...

(4.5) **Société** Générale
Banque Nationale de Paris
MAC API **Sarl**
Jobanova **GmbH**

(4.6) **Mer** du sud
Aire de Thionville

Les prénoms peuvent également être considérés comme des preuves internes pour le repérage d'une personne.

(4.7) **Pierre** Mendès France
Françoise Lemenier

La majuscule joue un rôle très important dans la plupart des systèmes d'extraction automatique surtout dans les langues comme le français ou l'anglais, où seuls les noms propres et les débuts de phrases sont écrits en majuscule. Cependant cette preuve interne montre ses limites dans les langues comme l'allemand où tous les substantifs s'écrivent en majuscule.

4.1.0.2 Les preuves externes

Les preuves externes sont les contextes droits et les contextes gauches dans lesquels apparaissent les entités nommées. Ils viennent introduire les noms propres ou les enrichir d'informations supplémentaires. Les preuves externes permettent surtout de lever les ambiguïtés de classification des entités nommées reconnues par leurs preuves internes.

Ainsi dans l'exemple 4.8 (a.), « Pierre Cardin » pris hors contexte peut être classifié aussi bien dans la catégorie *Organisation* que dans la catégorie *Personne*, alors que dans les deux exemples suivants, et de par la présence des termes significatifs à gauche, il n'y a plus lieu de douter. L'abréviation « Mr. » (exemple 4.8 (a.)) précédant un nom de personne potentiel certifie bien qu'il s'agit d'une Personne et de même pour le second cas où le terme « groupe » lève toute ambiguïté sur la classification de l'entité nommée « Pierre Cardin ».

(4.8) (a.) **Pierre Cardin**
(b.) Le défilé de mode proposé par **Mr. Pierre Cardin**...
(c.) Le **groupe Pierre Cardin** est assigné en justice...

Les contextes externes sont réunis à travers un apprentissage sur des corpus de grandes tailles, par *Bootstraping* [62] par exemple, une méthode que nous détaillons plus tard et que nous prenons d'ailleurs comme point d'attache dans le développement de nos grammaires locales d'extraction d'information dans le domaine de l'emploi. [61] propose une description détaillée des contextes externes aux noms des organisations dans les dépêches journalistiques de langue anglaise. Elle étudie un grand nombre de verbes mettant en relation deux ou plusieurs organisations (le rachat, la cession, la fusion) ou alors les relations entre les personnes et les organisations (professions dans une organisation). Elle étudie également les descriptions venant enrichir les noms des organisations (le secteur d'activité, le marché et la clientèle visés) toujours pour la langue anglaise.

(4.9) **Abit**, en difficulté financière **fusionne** avec **USI**...
Trois mois après **le rachat** de **CitéFibre** par le **groupe ILIAD**...
Ansell Healthcare est un **leader mondial dans le domaine** des produits de protection...

Certaines catégories peuvent être à la fois des preuves internes et externes comme dans l'exemple la « **Société** Générale », le terme « Société » appartient au nom de l'organisation bancaire, alors qu'il joue un rôle informatif externe dans le cas de « La société Microsoft » où le nom de l'organisation se limite à « Microsoft ».

L'hypothèse de base de tout système d'extraction d'EN est le fait que tout nom propre n'apparaît pas isolé dans un texte mais il est introduit dans une description de caractéristiques, de faits ou d'évènements. La présence de ces caractéristiques, de ces verbes de relations ou de ces évènements dans le contexte environnant les noms propres est alors utilisé comme déclencheur pour la reconnaissance et la classification de ces derniers dans les catégories sémantiques pré-identifiées.

Notre système use des contextes internes et externes enrichis pour extraire les informations intéressantes dans le sous-langage des offres d'emploi.

Les métriques d'évaluation

L'évaluation d'un système d'extraction d'information se fait par la comparaison des résultats retournés par le système avec ceux étiquetés manuellement dans un même corpus de test. Au vu de cette comparaison un SRI peut produire deux types d'erreurs [54]. Le système peut retourner des résultats n'appartenant pas à l'ensemble des réponses correctes, ce phénomène est connu pour être le « Bruit », ou il peut

retourner des réponses correctes mais en omettant certaines possibles, c'est le phénomène de « Silence ». Ces deux erreurs peuvent être évaluées par les deux mesures « Rappel » et « Précision » que nous avons détaillé au niveau de la section sur les méthodes d'évaluation des systèmes de recherche d'information. Dans le cas présent, le rappel est le rapport entre toutes les entités extraites correctement par rapport à toutes les entités existantes dans les textes et la précision représente le rapport entre toutes les entités extraites correctement par rapport à toutes les entités extraites.
Une troisième mesure basée sur les deux citées précédemment est souvent utilisée pour évaluer les systèmes d'extraction d'information. Il s'agit de la « F-mesure », elle combine les deux mesures de précision et de rappel pour ainsi diminuer les grands écarts entre ces deux valeurs.

Propriété 2 *La F-mesure*

$$F - mesure = \frac{2.Precision.Rappel}{Rappel + Precision}$$

L'intérêt de disposer d'un système performant de reconnaissance et de classification des entités nommées est un enjeu pour plusieurs applications plus générales. Des textes étiquetés par les entités nommées permettent par exemple d'améliorer les performances de systèmes de classification automatique. En effet des dépêches financières auraient tendance à citer beaucoup de noms d'organisations alors que des textes littéraires n'en citeraient pas. Un tel pré-traitement permet également de structurer des documents textuels bruts et améliorer par conséquent la qualité des systèmes de recherche d'information mais il peut aussi être utile dans un système de traduction automatique, où les entités nommées ne sont pas à traduire mot à mot mais sont retenues telles quelles comme unités de sens.
L'extraction des entités nommées s'est vu évoluée ces dernières années en particulier dans le domaine biologique et médical, où il s'agit de reconnaître automatiquement les noms de gènes, de maladies, de symptômes dans des rapports scientifiques rédigés en langue naturelle. Cela permet entre autre de construire des bases de données médicales très riches utiles pour détecter certaines pathologies ou symptômes particuliers à une tranche d'ages ou à l'un des sexes ou à détecter d'autres facteurs extraits à partir des textes et non visibles dans la masse de documents bruts. Dans ce cas on parle de pré-traitement utile pour des applications de *Datamining*.

Les systèmes d'extraction d'information et particulièrement les systèmes d'extraction d'entités nommées sont de trois types. Les systèmes

à base de patrons d'extraction, les systèmes basés sur des calculs statistiques et les systèmes hybrides combinant les deux techniques précédentes.

4.1.1 Les approches à apprentissage

Face au coût élevé des systèmes à base de règles et face à leur spécialisation à des domaines d'application particuliers, les approches à apprentissage se sont multipliées, ayant pour but commun de modérer le besoin d'intervention des spécialistes dans les différentes phases de développement et de maintenance du système, mais aussi de développer des systèmes plus flexibles et transposables sans trop de modifications sur de nouveaux domaines ou de nouvelles langues.

[50] affirme que les interactions complexes faisant des systèmes à base de règles des systèmes difficiles à développer et à maintenir, peuvent être apprises automatiquement à partir d'un corpus d'apprentissage. Chaque système apprenant est cependant fortement dépendant de la taille du corpus d'apprentissage ainsi que de l'homogénéité de son contenu. D'après Miller les performances des systèmes à apprentissage augmentent d'une manière logarithmique avec la taille du corpus d'étude.

Les systèmes à apprentissage nécessitent ainsi un corpus d'apprentissage homogène et pré-traité. La simple collecte d'un ensemble de documents électroniques appartenant à une catégorie sémantique proche ne suffit pas pour atteindre des performances d'extraction comparables avec des extractions humaines ou des extractions par des systèmes à base de règles. Le corpus doit être pré-traité tel que : segmentation en phrase, en unités lexicales, annotation morpho-syntaxique et sémantique. Même si le corpus d'apprentissage a subi un pré-traitement manuel irréprochable, un simple écart de domaine de quelques textes peut causer des dégradations importantes de performances.

Les modèles statistiques les plus utilisés en traitement automatique du langage naturel et particulièrement en extraction d'information sont les chaînes de Markov cachées, l'entropie maximale et les arbres de décision.

4.1.1.1 Les modèles de Markov cachés

Les chaînes de Markov cachées [2] ont été découvertes au début du 20ème siècle par Andrei A. Markov (1856-1922). Plusieurs champs d'applications ont très vite adopté ce modèle qui s'est avéré efficace pour

[2]Hidden Markov Model ou HMM

la classification de séquences en analyse du langage, en reconnaissance des formes et en décodage de la parole en particulier.

Une chaîne de Markov cachée est d'une façon simplifiée un processus stochastique ou aléatoire dont l'état à l'instant $t+1$ dépend uniquement de son état à l'instant t. Elle peut être représentée par un automate à états finis composé d'états, de transitions et d'un ensemble de distributions de probabilité de transition. A chaque transition est associé un symbole de l'alphabet et à chaque état est associée une distribution de probabilité sur l'ensemble des symboles finis de l'alphabet [55]. Formellement une chaîne de Markov cachée est définie par le quintuplé *(S,K,*Π*,A,B)* :

$S = S_1, ..., S_N$: ensemble de N États

$K = k_1, ..., k_M$: ensemble de M symboles d'émissions

$\Pi = \pi_i, i \in S$: ensemble des probabilités des états initiaux

$A = a_{ij}, ietj \in S \qquad \sum_{j=1}^{N} a_{ij} = 1$: probabilité des transitions partant d'un état i à un état j $S_i \rightarrow S_j$

$B = b_{ijk}, ietj \in Setk \in K \qquad \sum_{k=1}^{M} b_{ijk} = 1$: probabilité d'émettre le symbole k étant dans l'état S_i et allant vers l'état S_j.

En extraction d'entités nommées, les états de la chaîne cachée sont représentés par les étiquettes sémantiques et les sorties observables sont représentées par les termes du document. La reconnaissance des entités nommées est alors réduite à un problème de classification où chaque terme d'un document est soit une entité nommée, soit une partie d'une entité nommée, ou alors n'est pas une entité nommée. Dans la phase d'apprentissage, les chaînes de Markov cachées nécessitent un corpus annoté pour pouvoir calculer les probabilités des mots et des transitions entre les mots.

[5] - BBN IdentiFinder

[5] développent une version d'un modèle de Markov caché qui attribut à chaque mot du corpus l'une des classes sémantiques des entités nommées recherchées ou encore la classe « Not-a-Name » représentant tous les autres mots.

Les états de cette version sont organisés en régions, une région par classe. Étant donné une phrase, l'état initial et l'état final sont représentés par le début et la fin de la phrase et les états intermédiaires sont représentés par les 7 classes représentatives des entités nommées à chercher, plus la 8ème classe pour les mots qui n'appartiennent pas à une EN. Dans chaque région, la probabilité d'occurrence des mots apparaissant dans cette région est calculée par un modèle statistique de bigrams. Cette probabilité d'occurrence d'un mot est alors calculée uniquement en fonction du mot précédent. Plus formellement chaque mot représente un état de la chaîne dans le modèle de bigrams et à chaque transition à partir du mot courant au suivant est associée une probabilité. Le calcul de la probabilité d'une séquence de mots $w_1...w_n$ est égale au produit des probabilités des différents mots $\prod_{i=1}^{n} .p(w_i|w_{i-1})$

La reconnaissance et classification des noms se fait à travers l'association de la séquence (Nom-Classe) la plus probable à une séquence de Mots W. L'algorithme de Viterbi [70] permet de réduire ce calcul de recherche du plus probable à une fonction linéaire de recherche du $maxP(Nom - Classe|w_i)$. Un certain nombre de critères contextuels sont introduits, afin d'éviter le traitement de tous les bigrams du texte qui ne porteraient aucune information contextuelle intéressante pour les EN.

- ◇ Les preuves externes comme les titres pour les noms de personnes ou les statuts juridiques pour les noms des organisations.
- ◇ Les mots vides comme les prépositions et les pronoms sont utilisés comme marqueur des bordures des entités nommées.
- ◇ Les preuves internes comme :
  - une séquence de 2 chiffres est interprétée comme une abréviation d'une *Année*
  - une séquence de chiffres contenant un ou plusieurs tirets est interprétée comme une *Date*
  - une séquence entièrement en majuscule est interprétée comme une *Organisation*
  - environ 20 descriptions de ce type sont retenues pour paramétrer les segments.

Le corpus de test de la MUC-6 est utilisé pour permettre une comparaison directe avec les systèmes antérieurs et surtout avec le meilleur système à base de règles. Le Système BBN obtient alors une F-mesure de 94 % face à 96,4 % pour le meilleur système à base de règles et obtient une F-mesure de 93 % sur un corpus entièrement en majuscule contre seulement 89 % pour ce même rival.

[76] NER using an HMM-based chunk tagger

Ce système propose un modèle de Markov caché et un étiqueteur de segments pour la reconnaissance et la classification des entités nommées. La différence majeure avec les autres systèmes implémentant des chaînes de Markov cachées est la théorie de probabilité admise. Dans ce système les auteurs admettent la théorie de l'information mutuelle entre les termes des documents contre la probabilité conditionnelle admise par les autres systèmes. La théorie de l'information mutuelle leur permet de mesurer la dépendance statistique entre deux termes tout en assurant la prise en compte d'informations contextuelles supplémentaires pour déterminer l'étiquette du terme courant. L'étiquetage des segments se fait en trois phases :

1. Identification des bordures des couples (terme, catégorie). Chaque terme prendra l'une des valeurs de BC suivantes : $BC = 0, 1, 2, 3$, où 0 signifie que le terme est une entité nommée à part entière et 1,2,3 représentent respectivement les positions (initiale, milieu, finale) d'une entité nommée.
2. Identification de la catégorie de l'entité nommée : $EC \bar{\{}Person$, *Organisation, Location...}* pour dénoter la catégorie de l'entité nommée
3. Identification des caractéristiques du terme WF

Une table de contrainte sur les bordures est ajoutée pour éviter les séquences mal coordonnées. Elle résume les combinaisons d'étiquette BC possibles entre les termes qui se suivent. Ainsi si l'étiquette de t_i est la bordure 2 (signifiant que le terme est un terme au milieu d'une entité nommée), le terme t_{i+1} ne peut pas avoir l'étiquette 0 ou 1 qui donnerait une séquence non valide de limites mal reconnues.

Les auteurs introduisent plus de 4 types d'informations contextuelles supplémentaires pour améliorer les performances du système :

1. Preuves internes simples : majuscules, digitalisation...
2. Déclencheurs sémantiques : suffixes et préfixes classifiés (River$\rightarrow$ *SuffixLoc* ou \$ $\rightarrow$ *PrefixMoney*...)
3. Ensemble de dictionnaires de noms de personnes, de noms d'organisations...
4. Évidences externes : identification des termes entourant les EN déjà connues et ceux candidats pour lever l'ambiguïté de classification de certaines EN.

L'approche a été apprise et testée sur les données de la MUC-6 et de la MUC-7, elle obtient une F-mesure de 96,6 % sur les données de MUC6, et de 94,28 % sur les données de la MUC-7.

4.1.1.2 L'entropie maximale

Ces dernières années et surtout depuis que la puissance de calcul des ordinateurs ne pose plus de grands problèmes, les méthodes à base du modèle d'entropie maximale se sont multipliées car elles appartiennent aux modèles exponentiels. L'entropie maximale est utilisée pour inférer des propriétés dont nous ne savons rien. En effet étant donnée une collection de faits, elle permet de choisir un modèle compatible avec tous les faits de la collection qui soit le plus général possible.

[39] Si un mot x apparaît $f(x)$ fois dans un texte de longueur n, la probabilité de choisir le mot x par hasard dans le texte est $p(x) = \frac{f(x)}{n}$. La quantité d'information portée par le mot x est alors $-\log p(x)$. Ceci signifie que plus $p(x)$ est petit, plus la quantité d'information est grande et plus $p(x)$ est grand, plus la quantité d'information se rapproche de 0. C'est le cas des mots vides comme les prépositions ou les articles qui n'apportent pas beaucoup d'information de par leur fréquence élevée. Si l'on ne connaît pas la probabilité de distribution d'une variable, il suffit d'appliquer le principe de l'entropie maximale qui représente la distribution maximisant l'entropie tout en assurant la neutralité.

La distribution de probabilité conditionnelle dans le cadre de l'entropie maximale s'écrit :

$$p*(o|h) = \frac{1}{Z(h)} \cdot \prod_{j=1}^{k} .\alpha_j^{f_j(h,o)}$$

$f_j(h,o)$ est une caractéristique binaire et α_j est le poids de chaque attribut

Les séquences invalides comme $person\_init, Organisation\_Fin$ sont écartées en utilisant la probabilité de transition $P(c_i|c_{i-1}) = 1$ si la séquence est valide et 0 si elle ne l'est pas, d'où la formule pour déterminer la séquence valide la plus probable

$$P(c_1, ..., c_n|s, D) = \prod_{i=1}^{n} P(c_i|s, D) * P(c_i|c_{i-1})$$

[7] - MENE

Le système MENE développé en 1998 par A. Botwick à l'université de New York est un modèle d'extraction d'entités nommées purement statistique, il n'utilise aucun patron d'extraction généré manuellement.

Soit un corpus segmenté en unités lexicales et un ensemble de n étiquettes sémantiques ($n = 7$) représentant des catégories des entités

nommées à chercher. Le problème de reconnaissance des EN est réduit à un problème d'attribution de l'une des $4.n + 1$ étiquettes à chaque unité lexicale du corpus. Pour chaque étiquette $x \in n$, x peut être dans 4 états possibles ($x_{init}, x_{continue}, x_{fin}, x_{unique}$). Une unité lexicale peut également être annotée par l'étiquette « Autre » signifiant que le mot ne fait pas partie d'une EN. Ainsi chaque unité lexicale du texte est initialement annotée par l'une des 29 étiquettes $(7 * 4 + 1)$ retenues. Dans un deuxième temps, les unités lexicales sont enrichies par des caractéristiques permettant d'améliorer les performances du système :

- ◇ Caractéristiques binaires : sont les évidences binaires historiquement sauvegardées pour une unité lexicale donnée.
- ◇ Caractéristiques lexicales : pour créer un historique lexical, les unités $w_{-2}...w_{+2}$ sont comparées avec le vocabulaire. Par exemple si le terme est « to » et que l'étiquette suivante est « $location_init$ », alors la preuve externe « to » augmente la probabilité de classification dans la classe « location ». L'utilisation de plusieurs marqueurs de discrimination faible ont été considérés et ont permis d'augmenter les performances du système.
- ◇ Caractéristiques régionales : elles font des prédictions en fonction de la section de l'article traitée comme le titre, les préambules etc.
- ◇ Dictionnaires : listes de prénoms, de noms d'organisations avec et sans suffixes, listes d'abréviations à deux lettres de noms de pays...

Ce système atteint la 4ème place dans le challenge de MUC-7 avec une F-mesure de 92,2 % composé d'une précision de 96 % et d'un rappel de 89 %.

4.1.1.3 Les modèles à base d'arbres de décision

Les arbres de décision sont des solutions optimales pour représenter des règles compréhensibles et facilement interprétables. Ils permettent de représenter graphiquement une procédure de classification. Les arbres de décision classifient des instances en les triant sous forme d'arbre, les noeuds y représentent les attributs, les branches y portent les différentes valeurs et les feuilles y représentent les classes dans lesquelles les instances ont été classifiées. Une instance peut donc être retrouvée en parcourant un chemin de la racine vers une feuille de l'arbre. L'apprentissage d'un arbre de décision nécessite un ensemble d'exemples d'apprentissage représenté sous forme de vecteurs d'attributs-valeurs appartenant aux différentes classes à apprendre.

L'idée est de diviser l'ensemble des exemples au fur et à mesure en fonction de tests sur les attributs pour obtenir des sous-ensembles homogènes à une même classe. L'arbre est terminé lorsque tous les exemples sont classifiés. A chaque étape de l'algorithme un attribut est choisi comme noeud à partir duquel émane autant de branches que de valeurs possible prises par les exemples. Un arbre est d'autant plus compliqué que le nombre de valeurs prises pour un attribut est grand. Le choix des attributs les plus discriminants doit être calculé à chaque étape car moins les attributs les plus discriminants sont choisis en haut de l'arbre moins l'arbre est compact.

[53] VIE NERC

[53] proposent d'utiliser l'induction des arbres de décision comme solution à la reconnaissance et à la classification des entités nommées dans un domaine très spécialisé. Les grammaires d'extraction sont construites par application de l'algorithme C4.5 d'apprentissage d'arbres de décision. Ces règles apprises sont compréhensives par les humains non pas comme celles produites par les chaînes de Markov cachées qui ne sont pas directement interprétables.

L'approche se base sur l'hypothèse que les entités nommées sont des groupes nominaux (GN), c'est pourquoi un analyseur grammatical permet dans une phase de pré-traitement d'extraire tous les GN du document. L'arbre de décision peut alors se focaliser sur ces GN et les catégoriser dans les classes des ENs prédéfinies ou dans la classe "Non-EN". Les auteurs se sont concentrés sur la reconnaissance et classification des noms de personnes et des noms d'organisations qui posent en général le plus grand problème dans les systèmes classiques, du fait que les noms d'organisations contiennent souvent des noms de personnes ou autres termes d'étiquettes grammaticales différentes.

Le but des auteurs est d'apprendre un système qui minimise le besoin en intervention humaine dans son adaptation à d'autres domaines d'application, pour cela ils introduisent un certain nombre d'informations supplémentaires. Ces informations sont obtenues par le passage d'un analyseur grammaticale, puis par le passage d'un comparateur entre les termes du document et ceux des listes des noms et des marqueurs disponibles. L'algorithme C4.5 utilisé nécessite des données représentées sous forme de vecteur d'attributs-valeurs.
Dans la phase d'apprentissage, les groupes nominaux représentant les organisations et les personnes extraites du corpus de la MUC-6 ont été mis sous forme de vecteur, dans lequel les preuves internes et externes (les deux termes à gauche et à droite accompagnés de leurs descriptions grammaticales) ont été ajoutées. Un vecteur de 28 attributs a

été définit. Pour que l'algorithme apprenne également à différencier entre les EN et les Non-EN, des exemples négatifs ont été ajoutés à l'ensemble des instances d'apprentissage. Ces exemples négatifs sont les groupes nominaux du corpus qui ne sont pas des EN. L'algorithme doit apprendre un arbre de décision qui classifie les exemples dans les 3 classes : Organisation, Personne, Non-EN.

Les résultats obtenus ne sont pas compétitifs avec les meilleurs systèmes ayant participé à la MUC-6 et la MUC-7 avec un rappel et une précision pour les Organisations de 69 % et 83 % et un rappel et une précision pour les Personnes de 84 % et 92 %. Cependant les auteurs insistent sur la lisibilité des règles et la possibilité d'interprétation de ces dernières permettant à des humains d'y apporter facilement des modifications.

4.1.2 Les approches hybrides

Les systèmes hybrides, sont des systèmes qui utilisent des patrons d'extraction pour les cas simples à reconnaître et des mesures statistiques pour lever les ambiguïtés rencontrées. Le meilleur système d'extraction automatique d'entités nommées gagnant du challenge MUC-7 est un système hybride, il s'agit du *LTG system* développé dans le groupe HCRC Language Technology Group de l'université d'Edinburgh, par [49].

[49] -LTG System

Le LTG system est le système vainqueur du concours d'extraction d'entités nommées de la conférence MUC-7, il obtenait sur les textes en langue anglaise un rappel de 93,6% et une précision de 95% sur les entités ENAMEX (Organisation, Personne, Lieu). L'approche d'extraction et de classification des ENs se déroule en 5 étapes :

1ère Étape : *Pré-traitement*

a- Chaque document est transformé par le module *LT-XML* en une structure XML qui permet de traiter les différentes composantes séparément et différemment selon les besoins.

b- *Lttok* découpe les éléments de l'arbre XML en unités lexicales. Différentes grammaires sont utilisées en fonction du segment de l'arbre traité.

c- *Ltstop* permet l'identification des fins des phrases. Des grammaires de reconnaissance sont initialement appliquées, accompagnées d'un modèle d'entropie maximale en cas d'ambigus.

d- *Lt pos* est un étiqueteur appris à partir d'un modèle d'entropie maximale.

e- Collecte d'informations supplémentaires :

i.) Vérifier si les mots en majuscule existent aussi en minuscule dans le reste du document ou dans la liste des EN connus.

ii.) Informations sémantiques :

- suffixes « yst » ou « ist » représentent des noms d'emplois.
- les mots dont les lemmes appartiennent à des noms de lieux et se terminant par « an » ou « ese » représentent des adjectifs locatifs.
- ...

2ème Étape : *Application des règles sûres à base de contextes internes sûrs*

- *Xxx + is a? JJ* Profession* → Nom de personne (*Yuri Gromov* is a former director)
- *Xxx +, ? whose Relation* → Nom de personne (*Nunberg*, whose stepfather)
- *Xxx+ area*s → Nom de lieu (*Beribidjan* area)
- *Profession of/at/with Xxx+* → Organisation (director of *Trinity Motors*)

3ème Étape : *Reconnaissance* partielle : tous les ENs reconnus à la phase précédente sont extraits dans le reste du corpus et chaque sous-élément est annoté par la catégorie correspondante à l'EN complet.

- « Lockheed Martin Production » est reconnu comme Organisation à la phase précédente

  « Lockheed », « Lockheed Martin », « Lockheed Production », « Martin », « Martin Production », « Production » seront annotées par la classe Organisation.

4ème Étape : *Application de règles relaxées*- ce sont des règles plus souples en terme de contraintes contextuelles. Si un prénom connu du dictionnaire est suivi d'un ou plusieurs mots en majuscule, le segment sera étiqueté par la classe Personne. L'ambiguïté de début de phrase est également résolue à ce niveau. Une règle du type : si une EN est candidate dont le premier mot apparaît en début de phrase et si ce mot survient ailleurs dans le document en minuscule, il sera supprimé de l'entité nommée.

5ème Étape : Nouvelle reconnaissance partielle : permet de décider avec un modèle d'entropie maximale des annotations ambiguës. Comme pour décider du cas de « Martin » qui a été annoté par la classe Organisation à la phase-3- mais qui appartient à d'autres classes dans des contextes différents.

6ème Étape : *Traitement du titre du document*- LTG traite le titre à part, car en anglais tous les mots du titre sont en majuscule et de plus les entités nommées n'y sont pas accompagnées de contextes externes identifiables et exploitables.

4.1.3 Les approches à base de règles

Les approches à base de règles tentent de décrire des comportements réguliers retrouvés dans les textes écrits en langage naturel. Elles partent du principe que les entités nommées sont entourées de preuves contextuelles répétitives pouvant être capturées et décrites dans une représentation formelle compréhensible par les machines.

Les approches à base d'apprentissage sont fortement dépendantes de la taille des données d'apprentissage et surtout de leur homogénéïté. Les données d'apprentissage doivent être étiquetées et correctement pré-traitées, ce qui est très coûteux en terme d'effort initial. Elles prétendent être des méthodes indépendantes des domaines, et plus facilement transposables sur des nouveaux thématiques ou de nouvelles langues, bien que le moindre écart sémantique entre les données d'apprentissage et les données de test provoque une régression rapide de leurs performances.

Les approches à base de règles en revanche n'ont pas l'ambition d'être indépendantes du domaine, elles visent à développer des systèmes très performants pour des domaines spécifiques intéressants.

[14] Mc Donald

[14] propose un système d'extraction et de classification sémantique de noms propres basé sur les deux notions de *preuves internes* et de *preuves externes* qu'il introduisait alors. Le système est composé de trois procédures :

1. La délimitation qui revient à la détection des bordures de début et de fin des noms propres ;
2. La classification des séquences trouvées par les classes des différents mots composants ;
3. Le recensement des noms et de leurs classes dans le modèle de discours.

L'algorithme de délimitation extrait initialement toutes les suites de termes en majuscule pouvant contenir certains caractères spéciaux comme le *&* (et commercial). La classe correspondante à chacune de ces séquences est alors cherchée en deux étapes. Une analyse grammaticale régulière est initialement appliquée sur les séquences, elle permet d'introduire des informations grammaticales et sémantiques aux mots de la séquence. C'est à ce niveau que les preuves internes sur lesquelles se base la première classification sont reconnues. Des heuristiques de priorités permettent de plus de lever l'ambiguïté des séquences se laissant classifier dans plusieurs classes, par exemple *Mr.* ou *Dr.* sont des preuves à forte priorité alors qu'un prénom connu du dictionnaire est plus ambiguë et nécessite des patrons d'extraction supplémentaires. Dans la seconde phase de classification, les preuves externes sont appliquées aux séquences ambiguës et à celles pas du tout classifiées. Dans le cas où aucune classification n'a été possible la séquence en majuscule sera annotée par la catégorie « Nom ».

Mc Donalds rapporte une performance se rapprochant des 100 % sur un corpus de « Who's News » pour lequel une grammaire complète fût élaborée.

[6] - FACILE

Le système FACILE est un système à base de règles développé en 1998 par [6] dans le département de l'ingénierie du langage UMIST à Manchester.

L'architecture générale du système se résume ainsi :

- Pré-traitement
- Normalisation
- Reconnaissance de formats spéciaux
- Segmentation
- Étiquetage grammatical
- Étiquetage sémantique des mots composés et mots simples connu des dictionnaires
- Reconnaissance et classification des entités nommées

Formellement chaque unité lexicale est représentée par un vecteur dans lequel un certain nombre d'informations sont stockées :

- si la première lettre du mot est en majuscule
- si le mot est entièrement en majuscule
- si le mot contient des lettres et des chiffres
- quels caractères spéciaux le précèdent et le suivent ?
- dans quelle zone le mot apparaît-il (titre, corps) ?
- la forme normalisée du mot
- l'étiquette grammaticale du mot

– l'analyse morphologique du mot
– la présence de certains suffixes

La reconnaissance des EN est basée sur des patrons d'extraction du type $A \Rightarrow B\ C|D$, où A est un ensemble d'attributs, d'opérateurs et de valeurs ; B et D sont des séquences représentant les contextes gauches et droits de l'unité lexicale C. De plus, des poids de certitude entre -1 et $+1$ viennent multiplier les parties gauches des règles. Concrètement la règle :

$$
\begin{array}{l}
[syn = NP, sem = ORG](0.9) => \\
\quad /[norm = "university"], \\
\quad [token = "of"], \\
\quad [sem = REGION|COUNTRY|CITY]/;
\end{array}
$$

signifie que si une séquence comme celle à gauche de la règle (University of Mannheim) est découverte dans le texte, elle sera annotée par l'étiquette syntaxique NP et par l'étiquette sémantique ORG. Les règles introduisent également des variables permettant de gérer des coréférences de noms dans différentes règles.
Ce système atteint un rappel de 92 % et une précision de 93 % sur le corpus de la MUC-7.

4.1.3.1 Les méthodes à base de grammaires locales

Les grammaires locales peuvent être vues pour l'instant comme un moyen différent de représentation, de visualisation et d'application de règles contextuelles d'extraction. Il n'existe malheureusement pas d'application des grammaires locales sur les corpus proposés par la MUC-6 ou -7, ce qui rend les comparaisons un peu plus difficiles, mais certains résultats obtenus sur d'autres corpus plus variés ont réussi à nous convaincre de l'efficacité des grammaires locales dans la tâche d'extraction d'information et des entités nommées en particulier. Nous avons par conséquent opté dans nos travaux pour une solution à base de grammaires locales.

[62]- Senellart

[62] décrit une méthode de Bootstarpping pour la localisation de groupes nominaux (GN) qui dénotent d'un lieu ou d'un nom de personne. Il justifie la nécessité d'élaborer des grammaires locales pour ce genre de GN par le fait que chaque phrase dans les dépêches journalistiques en langue française contiennent ce type de phrases et ce

serait une grande perte d'essayer de faire une analyse grammaticale, syntaxique d'un tel texte sans avoir primordialement reconnu ce genre de séquences répétitives semi-figées.

Le but de l'auteur est de présenter un moyen de construire, de maintenir à jour et de gérer une grande base de données de transducteurs à états finis. Il utilise un corpus de plus de 10 millions d'unités lexicales, représentant 1 an des dépêches du *International Herald Tribune*.

L'auteur propose de décrire la classe sémantique *« Officer »* . Pour cela il commence sa recherche par le mot *officer* dans le corpus. Il trouve 85 concordances qu'il trie selon les contextes gauches. Il extrait ensuite tous les noms et adjectifs qui explicitent le métier dans l'armée, la marine, la police... Il regroupe ensuite les indices collectés dans 4 classes sémantiques distinctes, puis les séquences sont de nouveau cherchées et annotées par la classe sémantique correspondante et ainsi de suite, en ajoutant au fur et à mesure des éléments du contexte droit et des éléments du contexte gauche ainsi que des synonymes, des hyponymes et d'autres noms sémantiquement liés pour élargir le champ d'action futur de la grammaire. Ce processus est très général mais il représente un moyen très efficace pour explorer des textes et construire des grammaires locales.

[61]- Schmidt

[61] propose un système d'extraction de noms d'organisations dans des dépêches journalistiques en langue anglaise à base de grammaires locales. Elle se spécialise sur l'extraction d'un seul type d'entités nommées et essaye dans sa thèse de décrire tous les contextes dans lesquels un nom d'organisation peut apparaître. Elle explore une multitude de scénarios à travers des verbes supports, qui permettent de reconnaître à coup sûr les noms des organisations dans des textes en langue anglaise. Elle considère également les différentes relations entre une organisation et une autre, entre une organisation et ses employés, entre une organisation et ses clients, pour proposer une extraction très efficace des noms d'organisations dans le contexte économique.

Elle testait alors son système sur 7 corpus collectés à partir de la version en ligne de 7 quotidiens différents et obtenait sur le corpus du *Financial Time*, par exemple, une précision de 94,47 % et rappel de 91,18 %, soit une F-mesure de 92,80 %. Un prototype peut être testé à l'adresse suivante :`http://www.cis.uni-muenchen.de/~schmidt`.

4.2 Les Grammaires Locales

Les grammaires locales sont des grammaires qui décrivent les contraintes locales associées à certaines unités de sens. Elles décrivent des phénomènes figés ou semi-figés. Une grammaire locale est avant tout une grammaire, qui d'après le trésor de la langue française informatisé est un « Ensemble de règles conventionnelles (variables suivant les époques) qui déterminent un emploi correct (ou bon usage) de la langue parlée et de la langue écrite » . C'est également selon la même source une « Étude objective et systématique des éléments (phonèmes, morphèmes,mots) et des procédés (de formation, de construction, d'expression) qui constituent et caractérisent le système d'une langue naturelle ». Elle est locale car elle explore les spécificités contextuelles locales entourant un phénomène linguistique particulier. Elle n'a pas la prétention de décrire la totalité des phrases grammaticalement correctes d'une langue mais se contente de décrire de manière détaillée et exhaustive le comportement locale du phénomène étudié. [56] explique dans [56] qu'il est possible de considérer les grammaires locales comme des petits sous-langages [56] et qu'un sous-langage serait par ailleurs un ensemble de grammaires locales étendues.

Les systèmes d'extraction d'information appris sur un corpus quelque soit sa taille, tentent d'être le plus général possible et en perdent par conséquent en précision. Pour un domaine donné, un sous-langage donné, il est possible d'élaborer un ensemble de grammaires locales étendues couvrant au mieux la totalité du sous-langage. C'est sur cette supposition que se base notre approche de construction automatique de base de données d'offres d'emploi que nous détaillons dans ce manuscrit.

Les grammaires locales que nous considérons sont sous la forme de réseaux récursifs de transition [64, 71], dont la manipulation est facilitée par les logiciels libres *Intex*[3] et *Unitex*[4]. Ces logiciels mettent à disposition des utilisateurs des interfaces graphiques permettant la construction, la compilation et l'application de graphes lexicalisés[64, 35] directement sur les textes de façon interactive et intuitive.

Chaque graphe a un état initial et un état final, un chemin entre l'état initial et l'état final représente une séquence de termes dans le texte. Ces graphes connus pour être des réseaux récursifs de transition permettent l'appel à des sous-graphes mais interdisent la récursivité et les chemins vides. Chaque noeud peut contenir soit un état terminal, soit un sous-graphe.

[3] http ://www.nyu.edu/pages/linguistics/intex/
[4] http ://www-igm.univ-mlv.fr/ unitex/

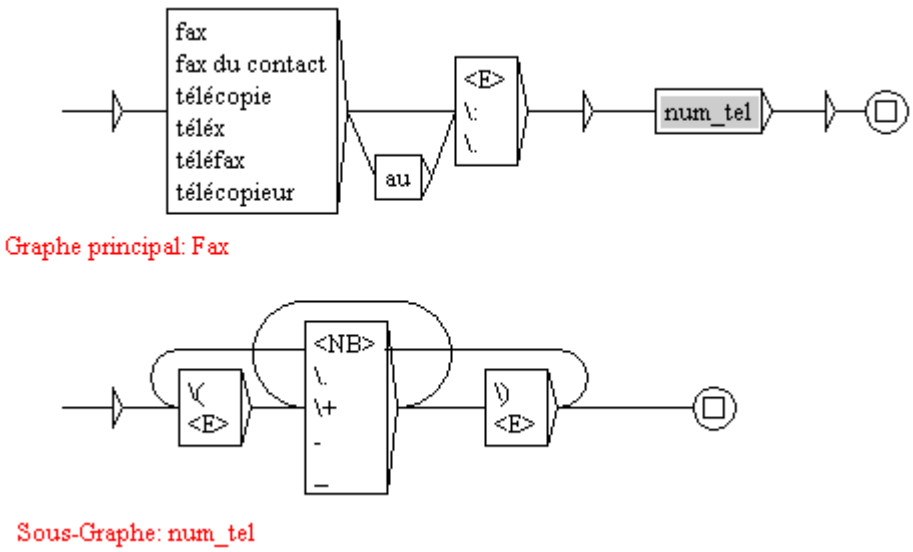


FIG. 4.3 – Grammaire locale pour la reconnaissance du numéro de fax

Formellement, une grammaire a un alphabet N de symboles non terminaux, un alphabet T de symboles terminaux (les symboles de T et N sont distincts), un ensemble de règles G représenté sous Unitex ou Intex sous forme de graphes et un axiome de départ g_0.

Dans les grammaires locales les symboles non terminaux sont les noms des sous-graphes et les symboles terminaux sont les unités lexicales, les symboles spéciaux et les signes de ponctuation. Dans l'exemple 4.3 on cherche à reconnaître les numéros de fax dans le texte à analyser, cette première grammaire fait appel au sous-graphe *num_ tel* en bas sur la même figure.

Des groupes de classes ont été définis pour diminuer le nombre de transition dans les graphes et pour permettre d'englober les symboles terminaux homogènes. ainsi :

<*voiture*> :	représente toutes les formes fléchies de la forme canonique voiture
<*N*> :	regroupe tous les mots du dictionnaire codé par l'étiquette syntaxique Nom
<*V :3s*> :	désigne toutes les formes verbales à la 3ème personne du singulier
<*NB*> :	représente n'importe quel suite contiguë de chiffres
<*!DIC*> :	représente toutes les unités lexicales inconnues des dictionnaires appliqués au texte

Quand une grammaire d'un champ sémantique quelconque est construite, il est très important de séparer les graphes dès qu'une variation sémantique apparaît afin d'assurer une construction modulaire où chaque graphe élémentaire peut être réutilisé dans une autre application distincte. Si un graphe contient deux ou plusieurs notions sémantiques différentes il est nécessaire de le décomposer en plusieurs automates représentant chacun une classe sémantique distincte. Cette séparation permet à côté de la réutilisation, une exécution plus rapide car plus lisible et moins encombrée. L'avantage majeur de cette notation est la clarté et la maintenance facile. Si une phrase du texte n'a pas été reconnue, il est alors très facile de se rendre au graphe correspondant et de le modifier. Les suppressions, mises à jour et changements sont rapides et intuitifs du fait que chaque graphe porte un nom sémantiquement significatif. Il est important pendant la phase de génération de décrire chaque occurrence une par une dans le contexte du champ considéré [27] afin de formaliser correctement les fonctions syntaxticomorphologiques du langage. Cette tâche est certes fastidieuse, mais elle a le mérite de produire des grammaires fiables et complètes, nécessaires dans le contexte actuel et prochaines où les méthodes d'apprentissage voient très vite leur limites face à une masse d'information croissante et à des utilisateurs toujours plus exigeants.

Il est également possible d'ajouter des étiquettes sémantiques à chaque séquence reconnue dans le texte, les graphes sont alors dit transducteurs à états finis [65].

Si on applique le transducteur à état fini de la figure 4.2 en mode fusion sur le texte de la figure 4.2, on obtient les concordances de la figure 4.6. On remarque qu'il est possible d'annoter le texte par des étiquettes sémantiques précises permettant des traitements ultérieurs sur le texte. Les chemins différents d'un même graphe peuvent ainsi donner des sorties étiquetées différentes en fonction du besoin de l'analyse.

La cascade de transducteurs est l'application de plusieurs transducteurs sur un texte selon un ordre particulier. L'ordre est celui de la fiabilité des règles : plus une règle est sûre plus elle est appliquée tôt

Mme Marie de la Marjolaine s'est prononcée sur la question de la restructuration sociale alors que le Directeur du groupe Mr. Gérald ne fût pas de commentaires devant les caméras.

FIG. 4.4 – échantillon de texte

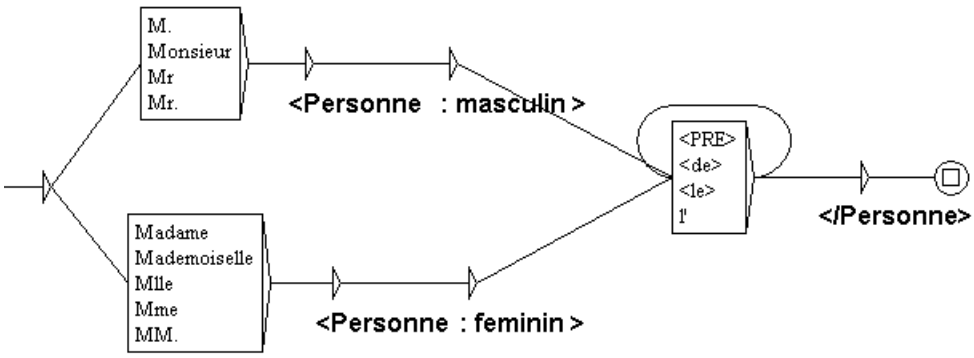


FIG. 4.5 – Grammaire locale autour de *Mr.* et *Mme*

Mme <**Personne : feminin**> Marie de la Marjolaine </**Personne**> s'est prononcée sur la question de la restructuration sociale alors que le Directeur du groupe Mr. <**Personne : masculin**> Gérald </**Personne**> ne fût pas de commentaires devant les caméras.

FIG. 4.6 – Echantillon étiqueté par le transducteur

dans la hiérarchie, à chaque étape le texte est transformé par les étiquettes de sorties du transducteur. Le texte ainsi transformé représente les données en entrée de la phase suivante, ce qui signifie que chaque transducteur est dépendant des résultats de l'étape précédente. Cette méthode s'est avérée améliorer les résultats dans une tâche d'extraction d'entités nommées [25], pour laquelle le texte s'enrichit au fur et à mesure d'annotations sémantiques permettant de diminuer l'ambiguïté et d'améliorer les reconnaissances dans les itérations avancées.

La description lexicale d'un texte ne permet pas d'atteindre des performances de compréhension automatique de textes proches de celle des humains, seule une analyse sémantique et syntaxique peuvent prétendre arriver à une telle solution.

4.2.1 Unitex

Unitex est une plateforme logicielle libre permettant l'analyse linguistique de textes en langue naturelle. Elle introduit des ressources linguistiques de type dictionnaires électroniques, grammaires et lexiques grammaires. Unitex fût développé au Laboratoire d'Automatique Documentaire et Linguistique (LADL) à Marne la Vallée en Ile de France. Les ressources linguistiques développées initialement pour la langue française au sein du LADL sous la direction du Professeur Maurice Gross (1938-2001), se sont vues très vite se multiplier pour les langues comme l'allemand, le russe ou le coréen à travers les partenaires du réseau RELEX des laboratoires de linguistique et d'informatique linguistique[5].

Le logiciel étant distribué sous licence GPL, chaque utilisateur peut l'adapter à ses besoins et développer des unités compatibles pour l'enrichir. Il permet aux linguistes de travailler aussi bien sur un niveau syntaxique que lexicale et morphologique dans leur analyse automatique de textes grâce à l'introduction de ressources électroniques et de grammaires flexionnelles et locales applicables aussi bien sur le caractère que sur l'unité lexicale.

L'outil permet grâce à une interface graphique conviviale de construire, modifier et appliquer des grammaires de reconnaissance sur un texte libre, de rechercher des patrons par le moyen d'un éditeur d'expressions rationnelles et d'appliquer des dictionnaires aux textes et reconnaître ainsi les mots inconnus. Un module d'élagage est également disponible permettant de lever l'ambiguïté dans les automates

[5]Laporte « Le réseau RELEX regroupe une douzaine de laboratoires, situés pour la plupart en Europe, qui collaborent à la constitution d'un inventaire d'informations linguistiques précises et exploitables dans les traitements automatiques, sur la base d'exigences méthodologiques : reproductibilité, exhaustivité, cumulativité »

du texte par des règles d'élagage modifiables facilement si besoin est.

L'interface d'Unitex est développée en Java et tous les autres programmes et modules de traitement en C++. Différentes distributions sont disponibles aussi bien pour les machines sous Linux, Mac Os que pour Windows 9x, NT, 2000, XP et ME. Les dictionnaires développés par les différents partenaires du réseau RELEX crée par Maurice Gross et son équipe sont conformes au formalisme DELA [12]. Les deux formats des dictionnaires DELAS et DELAC pour les mots simples et composés sont détaillés dans le chapitre suivant. Chaque entrée du dictionnaire est représentée sur une ligne où le mot sous sa forme fléchie est suivi de sa forme canonique, de sa ou ses catégories grammaticales, sémantiques et morphologiques. La dernière version d'Unitex (Unitex 1.2, dernière mise à jour 24 Juillet 2006) contient des ressources pour l'anglais, le finnois, le français (de France et de Québec), l'allemand, le grec (moderne et ancien), l'italien, le coréen, le norvégien, le polonais, le portugais (du brésil et du Portugal), le russe, le serbe, l'espagnol et le thaïlandais.

Nous résumons dans la figure 4.2.1 le processus suivi par Unitex lors de l'application d'une grammaire d'extraction sur un texte en langue naturelle.

Dans cette figure on distingue deux phases de traitement, la première consiste en une phase de pré-traitement du texte en entrée, nous insistons sur le fait qu'il est possible d'analyser des corpus textuels de plusieurs Giga-Octet. Dans cette phase, le texte sera initialement converti en Unicode, puis les espaces, les caractères spéciaux et les formes ambiguës y seront normalisés, ensuite les bordures des phrases seront reconnues par les transducteurs adéquats et le texte sera décomposé en phrases puis en unités lexicales et ce par le moyen des signes de ponctuation et des espaces. La dernière étape de la phase de pré-traitement est l'application des dictionnaires aux unités lexicales trouvées. A la fin de cette première phase on dispose pour un document donné, du nombre de phrases qu'il contient, ainsi que de toutes les unités lexicales simples et composées connues des dictionnaires, annotées par leurs informations grammaticales, sémantiques et morphologiques ; on dispose de plus de la liste des unités lexicales inconnues des dictionnaires. La seconde phase d'extraction d'information peut alors être appliquée, il suffit alors d'utiliser le module *« Locate »* auquel on passe la grammaire de reconnaissance en paramètre. La liste des concordances trouvées peut alors simplement être affichée dans un éditeur ou alors venir étiqueter les passages concernés dans texte.

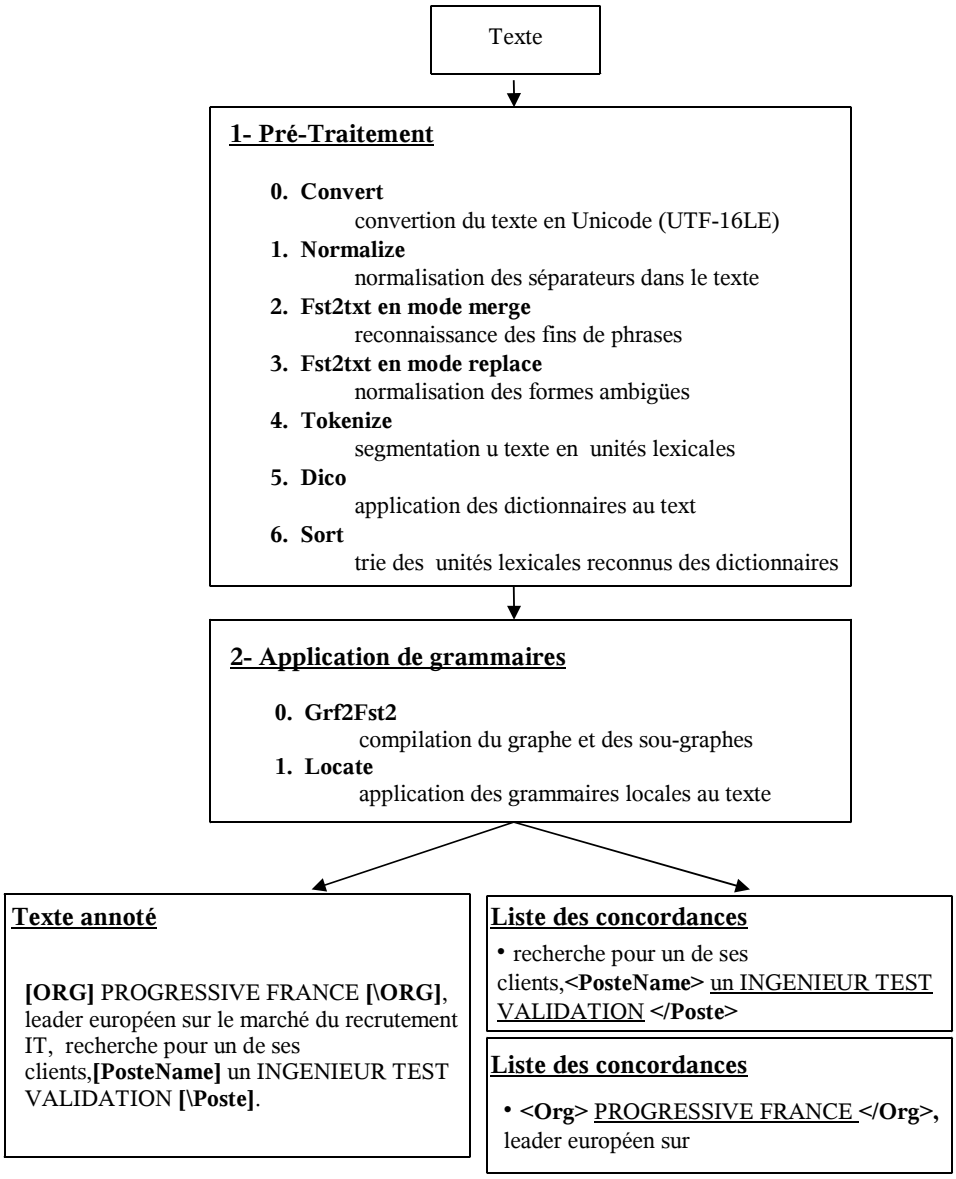


FIG. 4.7 – Processus d'extraction d'information par Unitex

4.3 Conclusion

La construction de grammaires avait comme objectif dans le passé la description générale d'un langage, le but était de couvrir toutes les phrases correctes d'un langage à un niveau plus formel.

Dans les modèles initiaux, les phrases étaient transformées tel que chaque mot était remplacé par sa catégorie grammaticale. En 1957 N. Chomsky montrait les limites de ces grammaires dites hors contextes pour le traitement des langages naturels. Il proposait alors avec Z. Harris un système combiné où prônait la notion de « phrase élémentaire » et où toute phrase complexe était une combinaison de plusieurs phrases élémentaires. Ils introduisaient deux types de transformations de la phrase élémentaire : La *règle unaire* transformant une phrase élémentaire en une autre phrase élémentaire et la *règle binaire* transformant une paire de phrases élémentaires en une seule phrase plus complexe. Or il s'est avéré impossible de construire une telle grammaire générale de par la présence dans les langues naturelles d'expressions idiomatiques et d'une multitude d'exceptions même face à des règles grammaticales claires. Ces idiosyncrasies ont des comportements particuliers dans des situations où une règle générale ne peut être véhiculée sans causer de troubles de logiques dans le système. Les grammaires locales sont alors la solution à la génération de grammaires descriptives de phénomènes locaux et précis. Etant donné que les grammaires locales sont un ensemble de graphes décrivant chacun un phénomène sémantique local et réduit, il suffit d'ajouter une transition supplémentaire pour chaque cas particulier dans le graphe de la classe correspondante pour couvrir les cas d'exception.

Depuis que les grammaires locales ont été implémentées sous forme d'automates à états finis, elles sont beaucoup plus performantes car leur sorties dépend linéairement de la taille de l'entrée. Les algorithmes de minimisation classiques assurent un gain d'espace mémoire énorme de ce type de systèmes.

Nous avons opté pour un système à base de grammaires locales pour le remplissage automatique de notre base de données d'offres d'emploi, car elles nous assurent une précision, une clarté et une maintenance facile et efficace dans un contexte où les règles grammaticales sont souvent ignorées.

CHAPITRE 5

Les Mots Composés en français

> *« La connaissance des mots conduit à la connaissance des choses »*
> Platon

Introduction

Au niveau de ce chapitre, nous résumerons les travaux antérieurs dans l'étude des mots composés français. Nous aborderons les unités polylexicales de différents angles de traitement. En effet, nous nous intéresserons d'une part à leur recensement et à leur reconnaissance ([60, 31, 32, 34]), d'autre part, à leur analyse syntaxique et typologique ([44, 30]), mais aussi à leur analyse morphologique et flexion automatique. Nous décrirons pour finir le format de présentation du dictionnaire électronique DELAC et de sa forme fléchie DELACF que nous prendrons ultérieurement comme représentation pour notre dictionnaire des noms de profession composés.

5.1 Mots composés vs. séquences figées

La notion de « mot composé » n'a pas de dénomination stable. Elle a fait l'objet de nombreuses discussions et représente encore un sujet de polémique. Le terme *mots composés* est utilisé dans la littérature, pour désigner des unités lexicales complexes analysées sous différents points de vue. Pour Silberztein [64] par exemple, toute séquence composée contient au moins un séparateur alors que pour [45] une séquence soudée de plusieurs concepts simples est un composé alors qu'une séquence comportant des blancs est une *« locution »*.

Benveniste dans [4] définit la composition tel que :

Définition 1 *il y a composition quand deux termes identifiables pour le locuteur se conjoint en une unité nouvelle à signifié unique et constant.*

Même si la définition de la composition n'est pas unanime, il existe cependant un certain nombre de critères sur lesquels la plupart des linguistes tombent d'accord. Il s'agit :

1. du figement
2. de l'opacité sémantique
3. de la grammaticalisation

5.1.0.1 Le figement

Le figement occupe une place importante dans les préoccupations actuelles des linguistes. L'équipe de saint-cloud, montre dans ses travaux sur les corpus textuels que les séquences figées couvrent jusqu'à 20 % des textes analysés. Le figement est une notion qui pose problème car le terme *figé* fait penser à un figement morpho-syntaxique, ce qui signifierait que les variations morpho-syntaxiques (déclinaison, conjugaison, transformation) seraient restreintes, alors que la définition est beaucoup plus complexe. Le figement peut être observé sur le plan phonétique [1], morphologique [46], syntaxique (degré de figement) [34] et aussi sémantique [47, 34].

D'après [34], une expression figée dépendrait soit d'un figement syntaxique, soit d'un figement sémantique. Il fait la distinction entre ces deux types de figement (déf. 2).

Définition 2 *Une séquence est figée du point de vue syntaxique quand elle refuse toutes les possibilités combinatoires et transformationnelles[...], Elle est figée sémantiquement quand le sens est opaque ou non compositionnel.*

La notion de figement est définit par J. Dubois dans [15] comme étant :

Définition 3 *le processus par lequel un groupe de mots dont les éléments sont libres devient une expression dont les éléments sont indissociables.*

C'est donc une lexicalisation dont la caractéristique principale est : « la perte du sens propre des éléments constituant le groupe de mots,

qui apparaît alors comme une nouvelle unité lexicale, autonome et à sens complet, indépendant de ses composantes »[1].

G. Gross [34] explique que les locutions « ne sont pas toutes figées au même degré ». L'indication du degré de figement, ajoute-t-il, « se reflète dans les possibilités transformationnelles ».

La détermination du degré de figement syntaxique d'une séquence consiste à dégager les restrictions plus ou moins nombreuses qu'elle connait par rapport aux transformations dont elle serait théoriquement passible en fonction de sa catégorie et de son patron syntaxique.

5.1.0.2 L'opacité sémantique

Le sens d'une séquence dans le cas traditionnel est le produit de celui de ses éléments composants. Ainsi le sens de la phrase « j'ai acheté un livre », est celui de la combinaison des sens respectifs du pronom personnel « je », du verbe « acheter » et du nom masculin « livre ». Cette phrase est ainsi sujette à une lecture compositionnelle. Le sens d'une expression figée, quand à lui, ne relève pas de ce type de lecture. En effet, le sens de la séquence « Cordon bleu », qui signifie un cuisinier exceptionnel, n'est pas le produit du sens combiné de ses lexèmes respectifs, elle se heurte donc à une opacité sémantique. Une expression figée possède un sens propre non déductible de celui de ses composants.

5.1.0.3 La grammaticalisation

Selon la définition extraite du dictionnaire français en ligne *LEXILOLOG*, la *Grammaticalisation* est « le processus selon lequel un terme ou un syntagme acquiert un statut grammatical, entre dans un système d'oppositions grammaticales »[...] et qui devient un « énoncé formé d'après les règles de production d'une grammaire donnée. ».

Pour la suite de ce mémoire, nous considérons tout syntagme à graphies multiples ayant un minimum de degré de figement, une représentation particulière dans le mémorial du lecteur et qui régit des règles de grammaires.

5.2 Analyse syntaxique des unités polylexicales

Les unités polylexicales ont fait l'objet de maintes études dans la littérature. La plupart des auteurs se sont concentrés sur une forme précise de composés comme [43] qui a initialement étudié les noms

[1] J. Dubois

composés à trait d'union ou [51] qui étudiait les composés de type « Nom Adjectif ».

La description des noms composés s'est très longtemps limitée aux six classes de composition suivantes :

- *Adjectif Nom* (belle-mère)
- *Nom Adjectif* (carte postale)
- *Nom Nom* (bateau-mouche)
- *Nom à Nom* (tir à l'arc)
- *Nom de Nom* (chemin de fer)
- *Nom Prep Nom* (saut en hauteur)
- *Verbe Nom* (ouvre-bouteille)

Cependant certaines analyses plus précises élaborées par [44] ont mis en évidence plus de 700 classes de composition dans la langue générale et les langues de spécialités.
Il explique que tous les composés complexes sont des extensions de l'un des types de base listé ci-haut. Dans cette classification, on retrouve des classes générales comme celle des « composés sur adverbes » qui elle même est subdivisée en sous-classes comme celle des « Adverbe de négation (ex- un non-lieu) » ou celle des « Adverbe de quantification (ex- la plus-value) » ou encore celle des « adverbe de manière (ex- le bien être) ». On y trouve également des classes de composés complexes comme la classe des « expansions de Nom Nom » tel que la forme « N/NA abrégé (ex- le bac math élém) » ou alors « NA / numero card (ex- l'ennui public numéro un) ». Certaines de ces classes contiennent des noms composés formés à partir de plus de 5 Noms et 3 Adjectifs comme dans le cas *« La servitude de libre passage sur les berges des cours d'eau non navigables ni flottables »*.

Nous étudions pour notre part dans le chapitre 7 des noms de profession composés qui s'avèrent souvent être très long, contenant des conjonctions de coordination, des signes de ponctuation (*« , / & »*, des surcompositions, des insertions, des abréviations. Nous reviendrons ultérieurement en détail sur leur typologie, leur composition et différentes formes.

5.3 Flexion des unités polylexicales

Le pluriel des mots composés est un des domaines les plus difficiles de la grammaire française selon [45]. Il existe certes des règles de transformations dérivationnelles des noms composés, mais il existe également un grand nombre d'exceptions dans chacun des cas.
Dans les composés de type « Nom Adjectif » par exemple, le pluriel est formé de telle sorte que les deux composantes se fléchies, sauf dans

« grand-mères » ou « petit beurres » etc. Pareillement pour les composés de la forme « Nom Nom » qui se fléchissent différemment en fonction de la tâche du second nom par rapport au premier. Ainsi dans le cas d'une apposition, les deux doivent se mettre au pluriel (ex : oiseaux-mouches) alors que s'ils sont dans une relation de compléments seul le premier nom est à fléchir (ex : timbres poste).

Ces composantes qui partagent les traits morphologiques du mot composé auquel elles appartiennent sont appelées les « têtes » ou les « constituants caractéristiques ». Pour le français et comme étudié par Savary dans sa thèse [60] les règles de reconnaissance des constituants caractéristiques pour les formes composées les plus étudiées dans la littérature sont :

$\underline{<Nom_1>}$ $<Nom_2>$
$\underline{<Nom_1>}$ $\underline{<\text{Adjectif}>}$
$\underline{<\text{Adjectif}>}$ $\underline{<Nom_1>}$
$\underline{<Nom_1>}$ à $<Nom_2>$
$\underline{<Nom_1>}$ de $<Nom_2>$
$\underline{<Nom_1>}$ prép $<Nom_2>$
$<$verbe$>$ $<Nom_1> \rightarrow$ pas de tête

Le problème de la flexion automatique des noms composés est traité dans le TALN par des règles générales où les exceptions ne sont pas décrites. Ainsi dans certains systèmes classiques les composés sont fléchis par la flexion de leurs composantes, ce qui comme nous venons de le monter ne reflète pas d'une correction grammaticale absolue. [60] s'est intéressée à la question pendant sa thèse, elle a proposé un nouveau formalisme qui permet de gérer les règles générales aussi bien que beaucoup d'exceptions afin d'éviter des fautes ou des insuffisances dans la forme fléchie des dictionnaires électroniques. L'algorithme de flexion automatique des noms composés proposé est une méthode universelle, indépendante de la langue. Le tout réside dans le prétraitement des fichiers d'entrée qui décrivent les particularités des composés dans chaque langue.

Pour chaque langue traitée, un premier fichier est construit dans lequel les types de flexion à utiliser sont présents accompagnés pour chaque type de l'énumération de ses formes. Pour le français ce fichier ressemblerait à :

N : s,p
R : m,f
P : 1,2,3
T : W,P,I,J,F,G,K,S,T,C,Y

Où les lettres à gauche des « : » représentent les catégories de flexion et les élements à droite, les formes possibles pour chaque catégorie. Dans cet exemple, la lettre « R » fait figure du « Genre » qui peut prendre deux formes, soit « f » pour le féminin, soit « m » pour le masculin.

Comme 2ème entrée, l'algorithme nécessite le fichier dictionnaire DELAC, lui-même des formes lemmatisées des noms composés. Dans cette approche, il y aura autant de fichiers que de règles de flexion des mots composés à recenser. Pour le cas régulier des composés du type « Nom prep Nom (NPN) », on aura un fichier DELAC qui ressemblerait à [60] :

```
# +/-/-
avocat(avocat.N32 :ms) de le diable,N+NdeN :ms/+N+G33
boîte(boîte.N21 :fs) à musique,N+NaN :fs/+N
champs d.honneur,N+NA :ms
frère(frère.N1 :ms) de lait,N+NdeN :ms/+N
preuve(preuve.N21 :fs) par absurde,N+NPrepN :fs/+N
...
```

La première ligne du fichier renseigne sur chacune des composantes, un « + » signifie qu'elle devra être fléchie, un « - » qu'elle devra rester invariante. Si une classe admet des irrégularités, celles-ci sont décrites au niveau de l'entête. Chaque fichier contient des entrées homogènes de flexion. Si une entrée se laisse fléchir dans plusieurs formes elle doit être présente dans tous les fichiers des formes en question.

Dans cet exemple de la forme NDN, la première ligne du fichier explique qu'il s'agit d'une forme où uniquement le premier terme sera fléchi. Et les deux lignes suivantes décrivent l'irrégularité qui est ici la manière d'obtenir le pluriel. Chaque entrée de ce fichier subira deux transformations au pluriel, la première est régulière, où seul le premier *Nom* est mis au pluriel tandis que la seconde forme met les deux noms du composé au pluriel, pour obtenir ainsi les deux sorties « blancs d'oeuf » et « blancs d'oeufs ».

```
# +/-/-
# p :p/-/-
# p :p/-/p
blanc(blanc.N1 :ms) d.oeuf(oeuf.N1 :ms), N+NdeN :ms/+N
chef(chef.N31 :ms) d.état(état.N1 :ms),N+NdeN :ms/+N+G
...
```

Les algorithmes ont été développés pour le logiciel *Intex*. Cette méthode très rigoureuse permet d'éviter certaines erreurs dans le dic-

tionnaire électronique des mots composés fléchis, néanmoins le prétraitement est très lourd, surtout pour les composés très longs et irréguliers comme ceux que nous avons recensés comme noms de profession composés.

5.4 Dictionnaires DELA

Il existe un nombre important de dictionnaires informatisés qui sont plus ou moins une forme électronique de leurs versions papiers classiques. Ces dictionnaires sont des ouvrages destinés, comme pour leur versions sur support papier, à des utilisateurs humains ayant des connaissances du monde, des capacités d'analyse, des facultés d'interprétation par analogie et par déduction.

Ainsi dans ce type de dictionnaires on trouvera des définitions de mots en fonction de synonymes ou d'exemples concrêts qui demandent à un être humain certaines connaissances préalables pour être capable de les interpréter. Une autre caractéristique de ces dictionnaires est que l'on y sous représente les formes répétitives, comme les adverbes en *-ment* (joli - joliment) ou les adjectifs en *-able* (manger-mangeable), du fait de leur dérivation intuitive et régulière.

Dès lors que l'utilisateur cible n'est plus un être intelligent ayant des capacités d'analyse et de raisonnement mais une machine incapable de déductions automatiques et encore moins de compréhension par analogie, une machine qui ne peut traiter que des informations présentées formellement, il a été nécessaire de penser un nouveau format de dictionnaire qui permettrait à un tel interlocuteur, un ordinateur en l'occurrence, d'analyser un texte écrit en langue naturelle et de retrouver explicitement ces informations habituellement cachées. L'équipe du Laboratoire d'Automatique Documentaire et Linguistique (LADL) a développé dans ce but le système DELA, qui est un système linguistique et informatique de représentation des unités lexicales d'une langue et de traitement.

Le système DELA

Le format DELA est une description formelle et une classification systématique des unités du langage. Chaque unité représente une entrée indépendante dans le dictionnaire et est accompagnée d'un ensemble d'informations supplémentaires d'ordre syntaxique, morphologique et sémantique exploitables par l'ordinateur.

Dans une tâche d'analyse syntaxique, ce dernier découpe le texte suivant les séparateurs usuels (espace, virgule, point, trait d'union) et

doit être capable de comprendre chacune des unités formées grâce à leur présence dans les dictionnaires disponibles. Dans le cas échéant il engendrera un échec d'analyse.

La configuration générale du DELA se compose d'un ensemble de ressources linguistiques

dictionnaires des mots simples DELAS, DELAF
dictionnaires des mots composés DELAC, DELACF
dictionnaires phonétiques DELAP, DELAPF

et d'un ensemble de programmes informatiques de traitement

programmes de génération de formes fléchies
programmes de génération des automates syntaxiques du texte
...

Dictionnaire de mots simples DELAS

Il nous semble important de présenter brièvement le format du dictionnaire DELAS, le dictionnaire électronique du LADL pour les mots simples, et ce car il fût la base de réflexion des dictionnaires qui ont suivi.

Blandine Courtois décrit en 1989 dans [12], le format du DELAS, et de sa version fléchie DELAF. Elle donne une nouvelle définition du **mot simple** et explique aux conservateurs de la langue l'intérêt d'estimer chaque unité graphique de la langue comme entrée indépendante, ce qui n'est pas le cas traditionnellement.

Un mot Simple comme compris par le DELAS est « toute unité de texte définie sur l'alphabet ASCII ou ABCDIC à 256 caractères, et ne comportant aucun séparateur ».

Les mots à trait d'union ou à apostrophe sont donc fractionnés et traités séparément même si les fragments ainsi obtenus n'ont aucune autonomie et aucune sémantique indépendante.

Les mots : « Prud'homme » ou « ping-pong » seront donc décomposés respectivement selon les séparateur *« ' »* et *« - »* et les quatre graphies simples *prud, ping, homme, pong* représenteront des entrées indépendantes du DELAS. Si les unités *prud ping* et *pong* ne sont pas des objets autonomes, considérés comme mots simples par les lexicographes et les grammairiens, l'équipe du LADL considère que ce sont néanmoins des objets formels bien définis. Étant donné leur faible proportion par rapport à toutes les entrées du DELAS, et du fait que pour un ordinateur « tout élément graphique constitutif d'une expression équivaut à un mot formel ordinaire et peut servir de clé d'accès à l'information stockée dans des articles de dictionnaires », elle justifie ce choix de parler

de mots simples et non plus de graphies simples du français.

Le DELAS est une base de données contenant près de 110 000 entrées de graphies simples du français mises sous leurs formes lemmatisées (le masculin singulier nominatif pour les verbes, le singulier pour les noms, le masculin singulier pour les adjectifs et déterminants, etc). A chacune des entrées est associée :

- un symbole de la partie de discours
- un numéro de code morphologique
- une ou plusieurs catégories sémantiques

lemme,.code flexionnel+Sem

Ces trois informations représentent un *« code flexionnel »* décrivant la façon d'obtenir toutes les formes fléchies de ce mot à partir de sa forme lemmatisée.

Prenons l'exemple de l'entrée du DELAS français suivante :

(5.1) administrateur,N43+(Hum+Profession)

Ce nom masculin est fléchit selon le code flexionnel *« N43 »* représenté par le transducteur (5.1). Ce dernier décrit que la forme au masculin singulier(*administrateur*) est identique au lemme, que pour obtenir le masculin pluriel il faut ajouter un « **s** » et que pour obtenir le féminin singulier(*administratrice*) il faut effacer les trois dernières lettres, à savoir « **-eur** » et les remplacer par le suffixe « **-rice** ». Les séquences précédant les deux points sont donc les transformations à effectuer sur le lemme et les séquences qui suivent les deux points sont les traits morphologiques des formes fléchies obtenues. Nous pouvons observer dans l'exemple ci-dessous (ex- 5.2) les formes fléchies obtenues après application du transducteur (5.1).

(5.2)

administrateur,.N+Profession+Hum :ms
administrateurs,administrateur.N+Profession+Hum :mp
administratrice,administrateur.N+Profession+Hum :fs
administratrices,administrateur.N+Profession+Hum :fp

L'alphabet d'entrée d'un transducteur est constitué du symbole vide (<*E*>), du symbole *L* pour le retour à gauche, de l'alphabet de la langue traitée et des codes des traits flexionnels. A côté de ces opérations observées dans notre exemple, il existe encore la lettre *C* et la lettre *R* pour la recopie et le passage à droite. Ces deux opérations supplémentaires sont nécessaires dans le cas où une voyelle change dans l'une de ses formes fléchies, comme dans le verbe *« acheter »* où le « e » sans accent est remplacé par un « è » à la troisième personne du singulier présent (*ach**e**ter -> ach**è**te*. Ainsi le chemin correspondant à ce dernier exemple est : LLLLèRes :P2s :S2s.

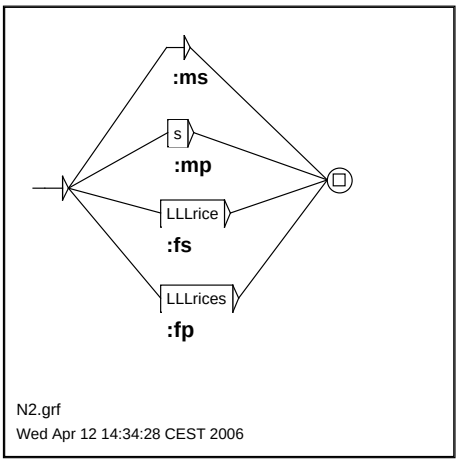


FIG. 5.1 – transducteur de flexion N2

Initialement	LLLLèRes	a c h e t e r _
Étape 1	**L**LLLèRes	a c h e t e r
Étape 2	**LL**LLèRes	a c h e t e r
Étape 3	**LLL**LèRes	a c h e t e r
Étape 4	**LLLL**èRes	a c h e t e r
Étape 5	**LLLLè**Res	a c h è t e r
Étape 6	**LLLLèR**es	a c h è t e r
Étape 7	**LLLLèRe**s	a c h è t e r
Étape 8	**LLLLèRes**	a c h è t e s _

TAB. 5.1 – Chemin de la grammaire V6

Dictionnaire des mots composés DELAC et DELACF

Le dictionnaire électronique DELAC est une liste des formes lemmatisées des mots composés. A chaque entrée sont associés comme pour le DELAS, les traits morphologiques, l'étiquette grammaticale, mais aussi la typologie du composé. Dans l'exemple 5.3, on observe une entrée du DELAC correspondant au mot composé « cousin germain ».

(5.3) *(cousin.N32 :ms) germain(germain.N32 :ms),N+NA :ms/++*

À travers cette description nous décelons qu'il s'agit d'un nom (N) de la typologie Nom-Adjectif (NA) et que cette forme canonique est au masculin singulier (ms). Les deux signes « ++ » apparaissant après la barre oblique signifient que ce nom doit se fléchir en nombre (singulier, pluriel) et en genre (féminin, masculin). L'alphabet possible à ce niveau est « +,- », un « - » dans la première position signifie que le composé ne doit pas se fléchir en nombre et un « - » dans la seconde position signifie que ce dernier ne se fléchit pas en genre.

Initialement, le DELAC contenait les 7 classes de noms composés *AN* (Adjectif Nom), *NA* (Nom Adjectif), *NAN* (Nom à Nom), *NDN* (Nom de Nom), *NN* (Nom Nom), *PN* (Préposition Nom), *VN* (Verbe Nom), ainsi que les adverbes composés recensés par M.Gross [3, 11] *PAC, PC, PCA, PCDC, PCDN, PCONJ, PCPC, PCPN, PDETC, PF, PJC, PPCO, PV, PVCO*, les conjonctions de subordination ainsi que toutes les formes étudiées par Colas-Mathieu [44].
Les adverbes et les conjonctions étant en principe invariables, le travail le plus ardue fût de décider de l'étiquette grammaticale de certaines composantes simples dans les composés, par exemple est ce que *« mort-né »* doit être classé dans le groupe *NA* ou dans le groupe des *AA* ou alors dans le groupe *NN*. Pareil pour les cas de *« marche avant », « marche arrière »*, est ce qu'il faut les ranger dans la classe des *NN* ou bien est ce que le second terme est un adverbe invariable. Ces décisions jouent un rôle très important dans la phase de flexion comme nous avons pû nous en rendre en compte au niveau de la section 5.3. Un second problème à gérer au niveau de la description grammaticale des noms composés est que la catégorie grammaticale du tout n'est pas forcément la même que celle de ses composantes, ce phénomène est rencontré surtout dans les composés sans tête. De plus il arrive que la description morphologique du nom composé soit différente en genre ou en nombre de ses composantes comme dans *« un peau rouge »* ou *« une deux chevaux »*. Il arrive aussi qu'un composé ait une forme féminine quand ses constituants n'en ont pas (un après guerre).

Nous proposons un échantillon du DELAC fléchi [64] dans ce qui suit :

cousins germains,cousin germain.N+NA+Hum :mp
criant de vérité,.A+EPN :ms
pommes de terre,pomme de terre.N+NDN+Conc :fp
tant et si bien que,.CONJS+3
tout à coup,tout à coup.ADV+PCPC
...

conclusion

Dans la suite de ce mémoire nous montrerons nos besoins en construction de dictionnaires électronique de mots simples et composés pour notre langage spécialisé des offres d'emploi. Nous avons opté pour un formalisme DELA dans les différents cas car nous usons également du logiciel *Unitex* pour construire nos grammaires locales et patrons d'extraction dans le corpus des pages d'accueil d'entreprises d'une part et celui des offres d'emploi d'autre part.

CHAPITRE 6

La reconnaissance automatique de sites Web d'entreprises

Introduction

Un système de recherche d'information se compose essentiellement d'une collection documentaire, à laquelle s'applique un certain nombre d'algorithmes de traitement de la langue naturelle et des algorithmes de traitement de textes non structurés. A cela s'ajoutent des algorithmes de recherche permettant à un utilisateur à la quête d'informations d'extraire les documents pertinents correspondants à sa demande. Un moteur de recherche, cas particulier d'un système de recherche d'information, dispose non pas d'une collection documentaire mais d'une collection de sites Web. La première question fondamentale qui se pose est : *Comment est ce qu'un tel système se ravitaille- t-il de cette collection ?*

La réponse est différente en fonction de la nature des moteurs de recherche. Les techniques de collecte de sites Web d'un méta- moteur de recherche (mMR) comme Google ou *Yahoo* se distinguent de celles utilisées par les moteurs de recherche spécifiques(MRs) comme les moteurs de recherche immobilier ou les moteurs de recherche d'emploi.

Pour ce qui est des méta-moteurs de recherche (mMR), ils sont en quête d'une collection de sites Web variée. Leur but est de disposer d'une base de données hétérogène en terme de sémantique portée par les documents. Plus les thèmes disponibles seront divergents plus les utilisateurs trouveront des documents correspondants à leurs divers besoins, plus ils seront satisfaits et par conséquent fidèles à ce seul moteur de recherche. C'est pourquoi il suffit à un tel système d'explorer le Web à l'aide d'un robot d'indexation, appelé aussi *« Crawler, Spider ou agent Web »* qui suit récursivement tous les liens hypertextes qu'il

trouve à des profondeurs pouvant aller de 1 à n. Cette exploration est lancée depuis une liste d'URLs initiale récupérée sur un annuaire Web, ou alors émise manuellement. Étant donné qu'une page Web peut référencer beaucoup d'autres pages qui n'ont aucun rapport sémantique avec elle, un tel Crawler aura collecté une grande quantité de nouvelles URLs à la fin de son traitement, qu'il est cependant incapable de classer dans des catégories sémantiques dépendamment de celles d'origines ce qui ne joue d'ailleurs aucun rôle, vu que le but essentiel est d'amasser un nombre très grand et très varié de documents.

Dans le cas d'un moteur de recherche spécialisé (MRs), cette démarche n'est plus valide car ce qui importe dans un premier abord est la thématique des pages Web à stocker et non pas leur quantité, bien que la quantité n'est pas à négliger non plus. Ces systèmes sont souvent dépendants des entreprises ou des particuliers qui leurs fournissent les documents à publier. Prenons l'exemple concret du moteur de recherche immobilier. La collecte des offres immobilières se fait de la manière suivante : soit un individu privé ou représentant d'une organisation lucrative, disposant d'un logement libre et désirant le louer ou le vendre. Il s'adresse au dit moteur de recherche et y dépose son annonce moyennant une somme d'argent dépendante de la durée de publication désirée. Nous pouvons comparer un tel système à la rubrique immobilière d'un journal. Plus le journal a une renommée importante et un grand quota de lecteurs, plus les annonceurs désirent voir leurs communiqués publiés dans ce dernier, car il leur garantie un grand nombre de lecteurs et ainsi d'acheteurs ou de locataires potentiels. C'est exactement ce phénomène qui se reproduit dans le Web. Un tel système est complètement dépendant de sa notoriété et des personnes qui lui confient leurs offres.

A partir de cette dernière constatation nous pouvons voir la faiblesse majeure de ces systèmes et montrons dans la suite de ce chapitre en quoi notre système se différencie de ses semblables dans ce point particulier de la constitution de la base de données des annonces.

Par ailleurs, le fait que ce service de publication des offres soit un service coûteux, engendre que certaines entreprises ou certains organismes préfèrent ne pas recourir à ces annuaires spécifiques et se contentent de publier leurs annonces sur leurs propres sites Web. Seulement pour que ces communiqués soient trouvés et visités, il est indispensable que le site hébergeur soit déjà indexé par certains méta-moteurs de recherche, ou qu'il soit assez connu pour que son URL soit entrée directement au niveau du navigateur Web. Uniquement dans ces deux cas, il est possible à un utilisateur à la recherche d'un appartement ou d'un emploi d'accéder aux annonces disponibles « cachées ».

Pour résumer : les annonces quelque soient leurs types, sont éparpillées sur le Web entre les différents systèmes de recherches spécifiques existants,et sur les différents sites Web personnels qui restent souvent non atteignables. Ainsi un utilisateur à la recherche d'un emploi par exemple se trouve confronté à deux situations : ① Soit il se tourne vers les différents moteurs de recherche spécifiques d'emploi et qui contiennent une quantité dérisoire d'offres par rapport au nombre réel de postes vacants publiés sur la toile. ② Soit il se tourne vers un méta-moteur de recherche et dans ce cas, il doit formuler sa requête de telle sorte qu'il n'y ait pas la moindre ambiguïté, vu la diversité sémantique des documents indexés. En effet un nom de profession aussi compliqué soit-il ne garantie pas de trouver des offres d'emploi quand il est soumis à un mMR.

A la suite de ces investigations et en tenant compte des différents inconvénients et points faibles des systèmes similaires, il nous a fallu développer, pour notre moteur de recherche d'emploi, une autre stratégie de recherche et de collecte de documents qui puisse nous assurer une certaine cohérence de résultats et qui nous évite de parcourir trop de pages Web non compatibles à nos besoins, tout en nous assurant une bonne couverture du domaine. Nous tentons ainsi de découvrir automatiquement les différentes offres d'emploi disponibles sur le Web pour nous dégager de la dépendance des entreprises dont nous avons parlé ci-dessus.

Nous nous sommes donc posé la question suivante : *Où est-ce que les offres d'emploi sont-elles publiées sur le Net ?* Elles sont publiées :

- soit directement sur les pages Web des entreprises proposant les postes
- soit sur les moteurs de recherche spécifiques à l'emploi
- soit sur la rubrique emploi de certains journaux en ligne
- soit sur des forums de discussions, ou des pages d'informations

Nous avons pu constater que les entreprises ont tendance à poster leurs propositions d'emploi sur leurs sites Web et ce, même si elles ont recours aux services de publication chez les spécialistes.

C'est ainsi que nous avons eu l'idée d'aller chercher les offres à leurs sources et avons mis en place deux processus de découverte d'offres d'emploi.

Au niveau du premier processus, nous recherchons les postes vacants dans les pages Web des entreprises et au niveau du second processus nous faisons appel à un agent Web spécialisé muni d'un ensemble de

locutions et formules typiques au sous-langage des offres d'emploi que nous avons constitué. Le fragment extrait de l'architecture générale de notre système et présenté à la figure 6.1 résume bien la phase de collecte de documents de notre système. On peut y voir un troisième aspect noté « Adaptateur Entreprise » qui est un module post-apprentissage pour la découverte des offres d'emploi dans les pages Web d'entreprises déjà connues par le système. En effet une fois les descripteurs internes d'une entreprise appris, il est inutile de lancer de nouveau le processus d'analyse entier qui peut s'avérer assez lent en fonction de la taille du site Web et le nombre de liens qui y sont présents.

Dans la suite de ce chapitre nous nous concentrons sur la première procédure, à savoir la collecte de documents dans les sites Web d'entreprises et en particulier sur la reconnaissance automatique de ces derniers.

6.1 Motivations et étude de cas

Un moteur de recherche ordinaire lance un Crawler auquel il passe en argument une liste quelconque d'URLs d'initialisation, cette liste est souvent récupérée aléatoirement sur des annuaires électroniques, ou à la suite de requêtes sur d'autres mMR. Comme le but de notre système n'est pas d'amasser toute sorte d'URLs, il est essentiel que la liste des URLs de lancement soit déjà filtrée. Le premier processus de collecte de documents dont nous parlons ci-dessus impose que la recherche se fasse dans un ensemble de sites Web d'entreprises, c'est pourquoi il nous est

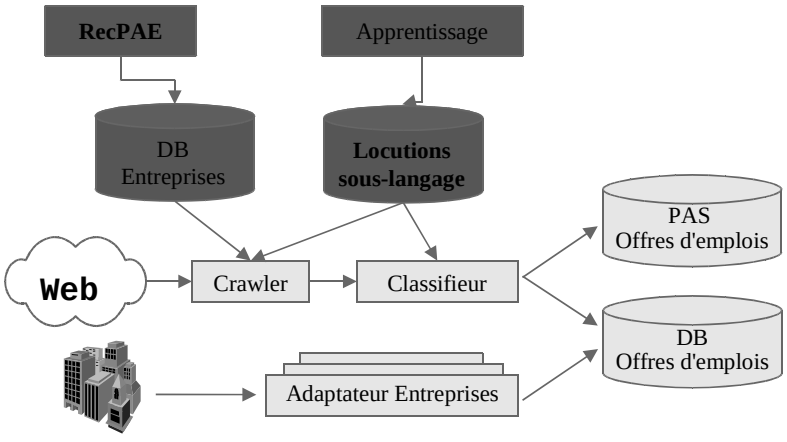


FIG. 6.1 – La collecte des offres d'emploi

indispensable de bénéficier d'une telle collection.

Or la première question qui se pose à ce niveau est : *comment réunir une liste d'entreprises qui peut augmenter de jour en jour et qui peut être très grande ?* et la seconde question est : *comment retrouver les offres d'emploi proposées par une entreprise en connaissant uniquement son adresse sur la toile ?.*

Il existe certes un grand nombre d'annuaires d'entreprises (les pages jaunes[1], Kompass [2], europass[3]) que nous pourrions prendre tels quels, ils sont toutefois confrontés au problème de la non exhaustivité. Il existe essentiellement deux genres d'annuaires. Le premier genre renferme les annuaires, où une entreprise voulant se faire connaître, décide de s'y inscrire. Celle-ci doit donc remplir un formulaire prédéfini et choisir entre autres la catégorie professionnelle à laquelle elle appartient. L'avantage d'une telle base de donnée, même si loin d'être complète, est la fiabilité des données existantes et surtout leurs catégorisations irréprochables. Dans le second type d'annuaires payants en général, c'est du personnel humain qui essaie d'évaluer, de rechercher et de classifier les entreprises. Ce genre de systèmes a l'avantage de couvrir une palette plus large que le précédent, mais la catégorisation n'y est que subjective. En effet du personnel non spécialiste, doit classifier quantité d'entrées souvent très ambiguës. Il existe aussi les registres de l'INSEE[4] qui répertorie toutes les entreprises enregistrées en France, cependant un certain nombre de critères est indispensable pour y avoir libre accès.

Vu l'évolution rapide du marché, on constate tous les jours la création nouvelle d'une quantité importante d'entreprises et la dissolution d'autres. Ainsi il est impossible de prétendre pouvoir tenir un tel registre à jour, ou bien même de détenir un registre couvrant tel ou tel domaine en se basant sur une identification et classification manuelle humaine.

C'est pourquoi nous avons choisi de développer un outil de reconnaissance automatique de pages d'accueil d'entreprises -**PAE**-. Ce système, que nous avons baptisé **RecPAE**, est une adaptation pour la langue française d'une première ébauche développée par *A.Filzmeyer* [22] dans le cadre de sa thèse et pour les pages d'accueil d'entreprises en langue allemande.

Dans [69], Svatek et al se sont aussi intéressés à cette question. Ils ont développé le système *Rainbow* de découverte de descripteurs d'organisations en se basant sur la structure HTML d'une part et d'autre part sur le contenu textuel. Cependant ils ne donnent pas de détail de

[1]http : //www.pagesjaunes.fr
[2]http : //www.kompass.fr/
[3]http : //www.europages.fr/
[4]INSEE : Institut national de la statistique et des études économiques

fonctionnement ni de précision sur les performances de leurs systèmes.

La question de la classification automatique des pages Web a été soulevée à maintes reprises dans la littérature. Il s'agit d'une part de la classification en genre des pages et de l'autre part de leur classification sémantique. La catégorisation sémantique peut être observée au niveau des différents annuaires Web tel que *Open directory*[5] ou alors son prédécesseur *yahoo*[6], où 14 catégories sémantiques principales sont mises à la disposition des internautes pour leur permettre une navigation plus ciblée à travers les documents pré-classifiés. Un papier très intéressant sur la catégorisation des pages Web de yahoo et de OPD est [41]. Pour de plus amples informations sur la catégorisation automatique de textes en général et de sites Web en particulier, nous proposons de voir [75], [74], [38] qui proposent un état de l'art sur ce domaine de recherches.

[2], par exemple s'est intéressé en 2001 à la classification en genre des pages Web, il a identifié 3 catégories dans lesquelles il affirme pouvoir ranger n'importe quelle page Web qui sont :

- Pages d'information : caractérisées par un logo en haut de la page et suivi d'un menu de navigation. Ces pages auraient un taux élevé de textes d'ancres
- Pages personnelles : caractérisées par les coordonnées de la personne, son nom et sa photo
- Pages de recherche : caractérisées par une quantité élevée de textes, d'équations, de graphes d'images

Bien que les techniques du Web aient évolué, certains critères restent invariants. En observant une page personnelle créée en 2006, elle sera certainement développée en *Flash*, remplie d'animations,et d'effets multimédias divers, mais elle contiendra tout de même une rubrique « Contact » où l'on retrouvera le nom, prénom et coordonnées du propriétaire.

Dans notre cas et étant donné nos besoins, nous nous intéressons à une classification binaire. Notre but est de décider en recevant une URL en entrée s'il s'agit d'une page d'accueil d'entreprise ou pas. Son domaine d'activité ainsi que la catégorie sémantique dans laquelle elle immerge, ne nous sont pas utiles. En effet, nous tentons de cerner les pages Web d'entreprises pour y retrouver les offres d'emploi potentielles. Or une entreprise de développement logiciel qui serait classée dans la catégorie sémantique « IT-Consulting » peut aussi bien manifester le besoin d'embaucher « un chef cuisiner » pour la cantine, ce qui

[5]http : //www.opendir.org, connu sous l'acronyme OPD
[6]http : //www.yahoo.fr

prouve que la catégorie de l'entreprise ou son activité principale ne permettent en aucun cas de classifier automatiquement les offres d'emploi qu'elle propose. Nous ne cherchons pas non plus à classifier en genre toute sorte d'URLs, comme c'est le cas de [2], nous nous contentons de reconnaître celles qui nous intéressent, à savoir les Pages d'accueil des entreprises.

Comment est-ce que nous, les humains, distingue-t-on une page d'accueil d'une entreprise par rapport à n'importe quelle autre page Web en admettant que le nom de celle-ci nous est inconnu ?

La réponse est que nous nous référons à un certains nombre de descripteurs, de syntagmes et de locutions types que nous nous attendons à trouver sur une PAE et qui ne sont pas forcement présents sur les autres types de pages Web. Nous proposons au niveau de la figure 6.2, la page d'accueil de l'entreprise « ADIFCO » , que nous avons choisi pour sa richesse en descripteurs.

En observant la PAE de la figure 6.2, nous sommes capables, nous

FIG. 6.2 – Page d'accueil d'une entreprise

êtres humains, d'affirmer sans peine qu'il s'agit bien d'un Site Web d'entreprise. Nous pouvons même identifier le nom de celle-ci et savoir d'ores et déjà quel genre de services et produits sont offerts.

Quels sont les critères, nous permettant d'assurer qu'il s'agit bien d'une PAE ?

- Lien : **Présentation Société**
- Lien : **Nos activités**
- Lien : **Recrutement**
- Lien : **Nous contacter**
- Lien : **Dossier de presse**
- Contexte : **Forme Juridique :** *SARL*
- Contexte : **Créée en** *2003,ADIFCO* **est une société**
- Contexte : **Capital Social :** *80000* **eur**
- Contexte : **Siège Social :** *14E rue Pierre de Coubertin, Parc Mirande, 21000 Dijon*
- Contexte : **SIREN :** *451 292 544 RCS Dijon*
- Contexte : **Copyright** © *ADiFCO, 2006.*

Les 10 critères énoncés, prouvent que les PAE se différencient des autres pages Web par leur structure HTML d'une part mais aussi pas leur terminologie technique d'autre part. Les 5 premiers indices précédés par l'étiquette *« Lien »*, représentent la structure interne du site Web, or l'analyse unique de cette structure ne suffit pas à la compréhension automatique du site Web par une machine, elle doit être accompagnée d'une analyse sémantique.
Comme on peut le voir, cette page Web renferme plus de liens hypertextes que les quatre listés ci-dessus : « ADIFCO », « SUD-Référencement », « SUD-CréationWeb », « Utilisable » sont aussi des liens hypertextes de cette page, seulement ils ne sont pas directement exploitables, car leur sémantique n'est pas porteuse d'informations communes repérables durant la phase d'apprentissage et car elles référencient aussi des pages externes. Elles ne font, par conséquent, pas partie de la structure interne du site Web

Sachant que l'URL correspondante à cette page Web est : `http://www.adifco.com`. Nous sommes capables, nous êtres humains, de faire une concordance entre le nom du domaine[7] (adifco) et la chaîne

[7]Le nom de domaine est l'adressage d'une machine sur Internet géré par les serveurs dits Domain Name Server (DNS).

ADIFCO, retrouvée à maintes reprises dans des contextes appropriés, pour en déduire qu'*ADIFCO* est le nom de l'entreprise de la page Web que nous visitons.

Nous avons usé de l'étiquette *Contexte*, pour référencer les syntagmes ou la terminologie typique aux entreprises.

Le numéro SIREN[8], par exemple, est un code INSEE[9], identificateur unique d'une entreprise française au niveau national. Ce numéro de 9 chiffres peut être combiné avec le numéro NIC[10] pour former le numéro de SIRET[11]. Ces deux identificateurs ne peuvent être attribués à n'importe qui, seule une personne juridique, physique ou morale, peut s'en voir attribuer après vérification par la CFE[12]. Cependant la chaîne de caractère « Siret » est aussi le nom d'une rivière de Moldavie en Roumanie. C'est pourquoi le fait de trouver le mot Siret dans une page Web ne suffit pas pour en déduire qu'il s'agit du Numéro de SIRET, identifiant national des entreprises.

Le *siège social* est un autre exemple de terminologie spécifique aux entreprises. Le siège social n'est rien d'autre que la désignation officielle de l'adresse postale de celle-ci. Ainsi une page Web autre qu'une PAE, désignera l'adresse postale par *« Adresse »* ou *« Adresse postale »* ou *« Domicile »* ou encore *« Domiciliation »* mais jamais *« siège social »*.

Pareillement pour la *« Forme juridique »*, seul une société ou une association peut avoir une forme juridique, une SARL[13] dans cet exemple.

Ces derniers exemples montrent l'importance de cerner la terminologie commune employée en général par les entreprises. Nous ne visons pas l'identification de la terminologie des différents domaines d'activités existants sur le marché, ce qui entrerait plus dans une étude sémantique ; nous nous intéressons plutôt à l'identification du langage spécial attaché à toute structure sociale dirigée par un entrepreneur et ce, quelque soit la taille, le capital, le nombre d'employés ou le domaine d'activité.

La notion de sous-langage soulevée par Harris [37] depuis les années soixante jusqu'aux années quatre-vingt-dix, semble bien s'accorder avec notre besoin. En effet, il ne s'agit pas pour nous de recenser uniquement la terminologie : les mots simples et composés utilisés mais aussi de construire des grammaires descriptives des structures syntaxiques et lexicales et des phénomènes linguistiques propres à ce langage spécialisé.

[8]Système d'Identification du Répertoire des ENtreprises
[9]INSEE : Institut national de la statistique et des études économiques
[10]NIC : Numéro Interne de Classement
[11]SIRET : Système d'Identification du Répertoire des ETablissements
[12]Centre de formalités des entreprises
[13]SARL : Société à responsabilité limitée

Selon Sager [57] :

Définition 4 *Le caractère distinctif d'un sous-langage, c'est que pour certains sous ensembles des phrases du langage, les restrictions de sélection, pour lesquels on ne peut pas fournir de règles pour le langage dans son ensemble, intègrent la grammaire.*

Selon Harris, ces sous-langages se caractérisent par un lexique limité et par l'existence de schémas de phrases en nombre fini.

Nous profitons de ces définitions et conséquences pour tenter de délimiter le sous-langage des entreprises. La méthode générale pour l'étude d'un sous-langage de domaine décrit par [57] est la suivante :

Définition 5 *Si l'on applique à un corpus de textes d'un secteur scientifique des méthodes de linguistiques descriptives similaires à celles utilisées pour le développement d'une grammaire d'une langue dans son ensemble, on obtient des motifs précis de cooccurrences de mots à partir desquels on peut définir des sous-classes de mots et des séquences de ces sous-classes qui sont caractéristiques (c'est-à-dire une grammaire). Ces catégories lexicales et formules syntaxiques de la grammaire du sous-langage sont étroitement corrélées aux classes d'objets du monde et aux compositions qui sont propres à ce domaine. Ils nous fournissent donc un ensemble de structures sémantiques pour refléter les connaissances de ce domaine.*

Dans un corpus traitant d'un domaine particulier, les mots appartenant au lexique du sous-langage forment la majorité des mots employés. L'étude des mots sémantiquement pleins[14] les plus occurrents dans ce corpus permet donc de former un lexique du sous-langage. Bien que Harris limite la potentielle existence de sous-langage aux domaines scientifique et technique et bien qu'il ne s'agisse pas dans notre cas de rechercher un lexique propre à un domaine scientifique, nous persistons à appeler notre sous-ensemble lexicale et grammaticale, un sous-langage et ce car vérifiant toutes les caractéristiques d'un sous-langage comme décrit par Harris et Sager.

Afin d'automatiser la tâche de classification des pages Web, en page d'accueil d'entreprise ou pas, nous avons développé un outil *multi-tâches*, qui extrait les différentes informations sur plusieurs niveaux d'analyses.

Étant donné une URL, nous interagissons tout autant avec la structure HTML et les informations cachées qu'avec le contenu textuel de celle-ci. Dans notre cas, il est très important de différencier entre le

[14] sémantiquement pleins par opposition aux mots vides, tel que les préposition, les articles, etc

squelette de la page et son contenu textuel. Et ceci car nous appliquons des traitements différents en fonction de la nature de l'objet.

Nous reconstruisons la structure de notre système de reconnaissance automatique de PAE, RecPAE, dans la figure 6.3.

Nous pouvons y distinguer les différentes analyses :

- ◇ niveau structurel : analyse de la structure HTML
  - analyse de l'URL
  - analyse des meta-informations : description, Keywords
  - analyse du titre
- ◇ analyse de la topologie des liens hypertextes
  - analyse des textes d'ancres
  - analyse de la structure des liens
  - analyse des scripts CSS et Javascript
- ◇ niveau contextuel : extraction d'information
  - extraction de l'adresse, du numéro de téléphone, du numéro de SIRET, ...

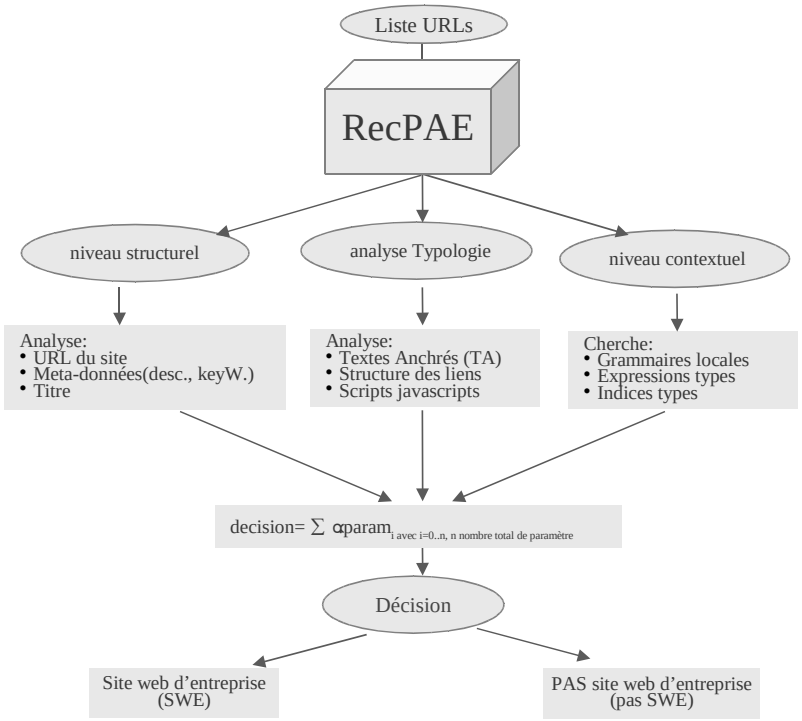


FIG. 6.3 – Différents niveaux d'analyse du Système **RecPAE**

– extraction de formules syntaxiques à structures sémantique précises

6.2 Terminologie

Nous appellerons dorénavant *cadre du document*, l'ensemble (URL, Titre, Meta-informations, Liens internes, Textes d'ancres, Objet référencés et descriptions liés) et *« Strippe »*, la version strippée de la page Web. Une version strippée d'une page Web correspond au texte pure de celle-ci, sans liens hypertextes ni images ni objets multimédias quelconques.

Dans la figure 6.4 , nous spécifions les différents constituants du cadre du document Web. Nous nous sommes arrêtés uniquement sur les éléments clés qui nous intéressent dans la suite de ce chapitre et dont nous avons besoin pour le développement de notre système RecPAE de reconnaissance automatique de page Web d'entreprises.

- le *titre* de la PAE : contient souvent le nom et le slogan de la

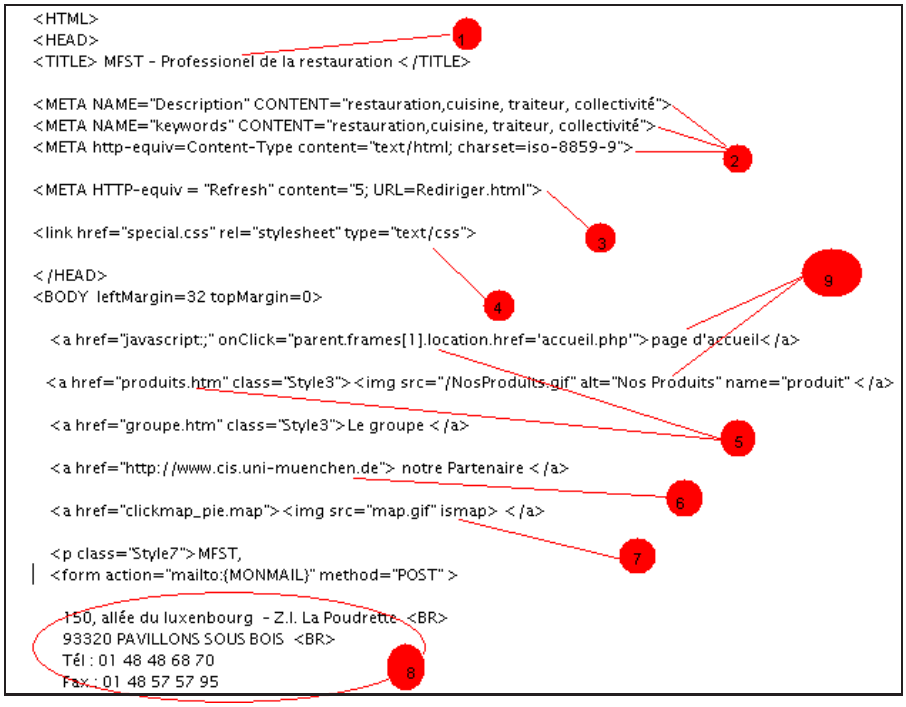


FIG. 6.4 – Cadre du Document

compagnie.

- les *meta-informations* (mots clé, description, adresse...) : décrivent plus amplement la compagnie, son domaine d'activité, ses services, ses produits.
- la *redirection* de la page : sert souvent à publier les offres exceptionnelles pendant quelques secondes avant de lancer la page d'accueil
- les *scripts* (Javascript ou CSS) : définition de certaines actions et du style de la page
- les ***liens internes*** : servent à la navigation interne dans le site de la PAE. Le lien peut être complet ou partiel(voir l'exemple).
- les *liens externes* : référencient des URLs extérieurs au site Web
- les *images* et les *clickmap* (les images cliquables) : détiennent des attributs descripteurs de l'image.
- les textes d'ancres : décrivent les liens référencés.
- frames

Pour reprendre une comparaison faite par Chi dans [9], le cadre d'un document Web est équivalent à la table des matières dans un document classique. Dans les manuscrits imprimés, une table des matières renseigne sur la structure du document, sur les différents points traités et la décomposition en chapitres, sections et sous sections. Le cadre du document est comparable, sauf qu'il propose encore plus d'informations. En effet, la topologie des liens internes renseignent sur la relation entre les différentes sections et les textes d'ancres peuvent être comparés aux titres et sous-titres dans les documents classiques. Ensuite la quantité d'information cachée à l'internaute, bien qu'elle soit cachée, renferme des informations de très grande importance pour la compréhension des pages Web. Ces Informations sont d'ailleurs plus importantes pour la compréhension de la page Web que ne l'est la table des matières pour la compréhension du document correspondant. Et ce, car les descriptions et meta-informations disponibles sur chaque page d'un site Web sont largement plus informatives sur la page même et sur le site en général. Les informations cachées à l'internaute bien qu'elles soient cachées sont très informatives, elles décrivent le but du site et résume son contenu, on y trouve aussi un ensemble de mots clés représentant la page Web et le site en entier.

6.3 Recensement des descripteurs

Dans une première phase d'apprentissage, nous nous sommes concentrés sur la collecte des indices structurels sémantiques propres

aux entreprises et particulièrement au langage des entreprises sur le Web. Même s'il n'existe pas de standards pour les pages d'accueil des entreprises, il existe un certain nombre de critères ergonomiques et de normes d'accessibilité pris en compte et respectés par la plupart des concepteurs de site Web d'entreprises.

La page d'accueil est l'image que l'entreprise donne d'elle même au reste du monde, l'enjeu est de taille, car pour de nombreux clients la consultation du site Web fait partie des démarches préalables à tout contact commercial, sachant que 10 secondes est le temps dont on dispose pour convaincre les visiteurs de rester sur le site. Les concepteurs veillent à présenter au mieux et le plus succinctement possible les informations sur les activités principales de l'entreprise, ses chiffres, ses produits et services, ses clients et ce en usant du maximum de mots clés possible repérables très vite par l'internaute, client futur potentiel.

C'est ce que nous avons tenté d'identifier et que nous décrivons plus en détail dans la suite de cette section.

6.3.1 Collecte et classification des textes d'ancres

Dans un premier temps, nous nous sommes surtout intéressés aux rubriques identifiables par les textes d'ancres et atteignables par les liens hypertextes correspondants. Ceux-ci ont été analysés et classifiés dans quatre catégories sémantiques. Chacune de ces catégories équivaut à un descripteur parmi les descripteurs fixés pour la prise de décision finale.

Afin de mener à bien cette tâche et d'essayer de réunir un ensemble significatif et représentatif du domaine, nous avons téléchargé 1200 PAE à partir de la rubrique « Business » de l'annuaire gratuit *Open Directory*[15]. L'avantage étant que la collection est hétérogène en terme de domaine d'activité. Nous avons ensuite identifié tous les liens internes[16] ainsi que leurs textes d'ancres correspondants. Dans le cas où une image vient remplacer le texte d'ancre, nous avons pris la description de l'image identifiée par l'attribut *src* comme représentant du lien. Nous avons pu rassembler un ensemble de textes d'ancres de longueurs et de types grammaticaux différents que nous avons tout d'abord normalisés puis analysés et classés manuellement dans cinq catégories sémantiques distinctes.

[22] avait identifié 9 classes pour ranger les textes d'ancres des entreprises repérés dans les sites de langue allemande, nous nous sommes

[15]OPD : http : //www.opendir.org

[16]un lien interne est un lien hypertexte ou un hyperlien dont la destination est un objet du même site Web. L'objet peut être un paragraphe dans le même document ou un autre fichier de la même hiérarchie

contentés d'en fixer six qui se sont avérées assez représentatives de l'ensemble des descripteurs pertinents découverts sur les PAE de langue française.

Catégories :

- Carrière
- Produits/Services
- Contact
- Société
- Presse
- Clients/Partenaires
- Non pertinent

Nous présentons au niveau du tableau 6.3.1 la distribution statistique de l'ensemble des textes d'ancres extraits à partir de cette collection de PAE de domaines divers. Il y a certes beaucoup de textes d'ancres classifiés comme étant non pertinents, ceci s'explique par des noms de produits ou les noms des sociétés qui sont souvent utilisés à la place de la locution « qui sommes nous ? ».

Carrière	85
Produits/Services	80
Contact	79
Société	335
Presse	30
Clients/Partenaires	77
Non pertinent	>2500

TAB. 6.1 – Distribution des textes d'ancres extraits à partir de 1200 PAE

Dans la table 6.3.1, nous avons listé des échantillons du contenu des différentes catégories.

L'hypertexte est un document informatisé, composé de noeuds reliés entre eux par des liens. La nature de ces noeuds peut être aussi bien textuelle, que visuelle, sonore ou encore audio-visuelle, on parle alors plutôt d'hypermédia. Or étant donné que nos recherches se limitent aux données textuelles et que de plus en plus de sociétés usent de la technique d'ancrage à base d'images ou autres, il nous a fallu remédier à cette perte d'information en nous intéressant directement à la topologie des liens en question dans le cas où le texte d'ancres est absent.

Carrière	nous recrutons jobs nos offres d'emploi joignez-vous à notre équipe ....
Produits/Services	Nos Produits accès à la boutique notre selection produits voir nos offres spéciales ...
Contact	Nous contacter comment se rendre à l'agence Pour venir nous voir nos coordonnées ...
Société	Notre Société qui sommes nous ? en savoir plus sur Entreprise ...
Presse	Communiqués de Presse La presse et nous presse infos média ...
Clients/Partenaires	Clients espace clientèle nos partenaires relation clients ...
Non pertinent	EuroDesign Prêts personnels souscrire en ligne cliquez ici ...

TAB. 6.2 – Échantillon des textes d'ancres

Le nom de la page liée et qui est symbolisé par la feuille dans la topologie d'un hyperlien, est souvent représentative du contenu qu'elle porte, c'est pourquoi nous avons tenté d'analyser le nom des feuilles et avons constitué un ensemble de catégories parallèles à l'ensemble des textes d'ancres classés décrits précédemment. Ce nouvel ensemble répertorie des chaînes de caractères souvent composés à partir des mots retrouvés dans les catégories des textes d'ancres.

Un lien hypermédia obéit à des règles syntaxiques strictes où les espaces, les apostrophes et les caractères spéciaux ne sont pas permis, c'est pourquoi pour nommer un document lié par le texte « qui sommes nous ? » il est usuel de remplacer les espaces par des caractères admissibles ou de les supprimer, ainsi le fichier correspondant se nommerait « qui_sommes_nous.html » ou « quisommesnous.html » ou encore « qui-sommes-nous.html » . Dans la même phase d'apprentissage, nous avons également répertorié la liste de toutes les feuilles des liens internes initialement extraits et avons essayé de les classifier dans les mêmes catégories sémantiques identifiées pour les textes d'ancre. Nous avons obtenu un peu moins de chaînes pertinentes que nous n'en n'avons trouvé pour les textes d'ancre, mais celles trouvées ont montré une grande importance dans l'amélioration des performances de notre système.

Carrière	*Contact*	*Société*	*Non pertinent*
30	37	98	>3000
nousrecrutons	contactez_nous	notre_soc	ps_124
jobs	itineraire	quisommesnous	interne9
...	...	...	...

TAB. 6.3 – Échantillon du contenu des noms de documents identifiés à partir de l'analyse des URLs

6.3.2 Grammaires locales de locutions types

La création d'une société qui consiste à donner naissance à une nouvelle personne, juridiquement distincte des associés fondateurs, que l'on nomme « personne morale », nécessite de lui donner un nom, appelé dans la terminologie officielle « dénomination sociale » ou encore « raison sociale », de la domicilier dans un local adapté nommé « siège social », de lui apporter un minimum d'argent et/ou de biens pour faire face à ses dépenses, ceci constitue son « capital social » , nécessite également la désignation d'une ou de plusieurs personnes chargées de

l'administrer et de la représenter, ce sont ses « dirigeants » et il faut en outre choisir sa « forme juridique ».

Ces données sont un point commun entre toutes les entreprises quelque soit le pays où elles ont été fondées, leurs tailles ou leurs capitaux. La présentation d'une entreprise contient donc au moins les données précédemment signalées. Certains autres paramètres variables, ne pouvant exister que dans le cadre juridique des organisations, peuvent aussi être décrits, c'est le cas du numéro de SIRET par exemple qui est le numéro d'enregistrement d'une entreprise française dans le Registre de l'INSEE, ce numéro identifie une société sur le territoire français uniquement, son équivalent en Allemagne est connu sous le numéro UstldNr.

Nous nous sommes consacrés à l'identification des descripteurs juridiques officiels mais aussi aux descripteurs locaux[17] afin d'augmenter les performances de notre système.

Dans cette phase il a été question de construire des grammaires locales pour l'extraction des informations citées précédemment. A côté de ces paramètres nous avons également identifié certaines locutions types statistiquement redondantes dans le corpus des PAEs. Au lieu de vérifier l'existence ou pas des locutions dans le contenu d'une page Web, nous avons opté pour la construction de grammaires locales, en effet elles nous assurent plus de flexibilité syntaxique, car les locutions sont souvent exprimées avec des variations lexico-syntaxiques minimales qui sont facilement repérables par un processus de Bootsrapping 4 au niveau de la construction et la maintenance des grammaires locales. Pour prendre un exemple, on retrouve souvent sur la page de présentation d'une société les phrases suivantes :

(6.1) a Notre société, leader mondial sur le marché [...]
b Notre société est leader européen dans le secteur [...]
c Notre société est leader sur le marché mondial [...]
d Notre société leader dans son domaine [...]
e Notre société en position de leader au niveau régional [...]

ces locutions peuvent être regroupées dans une grammaire locale qui ressemblerait à celle de la figure 6.5.

Bien que minimalement différents en terme lexico-syntaxiques, ces 5 exemples décrivent un contexte comparable autour du mot *« leader »*. Dans un processus d'extraction automatique de concepts basé sur des méthodes statistiques, on risque d'ignorer de telles entrées car trop

[17] Nous appelons descripteurs locaux, les paramètres juridiques propres aux entreprises et locaux à un pays particulier, comme le numéro de SIRET ou le numéro SIREN

peu fréquentes, bien que très fréquentes à quelques variations lexico-syntaxiques près.

Deux phénomènes linguistiques comparables peuvent être cités à ce niveau, il s'agit de la paraphrase linguistique et de la dérivation sémantique. Les dérivations sémantiques sont des phrases dont le sens est préservé mais dont la structure lexico-syntaxique est différente [16] :
Exemple :

- AOL a acheté Netscape
- l'acquisition de Netscape par AOL

Fuchs décrit dans [26] les paraphrases comme des phrases dont le sens linguistique dénotatif est équivalent. Toujours d'après Fuchs :

Définition 6 *Une phrase ou un texte Y constitue une paraphrase d'une autre phrase ou d'un texte X lorsqu'on considère que Y reformule le contenu de X*

Néanmoins comme l'affirme B. Levrat et T. Amghar, « l'intuition de l'identité sémantique ne suffit pas à faire de deux phrases des paraphrases linguistiques » [40]

Il existe des différences fondamentales dans la définition de la paraphrase et de l'équivalence sémantique dans les nombreux courants linguistiques présents. Certains linguistes maintiennent l'idée que toute différence de forme induit une différence de sens, ils considèrent que tout choix fait au moment de la construction d'un texte a de l'importance, aucune unité de texte ne peut être considérée comme une unité de substitution à une autre. Ils nient ainsi l'existence de l'identité sémantique. L'identité de sens serait d'après ces derniers résultante d'un manque de finesse dans l'analyse sémantique. Un exemple proposé dans [16] en témoigne bien :

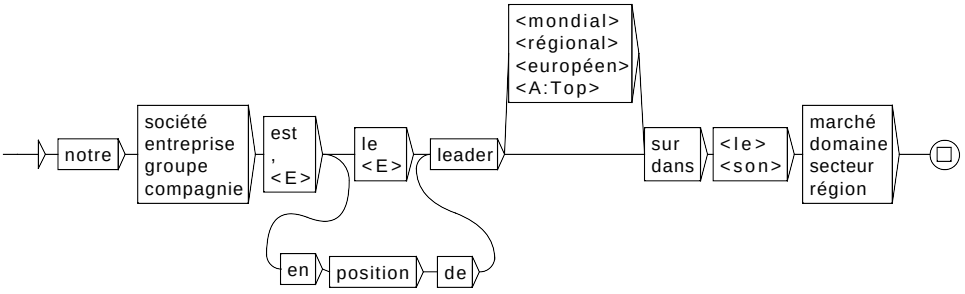


FIG. 6.5 – Grammaire Locale représentative de l'exemple 6.1

(6.2) 1. Jean a fait une tentative pour arrêter de fumer
2. Jean a tenté d'arrêter de fumer
3. Jean a fait deux tentatives pour arrêter de fumer
4. Jean a fait plusieurs tentatives pour arrêter de fumer

A première vue, il est possible de considérer (1) et (2) comme des paraphrases, elles véhiculeraient le même sens commun d'arrêter de fumer, or il est aussi possible de considérer (3) et (4) comme paraphrase de (2). Il est cependant impossible de voir (3) et (4) comme paraphrases de (1), ce qui serait inféré si la première supposition était vraie. Shopen [63] affirme : « je crois qu'il existe très peu (si même il en existe) des couples de phrases ayant une forme grammaticale différente et qui soient sémantiquement équivalentes, c'est-à-dire qui aient la même structure sémantique ».

Face à cette négation de l'identité sémantique entre phrases par cette école, certains linguistes comme Mel'cuk, considèrent qu'il s'agit d'une affaire d'approximation liée à l'imprécision inévitable de tout instrument de mesure, « Les distinctions sémantiques dont il est question peuvent être trouvées si l'on fait très attention à ce qu'on dit, mais de façon générale, les locuteurs ne font pas très attention à ce qu'ils disent, ils ignorent tout simplement beaucoup de distinctions latentes. Dans la pratique quotidienne de la parole, on ne se soucie pas de différences potentielles dont on n'a pas besoin dans un contexte donné » [48], il dit aussi « on ne prend pas une balance trop précise pour peser des légumes dans une épicerie » pourquoi il en serait autrement dans un contexte où l'instrument de mesure serait « l'intuition linguistique du locuteur ».

Mel'cuk affirme en ce qui concerne les équivalences sémantiques « les phrases P1 et P2 sont jugées synonymes si l'ensemble de toutes les inférences que P1 entraîne et l'ensemble de toutes les inférences que P2 entraîne sont identiques » et Fuchs parle de « noyau sémantique commun » nécessaire entre deux phrases pour admettre l'équivalence sémantique même si le sens exact n'est pas conservé. Ce sont ces définitions flexibles de la paraphrase qui sont le plus souvent adoptées dans les systèmes de reconnaissance automatique de paraphrases [19, 16, 20] utiles aux besoins d'extraction d'information ou de systèmes de Question-Réponse en particulier.

Bien que nous soyons également à la recherche de contextes associés à certains mots ou syntagmes qui pourraient jouer le rôle de « noyaux sémantiques communs » décrit par Fuchs, ce qui nous importe est moins l'identité ou l'équivalence sémantique mais plutôt la syntaxe grammaticale et la sémantique contextuelle approximative véhiculée. En effet

nous recherchons à décrire les contextes grammaticaux pouvant entourer les mots et expressions extraites statistiquement de la terminologie des entreprises.

La technique des grammaires locales s'avérant mieux répondre à nos besoins, nous avons opté pour cette technique pour la description contextuelle spécifique des locutions terminologiques types.

Les grammaires locales, décrites au niveau du chapitre 4, peuvent être vues comme des mini-cartes topologiques de la langue [61], elles décrivent des séquences de mots formant des unités sémantiques identifiables. Ces unités y sont également décrites sur un plan morpho-syntaxique. Un des avantages majeurs des grammaires locales est qu'elles permettent aussi bien la description des contextes positifs que la description des contextes négatifs. Il est souvent plus facile d'identifier les exemples négatifs dénombrables que de lister toutes les solutions possibles en évitant d'utiliser des étiquettes grammaticales générales. Supposons que tout Adjectif serait acceptable à part ceux ayant le suffixe *iste*, il est dans ce cas plus adéquat de formuler cette négation, que de lister tous les Adjectifs de la langue ne comportant pas le suffixe *iste*.

Dans notre cas il s'agit de trouver des variations lexico-syntaxiques où le sens n'est pas forcément préservé, comme c'est le cas dans l'exemple 6.1 où *leader mondial* a la même structure syntaxique que *leader régional* mais est sémantiquement différent. Nous avons développé dans cette phase de délimitation du sous langage plus de 15 grammaires locales développant les structures terminologiques suivantes :

- Renseignements juridiques
- Forme juridique
- Capital
- Chiffre d'Affaire
- Description de l'entreprise
- Bienvenue sur le site de la société
- Xxx SA et son équipe vous proposent leurs services
- Copywrite, ©
- Numéro de SIRET, Numéro SIREN
- Rejoindre notre équipe
- Siège social, Téléphone et Téléfax
- notre équipe se met à votre disposition
- notre offre couvre
- nous sommes leader dans le domaine...

Nous proposons la grammaire locale d'extraction du numéro de SIRET au niveau du graphe 6.6. Nous avons incorporé, dans cette grammaire 6.6, les différents numéros d'identification officiels d'une entreprise, tel que le numéro SIREN, le code ape/naf, le numéro de Cenil (une liste plus complète est disponible au niveau de la grammaire de la figure 6.8).

Ces identificateurs ont été réunis dans une même grammaire locale, car les contextes d'apparition dans les PAE sont similaires. La figure 6.7 et la figure 6.8 représentent les deux sous-grammaires désignées dans le graphe principal (fig-6.6). La figure (fig. 6.9) affiche un échantillon des concordances [18] de la grammaire du numéro de Siret.

6.3.2.1 Collecte des descripteurs juridiques

Grâce au développement d'Internet et l'évolution toujours croissante de son utilisation, l'activité de croissance des entreprises n'est plus limitée aux clients de la région de localisation de celle-ci comme c'était le cas pour la plupart des PME/PMI il y a 10 ans, en effet les PME/PMI en particulier, se voient clore des contrats avec des clients à des kilomètres de chez eux, les vidéoconférences, en plus d'Internet rendent possible à des petites entreprises de coopérer avec des clients dans le monde entier, il devient alors possible à des PMEs de développer des coopérations internationales. Ce qui en découle que les pages d'accueil des entreprises se trouvent traduites en plusieurs langues et entre autres en français. Bien que les entreprises traduisent leurs sites Web en des langues diverses, leurs dénominations sociales restent le plus souvent dans la langue d'origine ; ainsi la société *Penta4U GmbH*, ne traduira pas sa forme juridique *GmbH* par *SARL* bien que ce soit l'équivalent français pour la forme de sociétés allemandes *GmbH*. En effet les descripteurs juridiques font souvent parti du nom des sociétés

[18] Une concordance est obtenue en appliquant un transducteur sur un texte : il s'agit de la liste des séquences de mots repérées par ce transducteur et présentées une par ligne.

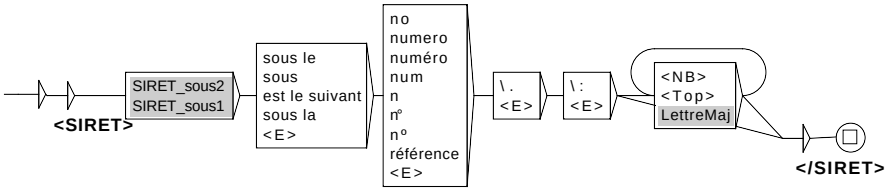


FIG. 6.6 – Grammaire Locale pour l'extraction du numéro de SIRET

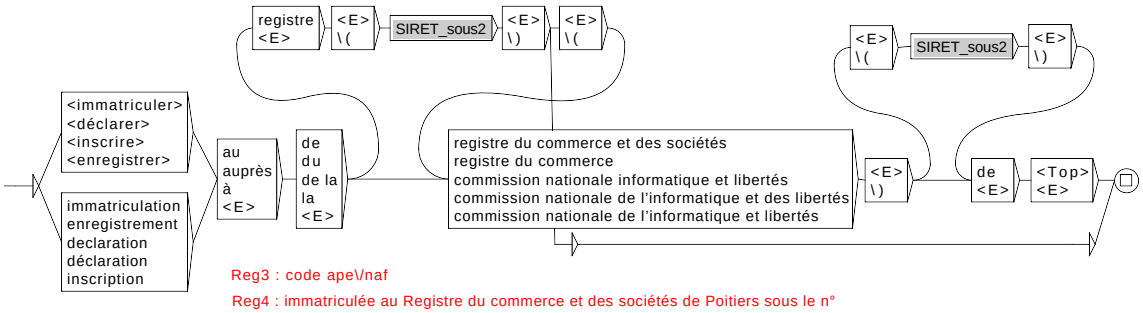


FIG. 6.7 – Sous-Graphe : Siret-sous1

et restent invariants.

Les descripteurs juridiques sont des indices très important pour la reconnaissance automatique des entités nommées et particulièrement la reconnaissance des noms des organisations, ils sont d'ailleurs utilisés dans la plupart des systèmes d'extraction automatique des entités nommées (EN) Le descripteur juridique suit la plupart du temps le nom de la société, mais il peut également se trouver dans la première position. Les formes juridiques représentent, également dans notre système de reconnaissance automatique de Pages Web d'entreprises, un descripteur très puissant car seule une personne juridique (entreprise) enregistrée peut faire suivre ou précéder sa dénomination par sa forme officielles telle que SARL, SA, SAS,etc....

Une liste exhaustive des formes juridiques des entreprises dans le monde est disponible sur le site Web de *Wright Inverstors' Service*[19]. Cette liste dans laquelle nous avons récupéré aussi bien les noms complets des formes que leurs diminutifs et acronymes a le mérite de ne pas nécessiter de maintenance, puisque les formes juridiques des sociétés n'évoluent pas très vite comme c'est le cas des entreprises.

Nous avons transformé les éléments extraient en un dictionnaire électronique respectant le format des DELA [20], dans lequel la forme étendue du descripteur juridique représente les formes de bases et leurs acronymes, les formes dérivées. Un échantillon du dictionnaire des formes juridiques qui contient dans les 400 entrées est proposé par la figure 6.10. Nous avons gardé les formes étendues dans les langue d'origine pour les raisons expliquées ci-dessus.

[19] http : //www.corporateinformation.com/defext.asp

[20] Format DELA de dictionnaires électroniques détaillé au chapitre 5

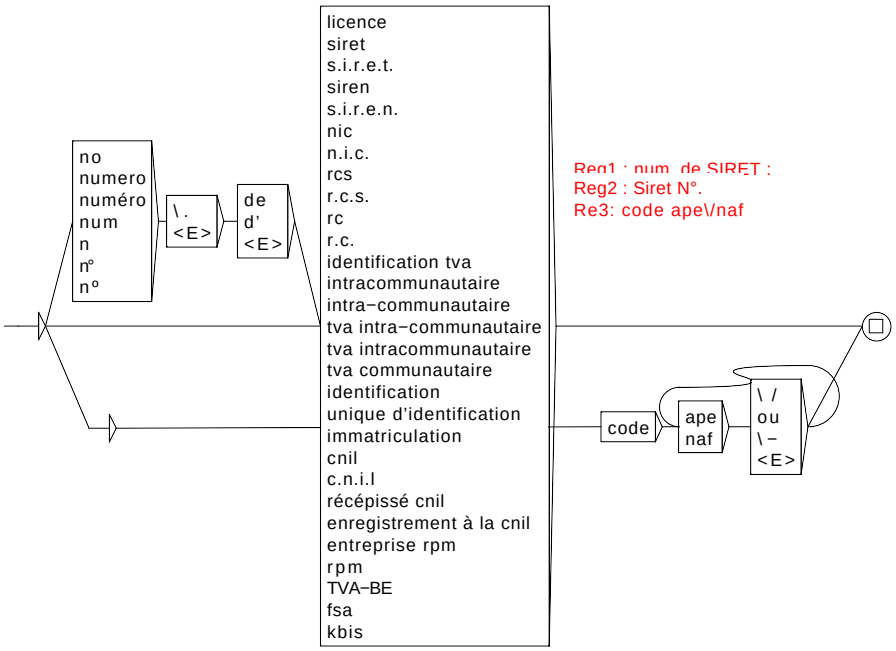


FIG. 6.8 – Sous-Graphe : Siret-sous2

6.4 Reconnaissance automatique du nom de la compagnie

Afin de maximiser les chances de notre système RecPAE de classification automatique de sites Web d'entreprises, nous tentons, en plus des descripteurs déjà décrits dans ce chapitre, d'extraire le nom de l'organisation. Ce descripteur s'est avéré avoir un pouvoir discriminant assez puissant dans la prise de décision finale du système.

L'extraction du nom des organisations, fût très souvent étudiée dans les systèmes d'extraction automatique d'entités nommées (voir chapitre 4). Déjà en 1995, durant la première conférence MUC6 [21], la tâche d'extraction automatique de la catégorie *ENAMEX* des noms propres et sigles fût définie. ENAMEX proposait alors une catégorisation grossière de trois sortes d'entités, à savoir les *Organisations*[22], les *Personnes*[23] et les *Noms de Lieux*[24].

[21] http ://cs.nyu.edu/cs/faculty/grishman/muc6.html

[22] comprend les noms des sociétés, gouvernements et autres entités organisationnelles

[23] comprend les noms de personnes ou les noms de familles, comme dans *les Kennedy*

[24] comprend les noms de villes, départements, régions internationales,hydronymes, montagnes

CNIL n°1139545 ...

Code APE/NA 923 B
Code APE/NAF 153 F

déclaré à la CNIL sous le numéro : 856553
enregistré à la CNIL sous le n° 1133602
Enregistrée au REGISTRE DU COMMERCE ET DES SOCIETES de Pointe à Pitre sous le numéro 433 423 720
enregistrée au Registre du Commerce et des Sociétés de Pontoise.

immatriculée SIRET : 435 059 696 00011
inscrite au registre du commerce RCS Bayonne: 433 926 904
intracommunautaire: 25 404 548 075

RCS : B447531369 ...
RCS : Beaune 410264527

Siret : 48264647800016
N° siret : 399 304 310 00028
SIRET : 490 551 876 00017

FIG. 6.9 – Les Concordances : Numéro de Siret

AE,Anonymos Etairia.FJ
AG,Aktiengesellschaft.FJ
société à responsabilité limitée,.FJ
Sarl,société à responsabilité limitée.FJ
SA,sociedad anónima.FJ
SA,Sociedade por Ações.FJ
SA,Societá in accomandita per azioni.FJ
SA,Societate pe actiuni.FJ
sa,Société Anonyme.FJ
S\/A,Société Anonyme.FJ
S\.A\.,Société Anonyme.FJ
SA,Société Anonyme.FJ
GbR,Gesellschaft burgerlichen Rechts.FJ
Gesellschaft mit beschränkter Haftung,.FJ
GesmbH,Gesellschaft mit beschränkter
Haftun.FJ
Gewone Commanditaire Vennootschap,.FJ
GIE,Groupement d'intéret économique.FJ
GmbH,Gesellschaft mit beschränkter Haftung.FJ

FIG. 6.10 – Échantillon du dictionnaire des Formes Juridiques(FJ) en format DELAC

Mallchok [61] a même consacré ses travaux de thèse à l'extraction exclusive des noms d'organisation dans les dépêches journalistiques de langue anglaise. Certains de ces systèmes ont atteint des résultats expérimentaux de haute performance, cependant il s'agit le plus souvent de méthodes fortement dépendantes de la langue de rédaction des textes ainsi que de leurs domaines d'interactions. En effet dans la rubrique financière du journal *Le Monde*, on trouvera un article parlant de la cession de Novartis à Nestlé de sa division aliments médicaux, alors que sur la page Web de Novartis on trouvera une phrase du genre : « Novartis est leader mondial dans le domaine... » . F.Mallchok [61] se consacrait aux dépêches journalistiques de langue anglaise, N.Friburger [25], elle, aux dépêches journalistiques de langue française et nous nous consacrons à un corpus de pages d'accueil d'entreprises. Bien que les systèmes d'extraction d'information basés sur des méthodes statistiques, développées pour la session « Language-Indépendant Named Entity Recognition » de la conférence CoNLL03[25], successeur de MUC, aient été développés pour être des systèmes indépendants des domaines et des langues et bien que certains comme ceux présentés par [23] et par [10] aient atteint des rappels et précisions supérieurs à 88%, ils restent très dépendants des corpus d'apprentissage.

Le développement du système RecPAE ne représente pas une finalité en soit , il est développé afin d'améliorer notre système de remplissage de base de données thématiques d'offres d'emploi. Il est donc nécessaire pour RecPAE, qu'il puisse classifier toute sorte d'URLs, indépendamment du domaine traité.

Nous avons mis au point deux méthodes distinctes pour l'extraction des noms des entreprises, d'une part nous utilisons le phénomène linguistique des grammaires locales, pour décrire les contextes d'apparition de ces derniers et d'autre part, nous avons développé un algorithme de segmentation et de reconnaissance de mots à partir de composés distribués dans le nom de domaine des pages Web traitées.

6.4.1 Extraction du nom de l'organisation : Grammaires locales

Plusieurs études ont été menées dans le domaine de l'extraction automatique des noms d'organisations à base de grammaires locales, citons entre autres les travaux de N. Friburger [25, 18] et F. Mallchok [61]. Cette méthode bien qu'elle permette une couverture exhaustive de la distribution locale d'un objet, elle a l'inconvénient d'être spécifique aux sous-langages décrits. Les grammaires locales comme l'indique leurs

[25] http : //www.cnts.ua.ac.be/conll2003/

noms permettent de décrire les contraintes locales de certaines unités. Elles permettent de diminuer l'ambiguïté en enrichissant les objets par leurs contextes distributionnels. Elles suivent le principe du plus grand contexte : « élargir le contexte pour lever l'ambiguïté »Ėn effet, il s'agit tout d'abord d'identifier les termes amorces, déclencheur d'un patron sémantico-syntaxique d'extraction et qui peut être développé par Bootstarping [62] pour décrire exhaustivement le contexte de l'amorce dans le corpus étudié. Dans la pseudo-phrase *elle réalise un chiffre d'affaire..*, *chiffre d'affaire* représente les termes amorces car porteur de l'information, alors que *réalise* ne représente ici qu'un verbe support, ce qui peut changer dans un contexte différent.

[14]identifiait déjà en 1996, en présentant un outil de reconnaissance et de classification des noms propres, les notions de preuves internes et de preuves externes. Deux notions importantes dans le domaine de l'extraction automatique d'information. Nous en donnons de plus amples détails dans le chapitre 4.

Les preuves internes se trouvant à l'intérieur même du nom propre, sont invariantes aux domaines d'application. Nous les détaillons au niveau de l'extraction du nom de la compagnie proposant l'offre d'emploi dans le corpus des offres d'emploi qui représente la plus grande partie de notre travail.

Nous nous sommes donc concentrés pendant cette phase sur le recensement et la description des preuves externes associées aux noms des organisations dans un contexte de pages d'accueil d'entreprises. Comme pour toute analyse strictement locale, nous avons agi par Bootstarpping [62] pour enrichir les contextes distributionnels des noms d'organisation dans le corpus des PAE.

Pour éviter d'alourdir notre système, nous avons également tenté de cerner les endroits possibles d'apparition du nom de l'entreprise sur son site Web. La tâche d'extraction nécessitant un prétraitement assez coûteux en temps, nous essayons de filtrer les passages, les pages où la recherche aura lieu.

Une analyse du corpus des PAE, analyse aussi bien textuelle que structurelle, nous a permis de constater que le nom de l'entreprise a de fortes chances d'apparaître dans les 5 positions suivantes :

- sur la page d'accueil entourée d'étiquettes de formatage spéciaux tel que : gras, italique, plus grand.
- sur les pages internes du type « notre société, nous connaître qui sommes nous »[26]

[26]les pages internes référencées par les textes d'ancres classés dans la catégorie « about_company »

- dans le cadre du document au niveau du titre

En effet si le nom de l'entreprise est encapsulé dans une image de format gif, jpeg ou autre, nous ne sommes malheureusement pas capable de l'extraire.

6.4.2 Extraction du nom de l'organisation : Algorithme de segmentation et de reconnaissance de mots

Pendant l'étude de cas traitée au début du présent chapitre, nous avons mentionné que les internautes sont capables de reconnaître le nom d'une organisation en observant sa page Web pendant un laps de temps très court, 10 secondes, temps maximum donné par un internaute à une page d'accueil pour le convaincre d'y rester.

En effet, ergonomiquement, il est conseillé aux concepteurs de sites entreprises, bien qu'il n'existe pas de conventions strictes, de respecter certaines règles, comme par exemple le fait de positionner le Logo ou le nom de l'entreprise dans le coin supérieur gauche de la page d'accueil, suivi d'une description concise de l'activité principale de celle-ci. En tant qu'être humain, avant même de lire le contenu de la page, nous sommes capables de trouver une ressemblance entre le logo ou le nom de l'organisation et le nom de domaine de celle-ci.

Le nom de domaine[27] en revanche se tient de respecter de nombreuses conditions restrictives au niveau du jeux de caractères. C'est pourquoi il n'est pas toujours possible de choisir un nom de domaine identique à celui de la société. De plus le nom de domaine souhaité et qui correspondrait parfaitement au nom de l'organisation peut être déjà attribué à une tierce personne. Dans un tel cas, nous avons pu observé deux situations : La première est de changer le nom de la société en fonction des domaines encore disponibles, c'est la solution la moins populaire, la seconde solution est d'essayer de composer un acronyme, un raccourci ou un « composé » à partir du nom de la société. Nous mettons « composé » entre guillemets car il ne s'agit souvent pas de mots composés correctement formés, ce sont souvent des raccourcis de chaque mot composant le nom de la société et qui grâce à une juxtaposition un composé non lexicalement connu par les dictionnaires d'une langue. Les concepteurs essayent sur le plan syntaxique, de rester le plus prêt possible du nom de leur entreprise et ce pour des raisons qu'il n'est plus nécessaire de préciser.

Il est facile, pour nous, de déduire que l'entreprise enregistrée sous

[27] Le nom de domaine est l'adressage d'une machine sur Internet géré par les serveurs dits Domain Name Server (DNS).

le domaine www.3i.fr correspond à Information Intelligence Interface s'il est écrit dans un autre format que le reste du texte de la page Web. Cependant ceci s'avère une tâche plus ardue, à un système automatique.

Notre heuristique principale est que le nom de domaine est en étroit rapport avec le nom de l'entreprise. Nous avons réussi à répertorier un nombre fini de modèles appliqués lors du choix du nom de domaine.

Ce qui nous intéresse est moins d'extraire le nom de l'entreprise que de pouvoir reconnaître une correspondance entre le nom de domaine et son nom. Bien que de nos jours, chacun peut acheter le nom de domaine qui lui convient et qui soit encore disponible et qui puisse correspondre à son nom de famille par exemple. Une correspondance entre le nom de domaine et celui du nom de l'entreprise reste un indice très puissant quand il est accompagné de certains des autres indices décrits plus haut.

Le nom de domaine est donc très souvent une variation du nom de l'entreprise. Ce premier ne pouvant pas contenir de blancs ni de caractères spéciaux, ceux ci sont alors généralement remplacés par les deux caractères « -, _ ». Les différents cas répertoriés sont :

- Le nom de l'entreprise contient des **blancs**, des **points** : ils sont simplement supprimés et/ou remplacé par « - ou _ ou . ».
  - *Venus en Mars* ⇒ *http ://www.**venusenmars**.fr*
  - *France Télécom* ⇒ *http ://www.**agence.francetelecom**.com*
  - *Von Roll Isola France S.A.* ⇒ *http ://www.**vonroll-isola**.com(*)*
- Le nom de domaine est une partie du nom de l'entreprise
  - *C2A Informatique* ⇒ *http ://www.**c2a**.com*
- Le nom de l'entreprise contient une **préposition**, un **article** et/ou une **apostrophe** : ils sont supprimés ou alors remplacés par « - ou _ ou . »
  - *L'Oreal* ⇒ *http ://www.**loreal**.fr*
- Le nom de l'entreprise contient des **caractères accentués** : ils sont remplacés par leur équivalent ASCII.
  - *Société Générale* ⇒ *http ://www.**societegenerale**.fr*
- Le nom de l'entreprise est très long : le nom de domaine est un acronyme(**).
  - *Société Générale* ⇒ *http ://www.**socgen**.com*

(*) Ce cas montre qu'il n'est pas nécessaire de choisir une unique fonction de transformation, mais qu'il est possible de les associer.

(**) Ce cas est un peu plus compliqué à résoudre car il n'y a pas de lignes claires pour la formation des acronymes et abréviations. Leur création est assez subjective dépendante du goût et du besoin de l'auteur.

(6.3) Pour Le nom du groupe *Société Générale*, ils auraient bien pu choisir le nom de domaine *SG* ou alors *socgen* comme dans l'exemple ou encore *socgenerale* , *societeg* , *societgener* . Il y en a en tout $|societe| * |general| \Rightarrow 7 * 8 = 54$ combinaisons possibles pour la formation d'un acronyme à partir du nom propre *Société Générale* et ce en supposant que seuls les préfixes sont permis.

En se basant sur ces variations possibles, nous avons développé un Algorithme de Segmentation et de Reconnaissance de Mots afin de reconnaître une corrélation entre le nom de domaine et celui de l'entreprise.

6.4.2.1 Problème de segmentation et de reconnaissance de mots dans les langues orientales

Dans les langues occidentales comme le français, les phrases sont séparées par des marqueurs de fin de phrases, tel le point, le point d'interrogation, le point d'exclamation. Les mots de la phrase, eux, sont séparés par les marques de ponctuation. Ces langues ne sont donc pas directement concernées par les méthodes de segmentation de séquences de caractères et d'identification automatique de mots. La situation est cependant différente quand on observe les documents électroniques disponibles sur le Web ; on y rencontre très souvent des mots écrits les uns après les autres sans aucun séparateurs. Cela rappelle le cas des langues orientales comme le chinois.

En chinois, une phrase est composée d'une chaîne contigüe de caractères n'incluant aucun des séparateurs connus dans les langues occidentales. L'absence de limites entre les mots pose de grands problèmes pour l'extraction automatique d'informations. De nombreuses recherches se sont intéressées à ce phénomène de reconnaissance automatique des frontières des mots dans les langues orientales, [72] propose un état de l'art assez étendu pour la segmentation automatique de textes chinois.

Les approches développées jusqu'ici dans ce domaine peuvent être divisées en deux grands courants : Les approches statistiques et les approches à base de règles, il y a naturellement aussi l'essor des approches hybrides.

L'idée de base de toutes les approches statistiques ([21, 66, 68, 17]) est la reconnaissance de suites de caractères montrant une grande affinité entre eux. En anglais par exemple il est statistiquement plus

fréquent de trouver la collocation des deux lettres « th » que la collocation « td » . En disposant d'un corpus d'apprentissage de grande taille, il est possible de construire un modèle statistique de langage qui peut être utilisé pour reconnaître les séquences à segmenter dans la chaîne. Le séparateur se situera alors au niveau des caractères de moindre affinité. Il y a deux paramètres à prendre en considération dans un modèle statistique, à savoir le nombre de caractères consécutifs à examiner et la méthode statistique utilisée pour la construction du modèle. Supposons que nous disposons de 26 lettres. Prenons un caractère, il existe 26 caractères possibles qui puissent venir se placer derrière lui. Ainsi pour une séquence de trois caractères il existe 676 caractères successifs possibles. Certains étant plus probables que d'autres. $t+h$ est plus probable que $t+d$ et $th+e$ est plus probable que $th+v$ [17].

Pour le chinois ces facteurs sont beaucoup plus grand, il existe initialement non pas 26 caractères mais 6.000 caractères, ce qui permet de déduire que pour une séquence de trois caractères il y aurait 36 millions de possibilités à chaque entrée. La méthode statistique la plus répandue est celle des modèles de Markov cachés. Dans un contexte de segmentation, le modèle de Markov décrit la probabilité qu'un caractère à la position p+1 sur un mot dépendamment du caractère à la position p. L'ordre du modèle décrit le nombre des états précédents à chaque position.

L'inconvénient des approches statistiques est que la performance des systèmes est fortement dépendante du modèle de langage construit au niveau de la phase d'apprentissage, ce phénomène est connu en anglais sous « data-sparseness problem » . En effet le sous langage utilisé dans les dépêches économiques d'un journal est fortement différent du sous langage de romans d'amours ou de discours politiques c'est pourquoi le modèle appris sur un corpus n'est pas forcément compatible avec un autre corpus plus spécifique.

Dans les méthodes à base de règles, dites aussi les méthodes à dictionnaires, il s'agit de trouver des correspondances entre les chaînes lues à partir du texte et les entrées du dictionnaire. Commençant à gauche et se déplaçant vers la droite, il s'agit de détecter les mots les plus longs connus du dictionnaire et ce jusqu'à la fin de la phrase. Cette technique est connue comme étant la technique de concordance-avant maximale, elle se base sur l'hypothèse qu'un mot long est plus spécifique et probablement plus correcte dans un contexte donné.

Alternativement, l'algorithme de recherche de concordance maximale peut commencer à droite et se déplacer vers la gauche (commencer à la fin de la phrase et avancer vers le début de celle-ci), il est alors appelé l'algorithme de correspondance-arrière maximale. Cette

technique bien que très répandue, ne s'applique pas en générale sans un ensemble de règles de résolution d'ambiguïté, d'une part car les dictionnaires quelque soient leurs tailles, ne peuvent être exhaustifs et d'autre part car il peut y avoir des cas où plusieurs possibilités de découpages s'avèrent possibles mais que dans le contexte donné il soit nécessaire de faire un choix unique. Cet algorithme a été adapté par [8] en y ajoutant 6 heuristiques afin de résoudre l'ambiguïté.

Bien qu'il existe une similarité entre ce problème de segmentation et d'identification de mots dans les langues orientales et la segmentation des composés dans les documents Web, le problème demeure tout de même fondamentalement différent. Ceci car la composition dans les documents Web dépend fortement de l'auteur et non pas de la validité des différents mots par un dictionnaire. En effet chaque domaine peut avoir sa règle de formation propre en acronyme et abréviations.

6.4.2.2 L'approche basique pour la segmentation et la reconnaissance de mots dans le Web

Nous nous sommes inspirés des approches à base de règles pour la segmentation et la reconnaissance des mots dans les langues orientales pour répondre à notre besoin de recherche de corrélations entre le nom de domaine et une séquence de caractères dans ce que nous avons appelé « cadre du nom ».

Les travaux antérieurs dans le domaine de segmentation dans les langues orientales ont très vite reconnu les limites de l'utilisation des dictionnaires pour la délimitation des mots dans les séquences de caractères, ils ont pour cela formulé un certain nombre de règles d'ordre morpho-syntaxique pour lever l'ambiguïté dans les cas où les dictionnaires permettraient plusieurs segmentations possibles mais que une seule serait grammaticalement et sémantiquement correcte.

Nous rencontrons également ce problème dans notre cas de segmentation des composés retrouvés dans les pages Web. Un deuxième problème s'ajoute dans ce contexte et qui est propre aux pages Web : C'est la segmentation de composés formés à partir d'affixes de différents mots. Ces collocations s'avèrent très courantes dans un contexte comme le notre, cependant ces collocations ne suivent aucune règle grammaticale, morphologique ou autres, elles sont formées à la guise de l'auteur.

Nous avons déjà analysé une instance typique de ce genre de composés, à savoir le cas de *www.**socge**.fr*. Ce dernier est formé par le préfixe *soc* du mot *société* et par le préfixe *gen* correspondant à l'adjectif *générale*. Ceux-ci ont été combinés pour former le mot *socgen* n'existant dans aucun dictionnaire commun ou spécialisé du français mais qui

porte une information précieuse pour notre système de reconnaissance de pages d'accueil d'entreprises.

D'après cet exemple, analysé en détail au niveau de l'exemple 6.3 et dans lequel on considère que seules les combinaisons à base de préfixes sont permises, on aurait 6 préfixes possibles pour le premier mot et 7 pour le second mot, en plus des deux mots initiaux, cela nous ferait déjà un total de 56 combinaisons. Or ce genre de restriction ne peint aucune règle de formation de ces composés trouvés sur les pages Web. L'exemple *www.edutainment.com* montre bien qu'il est également fréquent de trouver des collocations à base de préfixes, de suffixes et d'affixes extraits à partir des mots de la séquence initiale dans le même composé. En effet dans *www.edutainment.com* le préfixe correspond au préfixe du mot *Education* et la seconde syllabe du mot correspond non pas au préfixe mais au suffixe du second mot de la séquence à savoir *Entertainment*.

Nous pouvons en déduire qu'il est impossible qu'un dictionnaire puisse couvrir la totalité des suffixes et des préfixes ainsi que leurs diverses combinaisons potentielles. C'est pourquoi nous avons développé une autre approche de segmentation et de reconnaissance qui puisse reconnaître les mots d'origines formant le composé inconnu.

6.4.2.3 Chi et Ding [9]

[9] se sont intéressés à la reconnaissance des mots composés inconnus trouvés dans le *Framework d'une page Web*. Ils identifiaient alors le Framework d'un document comme étant l'ensemble des liens, des textes d'ancres et du titre d'une page Web. Ils considèrent que la compréhension d'une page Web se fait par le biais de son Framework. Leur démarche est fortement inspirée par les approches de segmentation dans les langues orientales et plus particulièrement par les travaux de [8] qui identifiait 6 heuristiques supplémentaires pour la levée d'ambiguïté dans un système d'identification de mots dans des textes en Mandarin chinois.

La démarche adoptée est la suivante : Dans un premier temps il s'agit des éléments constituant le framework. Après l'élimination des mots vides à partir d'une liste prédéfinie, chaque token restant sera recherché dans le dictionnaire. Si le mot est inconnu du dictionnaire utilisé il sera considéré comme étant un « composé » potentiel et sera envoyé au processus de segmentation. Les deux étapes de concordance -avant et concordance -arrière sont alors appliquées pour obtenir un ensemble de composants candidats. Ce n'est qu'alors que l'heuristique du maximum sera appliquée pour résoudre l'ambiguïté des candidats retenus. Ils ajoutent 6 heuristiques supplémentaires qui permettent de

prendre une décision dans des cas particuliers d'ambiguïté. Ces opérations se répètent tant qu'il existe des token inconnus non encore traités. Les composés pour lesquels aucune segmentation n'a pu être possible seront traités manuellement.

6.4.2.4 Notre démarche

La démarche que nous adoptons est différente de celle de [9], d'une part nous ne disposons pas de dictionnaire et d'autre part nous ne partons pas du texte pour y détecter les mots inconnus . Ce qui nous intéresse est plus de trouver une corrélation, une ressemblance entre le nom de domaine et une séquence de chaînes de caractères dans le site Web. Nous n'avons pas besoin de comprendre la séquence en question, car les noms propres non listés dans les dictionnaires ne portent pas en générale une sémantique déductible de leurs composants. Mis à part l'exemple traitant le nom du groupe *société générale* qui est effectivement composé de mots identifiables, la plupart des noms des organisations sont formés à partir de mots inconnus, c'est pourquoi au lieu d'user d'un dictionnaire, nous extrayons toutes les chaînes de caractères de la page Web traitée et en constituons une *liste de référence* dans laquelle notre algorithme essayera de reconnaître les mots d'origines constituant le composé à segmenter.

Définition 7 *La liste de référence comprend toutes les unités lexicales du document traité. Les unités lexicales sont reconnues selon les marques de ponctuation, les apostrophes, les espaces, les retours à la ligne mais aussi le changement de type, une suite de caractères alphabétiques et numériques sera découpé au niveau du changement de type de caractères.*

Ayant déjà identifié les endroits potentiels d'apparition du nom de l'organisation sur son site Web, cela nous permet de restreindre le champ d'application de l'algorithme de segmentation et de reconnaissance.

Étant donné un composé, la question qui se pose est de savoir *comment est ce que les composants initiaux peuvent-ils être reconnus ?*

Ceci n'est pas une tâche triviale du fait qu'il puisse d'une part exister plusieurs mots reconnus dans la liste à une position donnée du composé et d'autre part les composés ne suivent pas une logique de constitution restrictive séquentielle. La seule sécurité est qu'il s'agit d'une juxtaposition entre plusieurs affixes de chaînes de caractères.

Nous ne proposons pas une approche mixte de concordance maximale pour résoudre ce problème, nous nous contentons d'appliquer l'al-

gorithme de concordance maximale-avant. *avant* car la recherche se fait à partir de la première position du composé à segmenter.

Étant donné un composé, il existe différentes manières de le segmenter en des sous chaînes de caractères. Une heuristique que nous adoptons est de favoriser la concordance la plus longue. Cette heuristique a été vérifiée empiriquement dans différentes approches de segmentation de textes en langue chinoise. Nous admettons également le fait qu'une composition de nom de domaine se fait dans le respect de l'ordre d'apparition des unités lexicales d'origines. Cette supposition n'est pas toujours vérifiée et particulièrement dans le cas des traductions des pages Web des organisations. Prenons l'exemple de l'*Organisation des Nations Unies* connue en français sous l'acronyme *ONU*. Cette organisation est connue dans le reste du monde comme étant *UNO* pour *United Nations Organisation*. Le domaine correspondant à cette dernière est http ://www.un-org.org. Sur la traduction française de la page qui admet le même nom de domaine, aucune similarité ne pourra être détectée avec nom de l'organisation, car le nom est également traduit et l'ordre de construction de l'acronyme ne suit donc plus la restriction que nous avons admise.

Le principe de base de la concordance maximale-avant est d'explorer séquentiellement la chaîne de caractère reçue en entrée à partir de la première position et de comparer le préfixe au fur et à mesure de l'avancement avec les mots du dictionnaire, dans notre cas nous comparons les préfixes avec les mots de la liste intégrant les mots de la page Web traitée.

S'il existe plus qu'une entrée possible dans la liste de référence correspondant à ce préfixe, le mot le plus long sera privilégié. Après l'application du scanneur avant, nous disposons en sortie d'un ensemble de mots candidats pour la segmentation de la chaîne. Certaines heuristiques sont alors utilisées pour sélectionner les candidats les plus probables. Il est important à noter que la correspondance avec la base de référence se fait au fur et à mesure de la reconnaissance et s'arrête dès qu'une quasi-adéquation a été trouvée.

6.4.2.5 Terminologie et définitions

Chi-Hung et Chen Ding [9] définissaient dans leur approche de segmentation de composés trouvés dans le framework d'une page Web la notion de *base de référence* d'une chaîne de caractères. Nous reprenons cette notion en l'adaptant à nos besoins et en présentons la définition suivante :

Définition 8 *La base de référence d'une chaîne C est l'ensemble des*

chaînes auxquelles on applique les 5 opérations $O_1..O_5$ suivantes :

O(C) :

substituer les caractères accentués par leurs équivalents ASCII

mettre toute la séquence en minuscule

supprimer tout caractère de ponctuation

O_1 : *concaténer tous les mots de la chaîne C.*

O_2 : *si la séquence initiale contient des apostrophes :*

supprimer tous les mots vides en préservant ceux associés aux apostrophes.

concaténer la chaîne restante

O_3 : *concaténer tous les mots pleins de la chaîne C.*

O_4 : *concaténer toutes les premières lettres de la chaîne C.*

O_5 : *concaténer toutes les premières lettres de tous les mots pleins de la chaîne C.*

O_6 : *si la séquence contient des chiffres :*

concaténer toutes les premières lettres de tous les mots pleins en gardant aussi les chiffres.

Prenons la séquence C = *« L'Oréal, groupe international n°1 dans le domaine cosmétique. »*

$O(C) \Rightarrow$ l'oreal groupe international n 1 dans le domaine cosmetique

$O_1(C) \Rightarrow$ lorealgroupeinternationaln1dansledomainecosmetique

$O_2(C) \Rightarrow$ lorealgroupeinternationaldomainecosmetique

$O_3(C) \Rightarrow$ orealgroupeinternationaldomainecosmetique

$O_4(C) \Rightarrow$ login1dldc

$O_5(C) \Rightarrow$ ogidc

$O_6(C) \Rightarrow$ ogi1dc

Le but de notre démarche est de trouver une ressemblance entre le nom de domaine et le nom de l'entreprise d'une page d'accueil. Or étant donné que nous ne connaissons pas le nom de l'entreprise et qu'à ce niveau nous ne savons même pas s'il s'agit d'un site Web d'entreprise ou pas, nous appliquons notre algorithme sur le *cadre du nom* qui représente les endroits potentiels d'apparition du nom de l'entreprise dans son site Web. L'algorithme de segmentation et de reconnaissance de mots n'étant pas un algorithme indépendant du système de reconnaissance de pages d'accueil d'entreprises, il profite des prétraitements de celui-ci pour augmenter ses performances. Il est donc appliqué dans les trois situations suivantes :

- ◇ Nous nous trouvons sur la page d'accueil du site Web : La recherche de correspondance entre le nom de domaine et le nom de l'entreprise se fait au niveau du titre de la page.
- ◇ Nous avons trouvé une adresse postale (siège social) grâce aux grammaires locales, nous isolons alors les 5 lignes précédentes et les 5 lignes suivantes de l'adresse. Il est très courant de précéder le siège social par le nom de la société, nous élargissons la fenêtre de recherche car nous avons remarqué pendant la phase d'apprentissage qu'il pouvait y avoir d'autres informations entre le nom de la compagnie et son adresse postale comme le nom du responsable, le nom du service à contacter etc...
- ◇ Nous appliquons également cet algorithme sur la page Web liée par un texte d'ancres appartenant à la catégorie "Société". Si le texte d'ancres menant à la page Web traitée est du type *qui sommes-nous ?* on s'attend à une description de la société et si le nom de la société n'a pu être détecté par les grammaires locales, nous appliquons l'algorithme de segmentation du nom de domaine 'la version strippée de cette page.

Tout au long de notre étude, nous avons établi un certain nombre d'heuristiques de segmentation automatique, pour éviter de passer par l'algorithme de concordance si cela n'était pas nécessaire.

- ◇ Si le nom de domaine ne contient pas de séparateurs $\in$ { _ - .} (*)
  1. rechercher le nom de domaine dans la *liste de référence*

  Si pas de correspondance

  2. rechercher le nom de domaine dans le *cadre du nom*

  Si pas de correspondance

  3. appliquer l'algorithme de segmentation et d'identification des mots au nom de domaine
- ◇ Si le nom de domaine contient des séparateurs
  1. segmenter la chaîne au niveau des séparateurs (**)
  2. pour chaque chaîne de caractère obtenue appeler la condition

  Si une des chaînes a été trouvée dans le *cadre du nom*

  3. sélectionner une fenêtre de longueur *nombre de chaînes de caractère obtenues + 10* autour du mot trouvé.
- ◇ Si le nom de domaine contient une chaîne de caractères contigüe contenant des caractères numériques
  1. segmenter la chaîne au niveau du changement de type de caractères

2. pour chaque chaîne de caractère obtenue appeler la condition

Notre algorithme de segmentation et d'identification de similarités, bien qu'il soit inspiré des approches à base de dictionnaires dites aussi à base de règles, ne fait pas l'usage de dictionnaires, il se sert à la place d'une *liste de référence*[28] variant en fonction de la page Web traitée. En effet nous ne cherchons pas à comprendre les pages Web en essayant de comprendre les mots inconnus de celle-ci, nous cherchons plutôt à établir une relation entre le nom de domaine qui représente pour nous l'inconnu et une chaîne de caractère similaire se trouvant dans la page Web.

Description détaillée de l'algorithme de segmentation et d'identification de similarités

L'algorithme prend trois paramètres en entrée : le mot inconnu, la *liste de référence* et la *base de référence*. On commence par scanner le mot inconnu m de gauche à droite à partir de la première position, à chaque étape on fait correspondre la sous-chaine sc lue avec la *liste de référence* à la recherche d'unités lexicales commençant par la sous-chaine sc en question, tant qu'on trouve des mots dans la *liste de référence* commençant par la sous-chaine, l'exploration séquentielle avant continue en limitant la recherche à la sous-liste de mots potentiels établie à l'étape précédente.

Dans le cas où la lecture d'un caractère supplémentaire engendre une correspondance nulle avec la liste de référence, on recule d'une position et on repère cette position comme étant un point de séparation. Le préfixe est alors remplacé dans le mot inconnu m au fur et à mesure par chacune des unités lexicales candidates retenues. Après chaque substitution, on compare la nouvelle chaîne avec les entrées de la *base de référence*, aussitôt une concordance trouvée l'algorithme s'arrête avec succés. Si par contre aucune correspondance avec les entrées de la base de référence n'a été trouvée, le segment-préfixe ainsi que ses candidats sont stockés dans une table de hachage et le segment-préfixe est éliminé du mot composé initial. Ces étapes sont répétées jusqu'à atteindre la fin du mot inconnu reçu en entrée ou alors avoir trouvé une concordance dans le document représenté par la base de référence.

Si la chaîne en entrée a pu être segmentée mais les juxtapositions candidates n'ont pu mener à des concordances avec la base de référence : cas d'exploration intégrale du mot inconnu, on recherche alors

[28]la liste de référence d'une page Web englobe la liste de toutes les unités lexicales de celle-ci, à compter aussi les unités apparaissant dans le *cadre du document* cachées à l'internaute

Algorithm 1 Algorithme de Segmentation et d'Identification de Similarités

Input: m : mot composé $\wedge$ LR : Liste de référence $\wedge$ BR : Base de référence

Output: booléen : *true* $\leftarrow$ *ressemblance détectée*, *false* $\leftarrow$ *ressemblance non trouvée*

```
1: position-depart ← 0 ;
2: position-courante ← 1 ;
3: while position-courante != fin(m) do
4:   sous-chaine ← m[position-depart ..position-courante]
5:   liste-concordance ← chercher(sous-chaine,LR) ;
6:   if liste-concordance != vide then
7:     position-courante += 1 ;
8:     liste-candidats ← liste-concordance ;
9:   else
10:    if liste-concordance = vide then
11:      if position-courante !=1 then
12:        position-courante ← position-courante-1 ;
13:        sous-chaine ← m[position-depart..position-courante] ;
14:        i ← 0
15:        while i < taille(candidats) do
16:          n-motcomp ← m ;
17:          n-motcomp ← substituer(n-motcomp,sous-
             chaine,candidats[i]) ;
18:          trouvee ← chercher(n-motcomp, BR) ;
19:          if trouvee = 1 then
20:            return true ;
21:          else
22:            i ++ ;
23:          end if
24:        end while
25:        m ← supprimer(m, sous-chaine) ;
26:        table-hash{sous-chaine} ← liste-candidats
27:      end if
28:    end if
29:  end if
30: end while
31: for i = 0 to taille(table − hash − 1) do
32:   position-fin-sequence ← -1 ;
33:   position-fin-sequence ← chercher(table-hash[i], BR) ;
34:   if position − fin − sequence! = −1 then
35:     for j = i + 1 to taille(table − hash1) do
36:       fenetre ← selectionne (BR , position-fin-sequence , 10 +
          taille(table − hash[j])) ;
37:       if chercher(table − hash[j], fenetre) then
38:         return true ;
39:       else
40:         j ++ ;
41:       end if
42:     end for
43:   else
44:     i ++ ;
45:   end if
```

les différents segments dans le document et on examine leurs contextes droits. Si le segment apparaît dans le document, représenté ici par la base de référence, on isole alors son contexte droit. Le contexte droit du segment est exprimé par la fenêtre de 10 caractères venant directement à droite du segment trouvé. Dans cette fenêtre nous cherchons à identifier le segment successeur listé dans la table de hachage résultant de la première phase de segmentation. Ceci nous permet de reconnaître une similarité de la séquence cherchée à 10 caractères près. C'est à dire si le nom de domaine est constitué de certaines unités lexicales du nom et non pas de toutes, on est capable de détecter une similarité à quelques unités près.

Étant donné que nous cherchons à trouver une similarité entre le nom de domaine et le nom de l'entreprise, ou tout simplement le nom de domaine et une chaine de caractères dans le cadre du nom d'une page Web, nous définissons la similarité tel que nous avons déjà montré les différentes fonctions de transformation usuelle pour la constitution du nom de domaine. Il n'est donc pas nécessaire que tous les mots du nom de l'entreprise soient présents dans le nom de domaine. Ceci suit la logique que le mot composé peut avoir été constitué d'une partie du nom de l'entreprise et pas forcément à partir de toutes les unités.

exemple applicatif

Soit l'exemple applicatif suivant :

URL : *http : //www.**socgen**.fr*

mot composé inconnu : *socgen*

cadre du nom[29] : *Banque Société Générale : services financiers, gestion d'actifs, banque d'investissement, siège social*

liste de référence : *actifs, banque, d, financiers, gestion, générale, investissement, services, siège, social, societe,...*

base de référence :

o1.banquesocietegeneraleservicesfinanciersgestiondactifsbanquedinvestissementsiegesocial

o2.banquesocietegeneraleservicesfinanciersgestiondactifsbanquedinvestissementsiegesocial

o3.banquesocietegeneraleservicesfinanciersgestionactifsbanqueinvestissementsiegesocial

o4. bsgsfgdabdiss

o5. bsgsfgabiss

Les étapes appliquées à *socgen* sont :

Le pointeur initial pointe sur la première position du mot à segmenter.

[29]le cadre du nom représenté ici se restreint au titre de la page d'accueil

1. ***s*** est lue et cherchée dans la liste de référence. Il existe plusieurs entrées commençant par ce préfixe, le pointeur avance d'une position.
2. ***so*** est lue et cherchée dans la sous-liste de la liste de référence. Il existe plusieurs entrées commençant par ce préfixe, le pointeur avance d'une position.
3. ***soc*** est lue et cherchée dans la sous-liste de la liste de référence. Il existe plusieurs entrées commençant par ce préfixe, le pointeur avance d'une position.
4. ***socg*** est lue et cherchée dans la sous-liste de la liste de référence. Il n'existe pas de mots commençant par ce préfixe. Le pointeur recule d'une position. Cette position est reconnu comme point de séparation. Les mots candidats de liste de référence sont retenus : (social, societe)
5. ***socialgen*** Le premier mot candidat « *social* » vient substituer le segment reconnu dans le mot inconnu à segmenter. La nouvelle chaîne « *socialgen* » est cherchée dans les entrées de la base de référence. Aucune concordance n'est trouvée. Le mot inconnu initial est rétabli et continue avec les autres mots candidats.
6. ***societegen*** Le second mot candidat « *societe* » vient substituer le segment identifié « *soc* » . La nouvelle chaîne « *societegen* » est cherchée dans les entrées de la base de référence. Une concordance est trouvée au niveau de *o1. ,o2. et o3.*. Une similarité à été établie, l'algorithme s'arrête et retourne *vrai*.

Dans cet exemple, une correspondance a été trouvée avec la juxtaposition du candidat et du reste du mot inconnu initial, l'algorithme a donc fini avec succès sans avoir eu besoin de segmenter la totalité du mot inconnu.

6.5 Reconnaissance automatique des pages d'accueil d'entreprises : L'algorithme

Il a été question tout au long de ce chapitre de détailler les divers processus nécessaires à notre système de reconnaissance automatique de pages d'accueil d'entreprises. Nous proposons dans cette section une vue d'ensemble de notre algorithme et montrons comment ces processus interagissent pour répondre à notre besoin de reconnaissance automatiquement de pages d'accueil d'entreprises.

Le but de notre démarche étant de constituer une base de données d'URLs d'entreprises pouvant être automatiquement générée et

automatiquement extensible. Cette base de données nous permettra de cibler notre recherche des offres d'emploi dans les sites Web des entreprises, au lieu de parcourir dans le Web des hyperliens se référençant les uns les autres sans qu'ils aient de rapport avec les emploi. Nous proposons donc dans cette section une vue globale du système RecPAE et montrons comment nous intégrons les divers processus décrits dans un mécanisme globale de reconnaissance automatique de pages d'accueil d'entreprises.

A la suite de l'étude d'un corpus de PAE de domaines d'activités hétérogènes, nous avons été capable de réunir un ensemble de descripteurs (voir exemple de descripteurs 6.2) spécifiques aux pages Web d'entreprises. Les descripteurs réunis sont sémantiquement chargés du lexique du sous-langage constitué selon la méthode décrite par [57], syntaxiquement correspondant aux grammaires locales identifiées et construites à partir de la phase d'apprentissage sur le corpus. Ceux-ci ont également été enrichis par des paramètres propres à la structure du document Web traité comme nous avons pu le voir tout au long de ce chapitre. Nous attribuons, de plus, à chacun de ces descripteurs un *poids* d'importance, dépendant du *pouvoir discriminant* du descripteur en question. Ces poids ont été attribués manuellement après la phase d'apprentissage.

Les couples $(descripteur_i, poids_i)$ sont par la suite mis sous forme d'une *expression logique*. Un tel système a le mérite de nous délivrer un coefficient numérique facilitant la prise de décision finale. La décision finale est une entité quantitative calculée à partir de la conjonction de tous les paramètres que nous avons identifié.

Définition 9

* *soit n le nombre total de paramètres fixés*
* *soit $p_i \in P$, le paramètre i ; où $P = p_1, p_2, ..., p_n$*
* *soit $w_i \in W$, le poids associé au paramètre p_i ; où $W = w_1, w_2, ..., wn$*

$RecPAE(URL) = w_1.p_1 + w_2.p_2 + ... + w_n.p_n$ *Si* $RecPAE(URL) >= 4 \Rightarrow return1$

$\Rightarrow$ *URL correspond à la page d'accueil d'une Orgnisation*

Sinon

$RecPAE(URL) < 4 \Rightarrow return0$

$\Rightarrow$ *URL ne correspond pas à la page Web d'une Organisation*

L'algorithme prend donc une URL en entrée et rend un booléen en sortie : *OUI* signifie que l'URL correspond à une page Web d'une entreprise, *NON* signifie que ça ne l'est pas.

Algorithm 2 pseudo fonctionnement du système RecPAE

Input: url

Output: booléen : *true* ← *url est une PAE*, *false* ← *url n'est pas une PAE*

while $UrlFile \neq vide$ **do**

1. ExtractSkeleton(url)

Titre

Liens hypertextes : -Liens Internes *LI*, -Liens Externes *LE*, -Emails

Objets référencés : liens et descriptions alternatives ancrées

Formes : noms et liens

Javascript : liens et contenus

2. Syntaxic-SemanticAnalyser

(a). ***UrlAnalyser*** ⇒ analyse sémantique et filtre des liens internes *LI*.

(b). ***AnchorTextAnalyser*** ⇒ analyse et filtre des textes d'ancres *TA* correspondants aux LI

(c). ***CompanyNameExtractor***

1. Grammaires Locales ⇒ extraction du nom de l'entreprise par l'intermédiaire de grammaires locales décrivant les contextes d'apparition dans les PAE

2. ***SegmentRecognitionAlgorithm*** ⇒ algorithme de segmentation et d'identification de mots.(à partir du domaine de la PWE)

(d). ***CompanyInfosExtractor*** ⇒ extractions à base de grammaires locales

Adresse Postale, Numéro de téléphone et de téléfax

Numéro d'enregistrement au Registre national SIRET

locutions types(SARL au capital, ...)

Réponse = DecisionEvaluation($\Sigma poids_i * descripteur_i$)

if Réponse = OUI **then**

return OUI

end if

if exist(LI) **then**

3. ajouter liste LI à la liste URLFile

else

last

end if

end while

4. Réponse = DecisionEvaluation($\Sigma poids_i * descripteur_i$)

return réponse

La 1ère Phase

La première phase de l'algorithme consiste à extraire le squelette de l'URL. L'extraction du squelette réside dans l'identification des différents constituants de la page Web traitée. Nous reconstituons le cadre du document, composé des différents liens internes ainsi que des textes d'ancres correspondants, mais aussi des méta-informations et du titre qui sont des informations cachées à l'internaute mais porteurs d'indicateurs très importants pour la compréhension de la page Web. Le cadre du document comprend également les différents objets multimédias accompagnés de leurs descriptions. Vu l'évolution des techniques du Web, il n'est plus trivial d'extraire le squelette d'une page Web, un analyseur syntaxique traditionnel ne délivre pas une représentation fidèle de celle-ci. On observe de plus en plus de liens cachés dans des scripts Javascript ou bien sous des images qui portent l'information sémantique que nous tentons d'analyser. Nous avons donc pris ces changements en considération lors du développement de notre analyseur et tentons à côté de l'analyse du code source de l'URL, d'analyser également les scripts Javascript référencés, et sommes par conséquent capables d'identifier les liens et leurs textes d'ancres quand ils sont masqués dans un menu Javascript. Dans le cas des images, nous tentons d'exploiter les descriptions associées à travers les attributs *nom* et *alt*. Dans cette phase il a été également important de prendre en considération toute sorte de redirection de l'URL initiale, il arrive que la page principale se contente de montrer une image publicitaire pendant quelques secondes, dans un tel cas, celle-ci ne doit être ignorée, il est seulement important que l'analyseur capture l'hyperlien de redirection. La même situation est appliquée lorsque la première page d'un site Web propose de choisir la langue de lancement du site ou propose un bouton pour entrer dans ce dernier.

La 2ème Phase

Dans une seconde phase, le code source de la page Web est strippé, ce qui revient à récupérer à partir d'une version codée en *Html* le texte de celle-ci sans les objets multimédias ni les liens hypertextes et où tous les caractères spéciaux sont substitués par leurs équivalents Ascii. Avant de supprimer toute trace de formatage du texte, nous sélectionnons et stockons certains passages se trouvant entre des marqueurs Html spéciaux, pour des analyses structurelles ultérieures. Il s'agit par exemple de garder une indication sur les passages écrits en gras, ou ceux exprimant des titres et se trouvant entourés des marqueurs *<h1...n>*.

Une fois le texte pur obtenu, nous en extrayons certaines informa-

tions que l'on trouve habituellement sur la page Web d'une organisation, à savoir son adresse, son numéro de téléphone et de télé-fax, son numéro de SIRET[30] ou de SIREN, Ces extractions s'effectuent par l'application des grammaires locales que nous avons construit à cet effet. Chacune de ces informations est représentée par un descripteur de ceux décrits tout au long de ce chapitre.

La 3ème Phase

La phase 3 du système RecPAE représente la phase la plus intéressante. On y analyse les liens internes et particulièrement les textes d'ancres associés pour reconnaître ceux porteurs d'information utile et qui doivent être explorés. Nous avons détaillé, la classification des textes d'ancres ainsi que des séquences appartenant aux feuilles des URLs dans une section ultérieur dans ce même chapitre, nous n'y reviendrons donc pas.

6.6 Limites

La limite majeure identifiée pendant le développement du système RecPae, est la mode d'utilisation des langages de scripts comme l' ActionScript, lu par les interpréteurs Flash ou les scripts AJAX d'interaction directe avec les internautes. Le nombre de sites Web usant de ces techniques se voit augmenter et nous restons incapable d'interpréter de tels sites web. Cependant, les moteurs de recherche de notoriété importante refusent de se lancer dans l'interprétation de ces scripts et incitent les entreprises qui veulent bénéficier d'un rang raisonnable d'user des techniques classiques de mots clés, de descriptions en mode Html. En vue de cette constatation, nous supposons que la plupart des entreprises désireuses d'avoir de bons rangs dans les moteurs de recherche, continueront encore longtemps à décrire les informations importantes sur leurs sociétés en mode textuel et notre système est alors capable de les analyser convenablement.

[30] Le numéro de SIRET est l'identifiant unique et officiel attribué à chaque entreprise par la chambre du commerce.

CHAPITRE 7

Dictionnaire électronique des Noms de profession

> *« L'élaboration d'un dictionnaire général de la langue exige un travail assidu, poursuivi durant de longues années. Il faut, pour s'y astreindre, une foi persévérante dans l'utilité de l'effort »*
> Paul Robert dans *Introduction au Grand Robert de la Langue Française*

Introduction

Le marché de l'emploi actuel est en mouvement permanent, les individus se voient, de nos jours, obligés d'occuper plusieurs postes tout au long de leur vie professionnelle, il est rare de rencontrer des personnes ayant intégré un poste à l'âge de 20 ans et y être resté jusqu'à la retraite. Le déploiement des technologies internet a permis le partage et la décentralisation des données ainsi que l'ouverture de la scène de l'emploi et du recrutement au grand public. Ce sont les raisons majeures pour le développement, ces dernières années, de structures nouvelles, diversifiées et multiples d'intermédiaires sur le marché de l'emploi : les agences de ressources humaines traditionnelles, les job boards, les agrégateurs d'offres d'emploi, ...

Cependant ces technologies semblent ne pas intégrer de modules de traitement automatique des langues naturelles pour améliorer les performances des services offerts bien qu'il soit primordial de s'y intéresser pour ne pas stagner sur des résultats certes satisfaisants mais insuffisants. Les internautes de plus en plus exigeants et la quantité croissante des documents disponibles rend indispensable le développement de mé-

thodes plus précises et plus pertinentes de compréhension et d'analyse des données textuelles.

Les individus à la recherche d'emplois sont dans un état émotionnel très fragile. Pour aspirer à une notoriété importante, les intermédiaires du marché se doivent d'offrir des services excellents, en effet, un individu dans un tel état d'âme est très vite agacé face à une liste de documents résultats bruitée. Prenons le cas simple d'une personne à la recherche d'un poste de « serveur ». En tant qu'être humain en entendant sa requête se limitant au terme « serveur » dans un contexte de recherche d'emploi, nous sommes capables de comprendre sans ambiguïté qu'il recherche un poste de garçon dans un restaurant, un hôtel, une brasserie ou autre. Cela n'est pas le cas des systèmes de recherche d'information même spécialisés dans les offres d'emploi qui utilisent des techniques statistiques d'indexation basées sur la présence ou l'absence des termes dans les documents ou basées sur la fréquence des termes dans les documents et dans la collection en général.

Face à une telle requête un moteur de recherche d'emploi retourne des annonces de postes de « gestionnaire de serveurs Web » ou encore de « technicien maintenance serveur » ou bien de « développeur SQL-Serveur » , or quelle personne cherchant un poste de « technicien maintenance serveur » ira taper une requête avec uniquement le mot « serveur » ? cette personne aura tendance à enrichir peut être sa requête initiale par le terme serveur, mais ne se contentera pas d'une requête avec ce terme qui précise ainsi l'outil de travail et non la tâche à accomplir.

D'où l'intérêt linguistique et social d'intégrer le lexique et la grammaire des noms de profession pour mieux gérer ce genre d'erreurs autant que la synonymie, l'hyperonyme[1] et l'hyponyme [2].

Certains auteurs comme [52] et [73] se sont déjà penchés sur la question. Le premier étudiait une façon automatique d'enrichir la classe d'objet <Profession> à partir du Web, il avait réuni à peu près 10 000 noms de profession complexes. Le second s'intéressait à un dictionnaire électronique multilingue des noms de profession pour le français, l'espagnol, le catalan et l'arabe.

Il existe également des classifications officielles française et canadienne des professions et métiers qui organisent les métiers en hiérarchie et proposent des noms normalisés pour environ 20 000 emplois. Ces noms sont certes à prendre en considération lors de l'élaboration d'un

[1] « terme générique dont le sens comprend celui d'autres termes plus spécifiques. Animal est l'hyperonyme de mammifère, ce dernier terme étant lui-même hyperonyme de chien. » selon MediaDico

[2] terme spécifique dont le sens est inclus dans celui d'un autre terme plus général : mammifère est hyponyme d'animal. » selon MediaDico

dictionnaire électronique cependant ils présentent deux inconvénients majeurs, le premier est la représentation des données non directement exploitable dans une phase d'extraction d'information car ne respectant pas un format de dictionnaire électronique utilisable pour l'analyse de texte et le second inconvénient, le plus important d'ailleurs est que bien que ces classifications nationales aient essayé de normaliser les appellations d'emploi pour homogénéiser l'information, les recruteurs continuent à user dans leurs annonces de séquences polylexicales complexes différentes de celles présentent dans ces standards officiels. L'intitulé du poste dans une offre d'emploi représente le fragment du texte lu en premier par les candidats et se doit d'être suffisamment expressif pour attirer l'attention des candidats intéressants et uniquement les candidats intéressants. Les recruteurs essaient alors de décrire dans un groupe nominal plus ou moins court les aptitudes du candidat idéal désiré ou les spécificités du poste à pourvoir. Ainsi au lieu de choisir le nom de la profession normalisée proposé par les classifications officielles : ex - « Analyste-programmeur » (comme c'est le cas à l'ANPE), le recruteur aura plutôt tendance à l'enrichir par « Analyste-programmeur J2ee spécialiste bases de données ». D'où pour nous et dans un but d'extraction d'information efficace l'intérêt immédiat de construire un dictionnaire électronique de noms de profession réellement usités.

Nous décrivons dans ce chapitre les différentes étapes de la construction de nos dictionnaires de noms de profession simples et composés contenant un total actuel de 80 000 entrées : La collecte, le nettoyage, l'étude de la typologie, le développement de grammaires locales et l'enrichissement automatique. Nous décrivons ensuite le compromis entre correction et complétude que nous avons adopté pour parer au problème de flexion des noms de profession composés.

7.1 Noms de profession : définition

La classe « Profession » est une classe d'objets renfermant un ensemble d'objets de types noms d'emploi et noms de métiers. Les classes d'objets sont « des classes sémantiques construites à partir de critères syntaxiques » [33, 13]. Ce sont des classes définies par des prédicats définitionnels sémantiquement homogènes de type verbes, adjectifs ou noms auxquels correspondent des domaines d'arguments. Ainsi « Lilas, Marguerite, Rose » sont des arguments de la classe « Fleur » comme « animateur, professeur, aide pâtissier » sont des éléments de la classe « Profession ». La classe « Profession » étudiée par [52], est l'ensemble des noms simples ou complexes répondant essentiellement aux prédi-

cats « gagner sa vie comme » ou « exercer la profession de ». Dans notre langage spécialisé des offres d'emploi, on aura des constructions de prédicats définitionnels du type « poste à pourvoir de » ou « pour renforcer notre équipe nous recherchons ». Tous les prédicats définitionnels aidant à la reconnaissance des noms de profession dans notre système d'extraction d'information sont décrits par des grammaires locales au niveau du chapitre suivant.

La classe « Profession » est constituée de noms simples comme « professeur, gardien, ingénieur » et de noms composés comme « administrateur de production dans le domaine hydropneumatique, opérateur de presse typographique à épreuves, régleur de machines à fabriquer les tiroirs et coulisses de paquets de cigarettes ».

Les noms de profession simples sont d'un nombre restreint, nous en avons répertorié 5000 sous leurs formes lemmatisées et avons procédé à la flexion automatique en adoptant le format de dictionnaire électronique de noms simple DELAS proposé par le LADL et décrit au niveau de l'état de l'art. Le fichier Delas-ProfSimple est un fichier dont chaque entrée représente un nom de profession simple lemmatisé (au masculin singulier pour la plupart) accompagné du graphe de flexion correspondant permettant d'inférer toutes les formes fléchies associées, lesquels on enrichit par des catégories sémantiques utiles dans des phases d'analyses ultérieures. Un tel fichier est alors fourni en paramètre à la procédure de flexion « Inflect » distribuée avec le logiciel Unitex pour donner en sortie le dictionnaire fléchi des noms de profession simples DELAF-Profession. Les listes suivantes présentent un premier extrait du fichier DelasProfSimple.dic suivi de sa forme après flexion.

acidograveur,NN4+Hum+Profession+ProfSimple
aciériste,NN6+Hum+Profession+ProfSimple
aconier,NN5+Hum+Profession+ProfSimple
acteur,NN2+Hum+Profession+ProfSimple
adaptateur,NN2+Hum+Profession+ProfSimple
adjuvant,NN8+Hum+Profession+ProfSimple
aérodynamicien,NN3+Hum+Profession+ProfSimple
aérostier,NN5+Hum+Profession+ProfSimple
...
...

⇓

Unitex : :Inflect
⇓

acidograveur,acidograveur.N+Hum+Profession+ProfSimple :ms
acidograveurs,acidograveur.N+Hum+Profession+ProfSimple :mp
acidograveuse,acidograveur.N+Hum+Profession+ProfSimple :fs
acidograveuses,acidograveur.N+Hum+Profession+ProfSimple :fp
aciériste,aciériste.N+Hum+Profession+ProfSimple :ms :fs
aciéristes,aciériste.N+Hum+Profession+ProfSimple :mp :fp
aconier,aconier.N+Hum+Profession+ProfSimple :ms
aconière,aconier.N+Hum+Profession+ProfSimple :fs
aconières,aconier.N+Hum+Profession+ProfSimple :fp
aconiers,aconier.N+Hum+Profession+ProfSimple :mp
acteur,acteur.N+Hum+Profession+ProfSimple :ms
acteurs,acteur.N+Hum+Profession+ProfSimple :mp
adaptateur,adaptateur.N+Hum+Profession+ProfSimple :ms
adaptateurs,adaptateur.N+Hum+Profession+ProfSimple :mp
adaptatrice,adaptateur.N+Hum+Profession+ProfSimple :fs
adaptatrices,adaptateur.N+Hum+Profession+ProfSimple :fp
adjuvant,adjuvant.N+Hum+Profession+ProfSimple :ms
adjuvante,adjuvant.N+Hum+Profession+ProfSimple :fs
adjuvants,adjuvant.N+Hum+Profession+ProfSimple :mp
adjuvantes,adjuvant.N+Hum+Profession+ProfSimple :fp
aérodynamicien,aérodynamicien.N+Hum+Profession+ProfSimple :ms
aérodynamicienne,aérodynamicien.N+Hum+Profession+ProfSimple :fs
aérodynamiciennes,aérodynamicien.N+Hum+Profession+ProfSimple :fp
aérodynamiciens,aérodynamicien.N+Hum+Profession+ProfSimple :mp
aérostier,aérostier.N+Hum+Profession+ProfSimple :ms
aérostière,aérostier.N+Hum+Profession+ProfSimple :fs
aérostières,aérostier.N+Hum+Profession+ProfSimple :fp
aérostiers,aérostier.N+Hum+Profession+ProfSimple :mp
...
...

Le fichier DELAF-Profession contient au total 11 000 entrées. Ces noms de profession simples ont été récoltés d'une part à partir de listes trouvées sur le net, et d'autre part à partir de l'analyse de plus de 50 000 noms de profession composés dans lesquels nous avons extrait les têtes. Celles-ci fût ensuite vérifiées dans les dictionnaires de la langue française communs. Leurs formes fléchies et surtout leurs formes au féminin suivent les règles de féminisation officielles canadiennes et françaises. Les deux pays n'ont pas toujours adopté les mêmes patrons de féminisation.

L'office québécois de la langue française dira d'une femme ingénieur : « une ingénieure » alors que la Commission générale de terminologie et de néologie usera de la séquence polylexicale « une femme ingénieur ou une ingénieur ». Nous avons opté pour l'introduction dans nos dictionnaires de toutes les formes reconnues officiellement par les autorités spécialisées des deux pays.

L'intérêt de ce chapitre se situe essentiellement dans l'analyse des noms de profession composés qui sont des séquences polylexicales complexes d'un nombre indéfini, du fait que toute combinaison de noms de profession simples avec un secteur d'activité, une spécialité, un produit, un service peut former un nom de profession composé. Notre intérêt premier est la reconnaissance de telles séquences polylexicales dans les offres d'emploi et l'identification correcte des bordures. Une reconnaissance partielle d'un nom de profession peut induire en erreur et viser un publique différent de celui initialement souhaité, soit le texte suivant :

> *Nous recherchons pour notre filiale en France / Paris, un chef de projet spécialisé dans la gestion de qualité (h/f).*

> Les noms de profession reconnus possibles sont :
>
> *chef de projet*
> *chef de projet spécialisé dans la gestion*
> *chef de projet spécialisé dans la gestion de qualité*

Les trois noms de profession cités sont des noms bien construits et aussi plausibles les uns que les autres, cependant le recruteur en précisant la spécialisation du candidat en *gestion de qualité*, a essayé de borner le domaine pour ne viser que les candidats intéressants.
Si notre système ne reconnaît qu'une partie du nom de la profession, il n'est alors plus fidèle à l'annonce initiale car élargissant le champ d'application de l'offre par rapport à l'ambition de l'auteur. D'où l'intérêt de l'analyse de la typologie des noms de profession et du recensement d'un maximum de noms possibles.

7.2 Recensement

Le dictionnaire des noms de profession composés a été constitué à partir de différentes sources telle que *La Classification Nationale des*

Professions du Canada [3] établie en 2001 et mise à jour en 2006 et *le ROME : le Répertoire Opérationnel des Métiers et des Emplois de l'ANPE* ainsi qu'à partir d'extractions automatiques à base de grammaires locales dans des offres d'emploi et d'extractions automatiques à base d'expressions régulières dans des listes d'offres disponibles sur les moteurs de recherche spécialisés à l'emploi.
Bien qu'il y ait des normes et des appellations contrôlées, les entreprises ne les respectent presque pas, hormis celles publiant leurs annonces à l'ANPE contraintes de choisir les titres des postes dans la classification ROME, les appellations d'emploi dans tous les autres job boards comme Monster ou Cadremploi ne sont pas contrôlées. Les recruteurs remplissent manuellement des formulaires où il existe entre autre un champs : « Nom du poste » dans lequel ils sont libres d'inscrire l'intitulé qui leur convient tout en respectant la charte de bonne conduite interdisant les propos racistes ou pornographiques. Du fait que le titre de l'offre est le miroir principal de l'annonce, les recruteurs ont tendance à enrichir les noms d'emploi plus ou moins simples par des informations supplémentaires sur la spécialité de l'entreprise, les outils à maîtriser, les services d'intégration, les compétences désirées, etc. Il est donc insuffisant de compter uniquement sur les listes officielles pour l'extraction des noms de profession dans les offres d'emploi. De là l'intérêt pour nous de construire des grammaires locales repeignant les différents prédicats définitionnels de la classe « Profession » dans le langage spécialisé des offres d'emploi et l'enrichissement de notre dictionnaire des noms de profession composés par des extractions à partir du Web.

7.2.1 Les classifications officielles

Les classifications officielles canadiennes et françaises sont des classifications informationnelles, où chaque appellation d'emploi est classée dans le groupe professionnel correspondant et est accompagnée d'une description détaillée du métier, du type d'entreprise le pratiquant et le type de formation à suivre pour l'exercer. Il est impensable de construire un dictionnaire électronique de noms de profession sans tenir compte de ces classifications officielles, nous en présentons ici même un aperçu.

La Classification Nationale des Professions(CNP) Canada

La Classification nationale des professions (CNP) est une classification qui décrit les professions observées au Canada. Elle définie

[3] http ://www23.hrdc-drhc.gc.ca/2001/f/generic/welcome.shtml

formellement 20.000 appellations d'emplois réparties dans 520 groupes professionnels organisés selon une hiérarchie à trois niveaux.

- 5125 int. a. (interprète agréé/interprète agréée)
- 5125 interprète en American Sign Language (ASL)
- 5125 interprète en ASL (American Sign Language)
- 5125 interprète gestuel/gestuelle devant auditoire
- 5 125 traducteur médical/traductrice médicale
- 5125 traductrice d'émissions étrangères
- 5125 réviseur/réviseure - traduction

Ces exemples ont été choisis dans un même groupe d'appellations d'emplois, celui *5125-traducteurs, terminologues et interprètes* qui est un sous-groupe de la classe *512-Professionnels/professionnelles de la rédaction, de la traduction et des relations publiques*, lui même sous groupe de la classe plus générale *51-Personnel professionnel des arts et de la culture*.

En observant ces exemples nous pouvons déjà voir les différences de formats présentes dans une même classification et ainsi imaginer la difficulté d'automatisation du nettoyage d'une telle liste pour l'obtention de noms composés homogènes. La barre oblique sépare tantôt les noms d'emploi au masculin de leur forme féminine(ex-5) tantôt uniquement les adjectifs du nom composé(ex- 4) et tantôt uniquement la tête. Les formes masculines et féminines des appellations d'emploi sont tantôt présentes dans une même entrée tantôt séparées dans deux entrées distinctes de la liste(ex-6). Pareillement pour la parenthèse qui tantôt devrait être supprimée car contient un acronyme et tantôt contient l'appellation elle même(ex- 1 et ex- 2).

Prenons le cas de l'appellation du nom de métier du dernier exemple : *réviseur/réviseure - traduction*, la question qui se pose à ce niveau est de savoir : quel est le nom de profession composé correspondant à cette entrée de la CNP ? Nous proposons ci-joint six possibilités formulées à partir de ce dernier :

(a) réviseur traduction.
(b) réviseur en traduction.
(c) réviseur en traductions.
(d) réviseur de traduction.
(e) réviseur de traductions.
(f) réviseur des traductions.

Parmi ces 6 propositions *((a)-(f))* laquelle est conforme aux règles de construction des mots composés du français et laquelle est la plus fré-

quemment utilisée. Nous mesurons la fréquence d'utilisation par des requêtes sur les trois moteurs de recherche les plus usités dans le monde à savoir *Google, Yahoo et Msn* et rendons compte du nombre de correspondances exactes trouvées dans leurs documents indexés respectifs. Selon ce principe le mot composé le plus fréquemment disponible est : *(c) réviseur de traduction.* alors que le premier représentant une entrée dans la CNP est très peu fréquent. au vu de ces détails, nous avons filtré et nettoyé la classification nationale des professions du canada manuellement pour la transformer en un dictionnaire électronique de la forme des DELA, nous reviendrons sur ce point ultérieurement dans ce chapitre.

La Classification ROME

Le ROME -Répertoire Opérationnel des Métiers et des Emplois de l'ANPE- propose un peu plus de 10 000 appellations d'emploi distribuées sous 466 fiches emplois/métiers contre 520 dans la classification canadienne. Chacun d'entre eux est identifié par un code de 5 chiffres appelé *code rome*. Chaque fiche renseigne sur la définition exacte d'un métier, sur le type d'entreprises où il s'exerce, sur les compétences qu'il faut posséder pour le pratiquer et sur les formations permettant de l'exercer. La version que nous avons obtenue de la classification ROME est sous la forme de hiérarchie telle que :

- 5369 / monteur en chambres froides et entrepôts frigorifiques / bâtiment, travaux publics et extraction / personnel du second oeuvre / monteur plaquiste en agencements / 42222
- 5594 / mosaïste (poseur) / bâtiment, travaux publics et extraction / personnel du second oeuvre / poseur de revêtements rigides / 42231
- 6462 / mécanicien d'entretien et de maintenance d'engins de chantier et des travaux publics / mécanique, électricité et électronique / personnel d'entretien, maintenance / mécanicien d'engins de chantier, de levage et manutention et de machines agricoles / 44316
- 8748 / moniteur (électricité-électronique) / maîtrise industrielle / agents d'encadrement de fabrication industrielle / agent d'encadrement de production électrique et électronique / 51112

- 9063 / modéliste des industries des cuirs, peaux et matériaux associés / techniciens industriels / techniciens de préparation de la production / modéliste des industries des matériaux souples / 52151
- 9404 / monteur-dépanneur en installations climatiques / techniciens industriels / techniciens d'installation, maintenance / maintenicien des systèmes thermiques, climatiques et frigorifiques / 52332

Les appellations d'emploi dans la liste ci-dessus sont à reconnaître entre la première et la seconde barre oblique de chaque entrée. Elle sont suivies des groupes et des sous groupes de la classification ainsi que du Code Rome associé à la fiche métier de l'entrée observée. Nous proposons dans la liste ci-dessous les noms de profession sans leurs classes d'appartenance car nous ne visons pas à élaborer une ontologie des noms de profession mais d'en recenser un maximum pour améliorer nos tâches d'extraction ultérieures. Les noms des métiers présents dans le ROME sont comme c'est le cas pour la CNP canadienne sous des formes très différentes rendant le nettoyage automatique difficile.

- modéliste des industries des cuirs, peaux et matériaux associés
- moniteur (électricité-électronique)
- monteur en chambres froides et entrepôts frigorifiques
- monteur-dépanneur en installations de froid et climatisation
- mosaïste (poseur)
- mécanicien d'entretien et de maintenance d'engins de chantier

Cette liste de noms de profession plus ou moins complexes nécessite des transformations pour être prise en compte dans un dictionnaire électronique. Ces transformations sont pour la plupart effectuées manuellement, car cette liste comme la précédente ne suit pas une logique uniforme analysable par un programme automatique. Prenons le premier cas par exemple, nous avons décortiqué ce dernier en plusieurs noms de profession composés tel que :

1. modéliste des industries de cuirs de peaux et de matériaux associés
2. modéliste des industries de cuirs

3. modéliste des industries de peaux
4. modéliste des industries de cuirs et de peaux
5. modéliste des industries de cuirs, peaux
6. modéliste des industries de peaux et de cuirs
7. modéliste des industries de peaux, de cuirs et de matériaux associés
8. modéliste des industrie de peaux, cuirs et matériaux associés

Toutes ces combinaisons sont des combinaisons possibles et correctes que nous nous devons d'introduire dans le dictionnaire pour servir au mieux le système d'extraction d'information automatique. Il est cependant impossible d'automatiser de telles transformations de par leur irrégularités, c'est pourquoi nous avons opéré manuellement pour extraire et transformer les appellations d'emploi complexes et toutes les combinaisons qui y apparaissent. Nous décrivons la phase de nettoyage et de transformation dans la section 7.3.

7.2.2 Grammaires locales

À côté des classifications officielles nous avons enrichi notre dictionnaire électronique des noms de profession composés par des extractions automatiques à base de grammaires locales sûres.
Selon la charte de rédaction des annonces d'emploi publiée par l'ANPE, il est par exemple prohibé de favoriser un sexe sur l'autre dans le nom de la profession. Il est proposé aux recruteurs pour éviter des retombées juridiques pour discriminations sexuelles d'ajouter entre autres la mention *(H/F)* pour Homme ou Femme dans les noms des professions. Ce marqueur représente pour nous une preuve externe très puissante. Étant placé généralement à la fin du titre du poste à pourvoir, il représente la borne droite du nom, nous avons alors ajouté une heuristique supplémentaire pour l'identification de la borne gauche du nom.
Nous supposons que tout nom de profession aussi complexe qu'il soit commence toujours par un nom de profession simple parmi ceux que nous avons déjà classifiés ou par un terme comme(aide, vice, ...) que nous avons également recensé et décrit dans des grammaires locales appropriées. Cette heuristique n'est pas toujours vérifiée surtout du fait du non respect des règles grammaticales de construction des composés par les recruteurs ; mais nous considérons que ces cas particuliers

sont des exceptions que nous sommes capables d'élucider par d'autres grammaires locales plus complexes dans la phase de reconnaissance automatique.

La première grammaire d'extraction des noms de profession utilisée pour réunir des noms de profession complexes est présente à la figure 7.2.2. Elle permet de reconnaître les séquences les plus longues commençant par un nom de profession simple et se terminant par le marqueur(H/F) ou l'une de ses formes usitées similaire.

Les noms bruts suivants ont été reconnus par la grammaire 7.2.2 sur un corpus d'offres d'emploi dressé à partir du Web.

- conseiller(ère) en séjours linguistiques HF
- Commercial Export (Langue maternelle anglaise ou équivalent) (m/f)
- Analyste développeur confirmé HR ACCESS (H/F)
- développeur crystal report H/F

Le fait de choisir la séquence la plus longue peut engendrer quelques erreurs surtout au niveau des noms de profession simples polysémiques et le manque de marques de ponctuation dans certaines offres. Ce manque est dû aux balises Html qui permettent une visualisation claire,

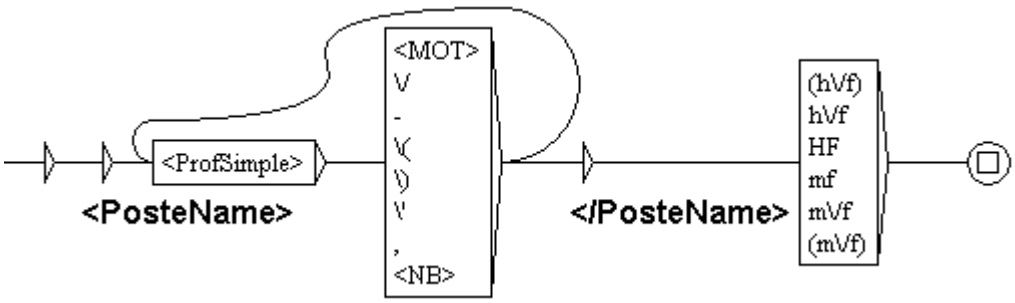


FIG. 7.1 – GL d'extraction de noms de profession complexes à partir d'une offre d'emploi

les auteurs ont tendance à omettre certains signes de ponctuation.
L'observation d'un nombre important d'offres d'emploi nous a appris que le titre de l'offre qui se confond le plus souvent avec l'appellation de l'emploi apparaît très souvent hors contexte, c'est pourquoi ces erreurs sont dérisoires par rapport à la qualité des extractions atteinte par cette grammaire dans un corpus d'offres d'emploi. Voici quelques exemples de fautes de reconnaissance de la grammaire présentée ci-haut et dûe à l'extraction de la séquence la plus longue :

(7.1) ... ,elle recherche dans le <**NomProf**> **cadre de son développement un Ingénieur Pilote** </**NomProf**> H/F

Ici la borne initiale fût reconnue pour être « Cadre » alors que le nom de profession devrait se limiter à « Ingénieur Pilote ». Ce type d'erreurs est géré dans la phase d'extraction automatique par la cascade de transducteurs utilisée, qui favorise la grammaire de

Xxx+ recherche dans le cadre de *yyy+* un/une *zzz+*
nous recherchons pour renforcer notre équipe *yyy+* un /une *zzz+*

(7.2) <**NomProf**> **Industriel de premier plan propose pour renforcer son équipe un poste de cadre études de prix** </**NomProf**>(H/F)

Dans ce deuxième cas, l'erreur également due au choix de la séquence la plus longue est gérée dans la phase d'extraction automatique à base de cascade de transducteurs où le patron d'extraction autour de la locution « proposer un poste » est prioritaire.
Les offres d'emploi sont souvent rédigées en bloc, sous forme de semi-formulaire, nous avons dans ce cas construit des grammaires locales supplémentaires pour l'extraction des noms de profession qui apparaissent suivant les locutions types comme :

poste de *Xxx+*
titre du poste : *Xxx+*
poste à pourvoir : *Xxx+*
titre de l'offre : *Xxx+*
poste offert : *Xxx+*
poste proposé : *Xxx+*

7.2.3 Extraction à partir du Web

Une méthode efficace pour recenser un nombre important de noms de profession composés réels est de parcourir les moteurs de recherche d'emploi disponibles. Ceux-ci disposent d'offres d'emploi structurées et postées manuellement par les recruteurs ce qui implique que les titres des annonces sont des noms d'emploi écrits par les recruteurs eux mêmes. Nous avons écrit des petit robots Web (Crawler) pour parcourir certains job boards et identifier grâce à des expressions régulières spécifiques aux différents moteurs la liste des noms de profession répondant à une série de requêtes.
Les requêtes correspondent tantôt à des noms de profession simples comme « ingénieur » ou « administrateur » et tantôt à des domaines d'activité comme « physique » ou « linguistique ». Le but étant de réunir un maximum d'appellations d'emploi compatibles avec le monde réel.

Cette méthode nous a permis d'assembler un grand nombre de noms composés nouveaux. La liste récupérée fût néanmoins très bruitée. Comme déjà mentionné plus haut les recruteurs ajoutent des informations supplémentaires sur la spécialité, le domaine d'activité ou même sur le lieu de travail ou le type de contrat dans le titre de l'offre afin d'attirer l'attention des candidats. Certaines de ces informations sont utiles à garder et peuvent être considérées comme faisant parti du nom composé, mais d'autres bien que très intéressantes répondent à d'autres catégories informationnelles à extraire dans des phases distinctes. De plus ces appellations d'emploi renferment un grand nombre de fautes d'orthographes, des accents manquants, des lettres permutées, des caractères spéciaux non indispensables ou encore des abréviations. et ne suivent pas une logique de format uniforme. Nous détaillons dans la section suivante la phase de nettoyage et de normalisation. Voici un échantillon des listes des appellations d'emploi brutes extraites auto-

matiquement à partir des moteurs de recherche :

- **orly -** exploitant**s** et assistant**s** d'exploitation
  → pluriel + lieu
- DESSINATEUR MECANIQUE/**TECH** MAINTENANCE (H/F)
  → abréviation
- Monteur en mécanique - **URGENT** (H/F)
  → adjectif additionnel
- **3** Ingénieur**s** Commercia**ux** CAO/Calcul/SGDT SolidWorks (H/F)
  → pluriel
- COMMERCIAL INTERNE H/F – **CDI**
  → type de contrat
- Chef de Projet Conception Système Mécanique **IDF**
  → lieu

7.3 Nettoyage

La phase de nettoyage des noms de profession fût une tâche ardue. Les données ont été récoltées à partir de plusieurs sources différentes et même celles provenant de la même source ne respectent pas un format uniforme qui nous permette d'automatiser la tâche de prétraitement, de nettoyage et de normalisation. Nous avons néanmoins traité les listes d'une même origine ensemble afin de maximiser les transformation automatiques.

La première phase de normalisation est une tâche automatique pour laquelle nous avons construit 3 listes en fonction de 3 types d'erreurs observés.

1. une liste des mots avec accents et leurs correspondances sans accents
2. une liste d'abréviations usuelle dans le domaine des offres d'emploi
3. une liste de préfixes employés avec ou sans trait d'union

La première liste des mots avec accents accompagnés de leurs formes sans accents correspondantes a été construite à partir des classifications officielles ROME et CPN qui sont orthographiquement irréprochables. Nous y avons pioché tous les mots avec accents et les avons mis en correspondance avec leurs formes incorrectes sans accents. Nous avons ensuite effectué un tri dans lequel nous avons éliminé les formes ambigües. Une forme est reconnue pour être ambigüe si sa forme sans accent correspond à un mot correct de la langue française ; « chargé » et « charge » est une entrée ayant été supprimée de la liste, car on ne peut pas remplacer automatiquement la seconde forme par la première si on la rencontre dans la liste des appellations, alors que remplacer « etude » par « étude » est sûrement correcte. Cette liste nous est donc très utile dans la phase de normalisation, dans laquelle les textes en entrée sont mis en correspondance avec la liste des mots sans accents recensés et sont remplacés par leurs formes correctes si besoin est. Cette phase s'avère essentielle pour le prétraitement des appellations d'emploi récupérées à partir des moteurs de recherche, car ceux-ci étant particulièrement bruités, ils ont tendance à être écrit tout en majuscule et sans accents. Les cas ambigües sont alors vérifiés manuellement dans une phase ultérieure. Des heuristiques supplémentaires spécifiques sont ajoutées pour minimiser le traitement manuel. Par exemple pour la cas particulier de « charge » nous avons opté pour l'heuristique qui dit :

Si	le nom de profession commence par « charge »
alors	remplacer « charge » par « chargé »
Sinon	rien faire

La seconde liste renferme un ensemble d'abréviations réunies à partir des différentes listes disponibles sur le Net et à partir du fichier des titres d'emploi initialement réunis. Les abréviations sont très souvent des noms de profession simples ou des domaines d'activités. Cette seconde liste nous sert également dans une phase de normalisation pour obtenir des noms de profession correctes. Un échantillon de cette liste est :

ing.	Ingénieur
comm.	commercial
commerc.	commercial
dev.	développement
nouv techno	nouvelles technologies
...	...

La troisième liste que nous avons construite est la liste des préfixes comme « technico, physio, micro, socio, ... » qui sont des préfixes pouvant être écrits collés aux mots qu'ils forment ou bien en tant que composé avec trait d'union. Dans ce cas nous avons dupliqué les entrées contenant ces préfixes une fois sous la forme collée et une fois sous la forme avec trait d'union (voir exemple suivant).

technico-commercial
technicocommercial
socio-éducatif
socioéducatif
maxillo-faciale
maxillofaciale
...

Afin de maximiser le traitement automatique, nous avons procédé par bloc. Tous les fichiers provenant de la même source ont été rassemblés et traités conjointement. Cela s'est avéré bénéfique surtout pour le traitement des signes de ponctuation comme les virgules qui tantôt devaient être supprimées, tantôt protégées ou les barres obliques qui dans certains fichiers séparent la forme masculine de sa forme féminine et dans d'autres séparent une suite de produits, d'outils ou de domaines.

Nous avons relevé un certain nombre d'irrégularités dans les listes recensées en provenance des classifications officielles et du Web. Des irrégularités que nous avons dû éliminer avant de les introduire dans le dictionnaire électronique des noms de profession composés. Nous récapitulons dans ci-après certains cas parmi ceux que nous avons eu à traiter dans les listes des appellations d'emplois recensés.

- *Les noms sont enrichis par des adjectifs ou des locutions non indispensables* : il faut s'assurer que la forme simple existe également

dans le dictionnaire.

◇ professeur d'histoire **au niveau secondaire**
- professeur d'histoire

◇ agent **professionnel** de magasinage
- agent de magasinage

◇ chef de travaux de formation pour adulte
- chef de travaux de formation

- *Les noms contiennent des conjonctions ou disjonctions de coordination* : il faut permuter les objets pour produire toutes les combinaisons possibles du nom et s'assurer que les noms simples avec uniquement l'un des objets existent déjà.

◇ entrepreneur en formage, façonnage et montage de pièces métalliques
- entrepreneur en formage de pièces métalliques
- entrepreneur en façonnage de pièces métalliques
- entrepreneur en montage de pièces métalliques
- entrepreneur en formage et façonnage de pièces métalliques
- entrepreneur en formage et montage de pièces métalliques
- entrepreneur en façonnage et formage de pièces métalliques
- entrepreneur en façonnage et montage de pièces métalliques
- entrepreneur en montage et formage de pièces métalliques
- entrepreneur en montage et façonnage de pièces métalliques
- entrepreneur en formage, montage et façonnage de pièces métalliques
- entrepreneur en montage, formage et façonnage de pièces métalliques
- entrepreneur en montage, façonnage et formage de pièces métalliques
- entrepreneur en façonnage, formage et montage de pièces métalliques
- entrepreneur en façonnage, montage et formage de pièces métalliques

- *Le nom renferme aussi bien le nom masculin que sa forme féminine* : Le problème à ce niveau est que le nom est tantôt exprimé dans les deux genres et tantôt uniquement certaines parties du mot sont exprimées dans les deux genres et le reste au masculin bien qu'il devrait également être fléchi.

◇ ergothérapeute évaluateur évaluatrice de la capacité du travail
- ergothérapeute évaluateur de la capacité du travail
- ergothérapeute évaluatrice de la capacité du travail

◇ photographe opérateur/opératrice-tireur
- photographe opérateur - tireur
- photographe opératrice - tireuse

◇ contrôleur(euse) financier(ère) en électricité industrielle
- contrôleur financier en électricité industrielle
- contrôleuse financière en électricité industrielle

◇ infirmier immatriculé immatriculée à l'hôpital
- infirmier immatriculé à l'hôpital
- infirmière immatriculée à l'hôpital

- *Les noms à trait d'union* : il est nécessaire de vérifier si les deux mots séparés par un trait d'union sont des noms de métiers simples, chacun d'eux doit alors être introduit dans le dictionnaire des noms de profession simples *Delas-ProfSimple* et chacun d'eux est augmenté par le reste du composé pour introduire les formes de base dans le dictionnaire.

◇ analyste-programmeur en informatique distribuée
- programmeur
- analyste
- programmeur en informatique distribuée
- analyste en informatique distribuée

- *Les composés NDN avec le second N au pluriel ou au singulier* : plusieurs noms composés de cette forme doivent être selon les règles grammaticales du français au singulier ou au pluriel, mais il s'avère qu'ils sont aussi souvent écrits de la première que de la seconde manière dans les offres d'emploi. Nous nous devons de gérer ces cas pour améliorer la performance du système d'extraction. Nous parlons plus en détail de ce cas dans la section sur la flexion automatique des mots composés.

◇ ingénieur d'étude en informatique
- ingénieur d'étude
- ingénieur d'études

◇ accordeur d'instrument de musique
- accordeur d'instrument
- accordeur d'instruments

◇ chef de produit de récupération
- chef de produit
- chef de produits

Après avoir nettoyé la liste des 50.000 noms de profession composés par le processus décrit ci-dessus, nous passons à l'étape de l'analyse de

la typologie des noms recueillis. Cette étape est indispensable pour :

1. La phase de flexion automatique des noms de profession complexes.
2. La phase d'extraction automatique des noms de profession dans les offres d'emploi.

L'étude de la typologie des noms de profession est importante dans la phase de flexion automatique car nous sommes alors capables d'appliquer les règles de flexion du pluriel sur des noms décomposés sous les 7 formes composées de base étudiées par Mathieu-Colas, il décrit dans [44] plus de 700 Patterns de composition de Mots Composés étant tous une composition des 7 formes de base à savoir : *NDN , NN, NA, AN, NAN, VN et NPN*

L'étude de la typologie nous a également permis de classifier les différentes composantes simples ou composées des noms de profession composés dans des catégories sémantiques distinctes, ce qui est très utile dans la phase d'extraction automatique des noms et surtout dans le cas de la non correspondance directe entre les entrées du dictionnaire et les noms disponibles dans le texte. En reconnaissant qu'un nom de profession peut être composé d'un nom de profession simple suivi d'un domaine d'activité (responsable du **service achats et contrats**) ou suivi d'une suite de produits (ingénieur d'étude *Java Jbuilder Swing*) nous sommes alors capables de cerner le nom de profession quand il apparaît dans un texte hors contexte ou avec des contextes externes non restrictifs.

7.4 Typologie des noms de profession composés

Après observation des appellations d'emploi dans le monde réel et après les avoir comparés avec ceux issus des sources officielles que nous décrivons au début de ce chapitre, nous avons été en mesure de conclure que les noms de profession cités dans les offres d'emploi, sont souvent écrits sans tenir compte des règles de composition de la langue française. Nous avons souvent été incapable de reconnaître et d'extraire ces entités dans les textes des offres d'emploi que nous téléchargeons tous

les jours en nous basant exclusivement sur notre dictionnaire. Ainsi et dans un but d'exhaustivité d'extraction, nous nous sommes concentrés sur l'étude de la typologie de ces derniers pour parvenir par la construction de grammaires locales de retrouver les noms de profession même s'ils ne sont pas encore connus par nos dictionnaires.

La conclusion que nous avons pu tirer est que ces entités sont exprimées sous forme de concaténation de plusieurs objets sémantiquement séparables. Pour plus de clarté, illustrons nos dires par un exemple. Dans notre Dictionnaire DNPC (Dictionnaire des Noms de Profession Composés), nous retrouvons le NPC(Nom de Profession Composé) suivant :

◇ gestionnaire de la paie et des ressources humaines

alors que dans les offres d'emploi, les recruteurs ont également tendance à rechercher un :

- gestionnaire paie et ressources humaines
- gestionnaire paie et RH
- gestionnaire de la paie et de RH
- gestionnaire RH et paie
- gestionnaire de RH et de la paie

A partir de ces exemples, il est clair que le nom cité par l'un ou l'autre peut être syntaxiquement différent mais que le message à passer reste dans ces 6 exemples exactement le même. C'est ainsi que nous nous sommes concentrés sur la décomposition des NPC, il nous a été possible de classifier les unités lexicales constituantes simples ou composés dans des classes sémantiques identifiables et dénombrables.

◇ **Produit /Matériel**
  - armoire de cuisine
  - chaussure
  - disjoncteur

◇ **Connaissance**
  - génie civil
  - maçonnerie
  - broderie

- ◇ **Processus**
  - construction de véhicules automobiles
  - conception de logiciel
  - réparation de poches de coulée
- ◇ **Etablissement**
  - école de danse
  - salle de commande centrale
  - gare de fret
  - bureau de poste
- ◇ **ServiceEnt**
  - paie et ressources humaines
  - service de renseignement des comptes
  - division industrie et sous traitance
- ◇ **DomaineEnt**
  - domaine réseau
  - hôtellerie
  - secteur agro-alimentaire
- ◇ **NomSpec**
  - dba oracle
  - sap rw/sp
  - réseau SNA
- ◇ **LangueMonde**
  - français
  - estonien
  - aranais
- ◇ **LangageProgrammation**
  - perl
  - 3D Studio Max
  - weblogic
- ◇ **MatiereEnseignee**
  - linguistique
  - science naturelles
  - anthropologie
- ◇ **SystemeExploitation**
  - linux entreprise
  - Mac OS X
  - Microsoft Windows millenium edition

Nous avons dans une seconde phase construit des grammaires locales de reconnaissance des noms de profession composés que nous décrivons en détail ci-après.

7.4.1 Catégorisation des unités lexicales dans les noms de profession composés

La classification des unités lexicales des noms de profession composés recensés dans le dictionnaire DNPC fût une tâche semi-automatique. Semi-automatique car nous avons écrit des grammaires locales spécifiques, en nous aidant d'Unitex et des transducteurs à états finis pour identifier les différentes catégories grammaticales et sémantiques des constituants des NPC, puis dans une seconde phase d'analyse nous avons parcouru manuellement les listes des concordances afin de réviser les erreurs de classification produites.

Dans la première phase de classification, nous avons pris en compte un certain nombre de critères syntaxiques et morphologiques comme par exemple les mots composés commençant par un mot ayant le suffixe *tion* ont été classifiés dans la classe « Processus ».

- <Processus> transformation du papier </Processus>
- <Processus> exploitation forestière</Processus>
- <Processus> construction mécanique</Processus>

Le composé de la classe « Processus » peut également être décomposable en des sous-unités pouvant appartenir à la classe « Produit » (voir l'exemple suivant).

<Nom de profession> :	manœuvre à la transformation du papier et du carton
<Processus> :	transformation du papier et du carton
<Produit> :	papier
<Produit> :	carton

L'extraction de la forme { Nom-Adjectif / Nom<tion$> tel que Nom se termine par le suffixe *tion* } a produit 420 correspondances dont 24 incorrectes et 17 nécessitant une manipulation supplémentaire. Voici un exemple des fautes observées à ce niveau :

- fonction publique,. $N + NComp + Processus + NA : fs$
- direction financière,. $N + NComp + Processus + NA : fs$
- directeur de la formation agricole,. $N + NComp + Processus + NA : fs$

Tous les composés de la forme { Nom_1-Adjectif ou Nom_1-Det-Nom / Nom_1<ement$> tel que Nom_1 se termine par le suffixe *ement* } ont été classés dans la catégorie sémantique « Produit » (500 sur 1500) et dans la catégorie « Établissement »(1000 sur 1500).

- équipement de transformation du plastique,.$N + NComp + Produit$
- instrument d'arpentage,.$N + NComp + Produit$
- établissement action spéciale,.$N + NComp + Etablissement$
- établissement bancaire,.$N + NComp + Etablissement$

Nous avons procédé sur le même principe avec les suffixes « tien, age, erie »qui nous a permis d'assembler d'une manière efficace et rapide des composés appartenant aux classes sémantiques citées ci-dessus.

- entretien d'aéronefs,.$N + NComp + Processus$
- soutien logistique intégré,.$N + NComp + Processus$
- usinage industriel,.$N + NComp + Connaissance$
- câblage en construction résidentielle,.$N+NComp+Connaissance$
- Pâtisserie,.$N + NComp + Etablissement$
- Pâtisserie,.$N + NComp + Produit$
- blanchisserie,.$N + NComp + Etablissement$

7.4.2 Grammaires locales spéciales pour les NPC

Une analyse détaillée de la composition dans un corpus d'appellation d'emploi, nous a permis de décrire ces derniers dans plusieurs grammaires locales. Comme nous l'avons déjà cité plus haut, notre conception de « la composition » est très large. Nous appelons *Nom Composé*, toute forme nominale non soudée (par non soudé, nous désignons les composés à graphies multiples) et présentant un certain degré de figement(voir chapitre 5). Nous nous sommes donc limités aux composés graphiques, à éléments disjoints ou articulés par un séparateur tel qu'un trait d'union, une apostrophe, une virgule, une barre oblique ou encore une conjonction de coordination « et, ou » en particulier.

Notre description syntaxique se limite aux noms composés commençant par un nom de profession simple en accord avec l'heuristique que nous avons admise antérieurement. Ces noms de profession simples sont recensés dans le Delas des Professions simples DelasProfSimple.dic dont nous montrons un échantillon au niveau de la section 7.1.
Les noms de profession à graphies multiples de l'exemple ci-dessous (7.3) et ne commençant pas par un nom de profession simple du DelasProfSimple, ont soit été inventoriés ou extraits par des descriptions de contextes droits et gauches très détaillées.

(7.3)

- **sage**-femme
- **prud**'homme
- **second**-chef
- **script**-girl
- **seconde** électricienne de plateau

L'hypothèse de base sur laquelle se construit cette description syntaxique est donc : « Un nom d'emploi est reconnu si et seulement si il commence par un nom de profession simple ». Nous avons pu recensé 5000 noms de profession simples, soit 2000 supplémentaires par rapport aux dictionnaires fournis par Unitex.

Comme l'indique le tableau suivant nous avons constaté qu'une appellation d'emploi composée est une association d'un nom de profession simple avec un nom de service d'entreprise, un nom de domaine d'activité, un nom spécial de marque, de produits, d'outils ou un acronyme.

Catégorie Sémantique	Signification
<ProfSimple>	Nom de profession simple : professeur, serveur, éboueur, ...
<ServiceEnt>	Services dans une entreprise : service commerciale, service des Ressources Humaines.
<DomaineEnt>	Domaine d'activité : agro-alimentaire, industrie automobile, ...
<NSpec>	Nom : langage de programmation, systèmes d'exploitations, logiciels spéciaux, techniques désignées par des acronymes ou noms de produits, ...
<ASpec>	Adjectif spécifique : senior, junior,confirmé, débutant, ...

Dans la suite de cette section nous nous concentrons sur la description syntaxique de certaines de ces jonctions pour ainsi montrer la complexité des noms de profession appris.

Les services d'entreprise

Nous avons assemblé une liste de 1350 noms de service dans les organisations. A côté des services classiques comme celui des ressources humaines ou celui financier, nous avons ajouté des services comme celui de l'outillage et des véhicules légers ou celui d'aide aux entreprises. Les noms des services collectés ont été fléchis conformément au format DELACF. Un échantillon de ce dictionnaire « ServiceEnt.dic »est disponible en annexe.

Les services des entreprises viennent souvent enrichir les noms de profession pour en constituer des plus complexes, c'est pourquoi nous nous sommes intéressés aux différentes combinaisons possibles à travers des grammaires locales nous permettant dans la phase d'extraction de retrouver de nouveau ces combinaisons dans le texte d'une offre d'emploi. La première grammaire construite permet d'extraire les concordances du type suivant :

<ProfSimple><ServiceEnt>
comptable service client
responsable ressources humaines

<ProfSimple><DET><ServiceEnt>
administratrice du service des avantages sociaux
assistant du drh

<ProfSimple><PREPDET><ServiceEnt><A>
préposé au service de l'outillage et des véhicules légers

<ProfSimple><ServiceEnt>&<ServiceEnt>
responsable paie & administration du personnel

<ProfSimple><ServiceEnt>et <DET><N><A>
président des ressources humaines et du développement organisationnel

Nom spécial : NSpec

Dans la classe *<NSpec>* nous avons regroupé toutes les unités lexicales simples et composées appartenant aux différentes classes : *« Produit, marque, domaine, langage de programmation, systèmes d'exploitation, noms de logiciels, noms de machines, noms de systèmes de gestion de bases de données... »*. Ces noms sont souvent des acronymes, c'est pourquoi nous avons aussi récolté la liste des noms entiers même s'ils ne sont que rarement utilisés. De même que pour les services de l'entreprise nous avons construit des grammaires locales décrivant les associations possibles des noms spéciaux avec les noms de profession simples pour former des noms de profession composés, dont un échantillon des concordances sur un corpus d'offres d'emploi est présenté ici.

<ProfSimple><NSpec>
programmeur Visual Age Generator (programmeur VGA)
analyste as400

<ProfSimple><NSpec><PREP><DET+Ddef><DomaineEnt>
expert sap dans le domaine logistique

<ProfSimple><NSpec><ASpec><NSpec>
ingénieur be confirmé Autocad (ingénieur bureau d'études confirmé autocad)

<PofSimple>-<ProfSimple><LangProg>
ingénieur-développeur C++

<PofSimple><N><NSpec>*
administrateur systèmes réseaux nt4 2000

Spécialiste ou spécialisé

Les recruteurs utilisent souvent les deux unités lexicales « spécialiste » et « spécialisé » dans les titres des emplois afin d'expliciter les spécificités techniques des candidats désirés ou alors le domaine d'activité particulier du poste à pourvoir. Ainsi dans le nom « ingénieur spécialiste ponts et chaussées » la précision vient enrichir sur le domaine de l'emploi de l'architecture en général, alors que dans le nom « ingénieur spécialisé en Java » on précise plutôt l'outil à maîtriser

par le candidat qui est dans ce cas un langage de programmation. Ces deux unités sont très courantes dans les noms de profession composés et doivent absolument être reconnues comme faisant partie du nom, car un « ingénieur spécialiste des métaux », n'est pas un « ingénieur spécialiste des réseaux sans fils ». Nous avons donc essayé de décrire le plus exhaustivement possible leurs grammaires locales dans un contexte d'offres d'emploi.

<ProfSimple>-spécialiste <PREP|DET><NDN><A>
Forestière-spécialiste en système d'information géographique
Ingénieur-spécialiste de la gestion de procédés alimentaires

<ProfSimple><A> spécialiste<PREP><N><A>
Travailleur social spécialiste en assistance individuelle
Diététicienne diplômée spécialiste en nutrition sportive

<ProfSimple>* spécialiste <DET><N>
Apprenti vitrier-mécanicien spécialiste du métal
Boulanger spécialiste de gâteaux

<ProfSimple> spécialiste <PREP><N>
spécialiste en neurologie
spécialiste en conception assistée par ordinateur

Nous avons traité dans cette catégorie, le nom d'emploi d'un spécialiste, en effet nous ne considérons pas spécialiste comme étant une appellation d'emploi simple valide, car cet objet n'est pas autonome, il a besoin d'un ou plusieurs attributs Ainsi *nous sommes à la recherche* ***d'un spécialiste en médecine du travail*** , **d'un spécialiste dans le domaine de l'agriculture biologique** ou **d'un spécialiste réseaux** et non pas *nous sommes à la recherche* ***d'un spécialiste***.

Nous devons encore décider si le spécialiste est spécialiste d'un domaine, d'un outil ou d'une branche.

A côté du spécialiste recherché, il est très courant de spécifier la spécialisation du candidat recherché, par la séquence « spécialisé en, spécialisé dans ». Ici aussi il est intéressant de distinguer entre la spécialité du domaine ou de l'outil requise avec l'adjectif souvent utilisé.

<ProfSimple> spécialisé <LangProg.Outil>
programmeur spécialisé java

<ProfSimple> spécialisé <PREP><Methode>
consultant spécialisé en risk managment

<ProfSimple> spécialisé <PREP><Technique>
ouvrier spécialisé en réparation de panneaux muraux
fermière spécialisé en culture du blé

<ProfSimple> spécialisé <PREP><Materiel>
ouvrière spécialisée en couvertures
ouvrier spécialisé en charpentes métalliques

7.5 Flexion automatique ou semi-automatique

La flexion automatique des noms composés étudiée rigoureusement par [60] pendant sa thèse est une méthode très coûteuse en terme de pré-traitement manuel. Celle-ci permet, de générer les formes correctes mais pas les formes les plus fréquentes. Nous avons pû observer que le pluriel des noms composés est souvent mal employé dans les documents du Web. Ne sachant pas la forme correcte entre « domaines d'activités » et « domaines d'activité » une personne qui se fierait aux nombres d'occurrences des deux chaînes dans les moteurs de recherche les plus usités, ou alors à la crédibilité des pages Web renfermant l'une des deux séquences, ne pourrait prendre une décision sûre sans se concentrer vraiment sur la qualité des sites proposés dans la liste des documents pertinents. En effet, les séquences apparaissent autant l'une que l'autre dans des sites sérieux avec des fréquences respectives de 596 000 et de 1 180 000 occurrences. Ces deux formes sont néanmoins correctes, car d'un point de vue sémantique, il s'agit d'un domaine composé de plusieurs activités et bien que la tête d'une telle forme (NDN) soit uniquement le premier nom, l'exception permet aussi bien les deux formes. Pour une génération automatique, il est impossible de faire une analyse sémantique des composants, soit on admet la règle d'exception qui fléchit les deux noms dans tous les cas de figures de la forme NDN ou alors on préfère l'hypothèse de la correction stricte et appliquons la règle régulière qui permet de fléchir uniquement le premier nom dans une telle forme composée. Ceci pour insister sur le fait que la génération unique des formes fléchies correctes de nos noms de profession composés, engendrerait des taux de reconnaissances médiocres par notre système d'extraction automatique d'information.

Afin d'obtenir toutes les formes fléchies des noms de profession composés, nous avons développé un programme qui génère toutes les formes possibles à partir d'une séquence en se basant sur un étiquetage syntaxique établi par des graphes Unitex. La liste résultante peut contenir des formes incorrectes, mais nous avons pu montrer tout au long de ce document que bien que les offres d'emploi reflétent l'image de l'entreprise, celles-ci sont très souvent rédigées sans tenir compte des règles grammaticales strictes.

Nous avons ainsi décrit dans plusieurs graphes les formes typologiques couvrant les noms de profession composés de nos listes initiales qui nous permettent d'obtenir à partir de chaque séquence, une séquence syntaxiquement étiquetée équivalente. Le programme de flexion automatique fait appel dans une première phase à l'analyse syntaxique de chaque nom. On obtient souvent en sortie de cette première étape plusieurs étiquetages possibles pour une seule et même séquence, dûe à l'appartenance des composants à plusieurs classes syntaxiques et sémantiques. Dans ce cas, nous pouvons privilégier des règles d'élagages que nous avons écrites, comme par exemple de choisir la séquence étiquetée la plus courte et qui revient à favoriser les sous-composés les plus longs déjà connus de nos dictionnaires ou alors de procéder à un élagage manuel que nous avons facilité par l'élaboration d'une interface utilisateur dans laquelle nous pouvons en quelques clics choisir l'analyse syntaxique correcte parmi toutes celles proposées par l'analyseur.

Une fois les séquences étiquetées, le programme de flexion génère toutes les formes à partir des formes fléchies des composantes simples identifiées et qui sont recensées dans les dictionnaires DELAS distribués entre autres avec Unitex. Un certain nombre de règles ont été mises en place pour assurer une cohérence en genre en particulier. Ainsi pour un composé de la forme « *<N :ms><A :ms>* » on ne génère ni la forme « *<N :fs><A :ms>* » ni la forme « *<N :fs><A :ms>* ».

Nous avons eu recours dans un premier temps à une méthode d'élagage des résultats obtenus en fonction de leurs fréquences d'utilisation rapportée à la fréquence de la forme de base. Celle-ci recherche la fréquence d'apparition de chaque séquence fléchie proposée par l'algorithme de flexion dans les documents des 3 moteurs de recherche universels les plus usités. Si la fréquence de cette dernière est au moins supérieur à 50% de la fréquence d'apparition de sa forme de base dans ces mêmes moteurs de recherche, alors la séquence est retenue sinon

elle est rejetée. Un tel élagage ne nous garantie pas une bonne qualité de résultats et handicape les phases de maintenance des grammaires locales décrites ultérieurement, car le Web contenant beaucoup de documents textuels non contrôlés comme les forums et les blogs, nous sommes passibles de supprimer de la liste des formes fléchies certaines séquences correctes car trop peu fréquentes et en garder en contre partie des fausses car assez fréquemment utilisées. La question d'utilisation du Web comme corpus d'apprentissage ou corpus de test dans les recherches scientifiques a été soulevée à mainte reprise dans la littérature et dépendamment de la communauté, cette technique est soit acceptée, soit tolérée, soit totalement refusée. Nous avons, en ce qui nous concerne, opté pour la solution de garder toutes formes générées même si elles sont linguistiquement incorrectes ou statistiquement peu fréquentes.

Nous proposons dans la figure (7.5), la capture d'écran de l'interface d'élagage des formes syntaxiques proposées par la première phase d'analyse du programme de flexion automatique des noms de profession composés, que nous utilisons par ailleurs pour augmenter le dictionnaire terminologique des composés fréquents et des mots inconnus extraits à partir des offres d'emploi.

Conclusion

Les dictionnaires élaborés des noms de profession simples et composés sont utiles dans la phase d'analyse et de changement de l'espace de représentation des offres d'emploi que nous détaillons au chapitre suivant. Le dictionnaire des noms composés est augmenté en permanence tout au long de la phase d'extraction d'information du fait des grammaires locales construites et qui permettent de reconnaître des appellations d'emploi présentes dans les offres d'emploi et pas encore classifiées connues.

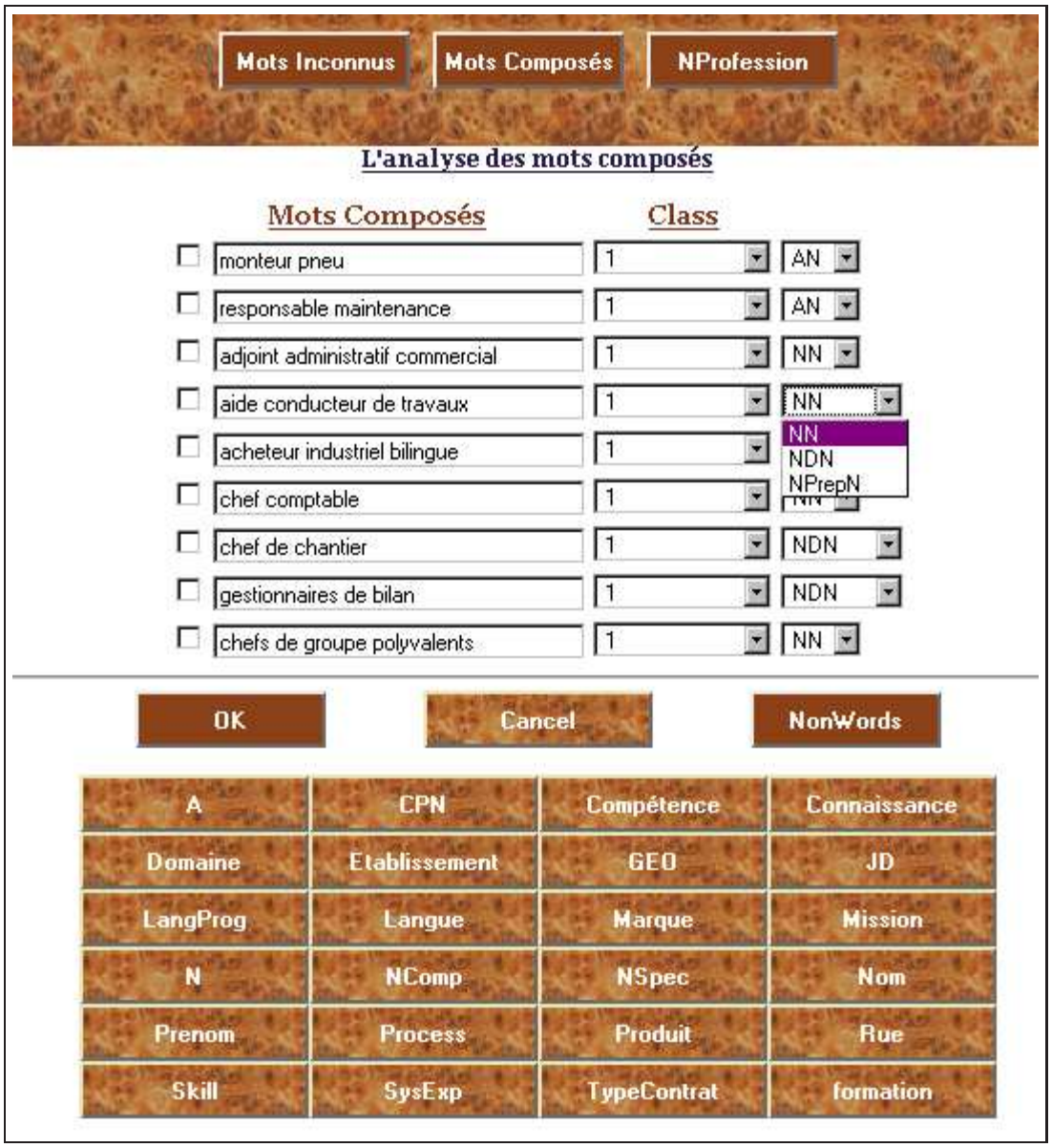


FIG. 7.2 – Capture d'écran : Interface d'élagage et de classification

CHAPITRE 8

Extraction d'information dans les offres d'emploi

Introduction

Ce chapitre représente l'élément clé de nos travaux. Il s'agit de la phase d'extraction d'information et de transformation automatique de l'espace de représentation des documents initialement en plein texte sous une forme plus structurée, chargée d'une sémantique interprétable par un système de recherche d'information. Nous nous sommes concentrés à ce niveau sur la construction d'un nombre important de grammaires locales couvrant au mieux les informations supports qui nous permettent de remplir automatiquement notre base d'offres d'emploi. La structure de la base de données peut être vue comme un formulaire permettant de structurer l'information présente dans chaque offre (*«l'intitulé du poste», «le nom de l'entreprise», «la date d'embauche», « l'expérience minimum requise »...*). Avant de nous pencher en détail sur les grammaires locales élaborées, nous présentons dans la section suivante une vue globale du fonctionnement du second sous-système de transformation et montrons son interaction avec les différents modules étudiés dans ce manuscrit et qui ont tous pour objectif de servir notre moteur de recherche d'emploi.

8.1 Transformation de l'espace de représentation

Notre système, comme déjà mentionné au niveau du chapitre 1, est constitué de 3 phases principales. Nous avons introduit dans le chapitre précédent, le module de collecte de documents qui représente la première phase et nous nous concentrons dans ce chapitre sur le second module d'extraction d'information et de transformation de l'espace de représentation des documents reçus de la première phase. La première phase de collecte de documents fournit une liste d'URLs d'offres d'emploi potentielles qui doivent être analysées et classifiées par la seconde phase étudiée tout au long de ce chapitre, avant de pouvoir intégrer la base de données. Le but de cette seconde phase est donc de lever l'ambiguïté occasionnée par la première classification et de transformer l'espace de représentation des offres d'emploi écrites initialement en langage naturel dans une représentation plus structurée. Afin d'aboutir à une telle transformation, nous appliquons différents modules de traitements aux pages Web reçues en entrée et dont nous présentons une vue d'ensemble dans la figure (fig. 8.1).

Comme énoncé dans la figure (fig. 8.1), notre système d'extraction, appelé *TransOE*, fait appel à 4 phases de traitement que nous résumons dans ce qui suit :

Pré-traitement

- Le contenu Html brut des sites Web pré-classifiés sont reçus en entrée. Chaque document est étiqueté par les balises structurelles sémantiques *«[TagMISSION]», [TagPROFIL]», [TagFORMATION]»,etc* correspondant aux 13 classes de structures retrouvées dans une offre d'emploi bien formatée et que nous avons présenté en détail au niveau du chapitre sur la description du projet, où nous avons d'ailleurs montré que certaines offres présentent peu ou pas du tout de structure sectionnelle. Dans un tel cas d'absence de structure, le marquage n'est pas possible.
- Une fois la structure de l'offre identifiée et marquée, le document est libéré de toutes les balises HTML et des différents marqueurs des langages de scripts et des langages de programmation utilisés. On obtient en sortie un document purement textuel où seules les balises structurelles sémantiques introduites à l'étape précédente sont retenues.

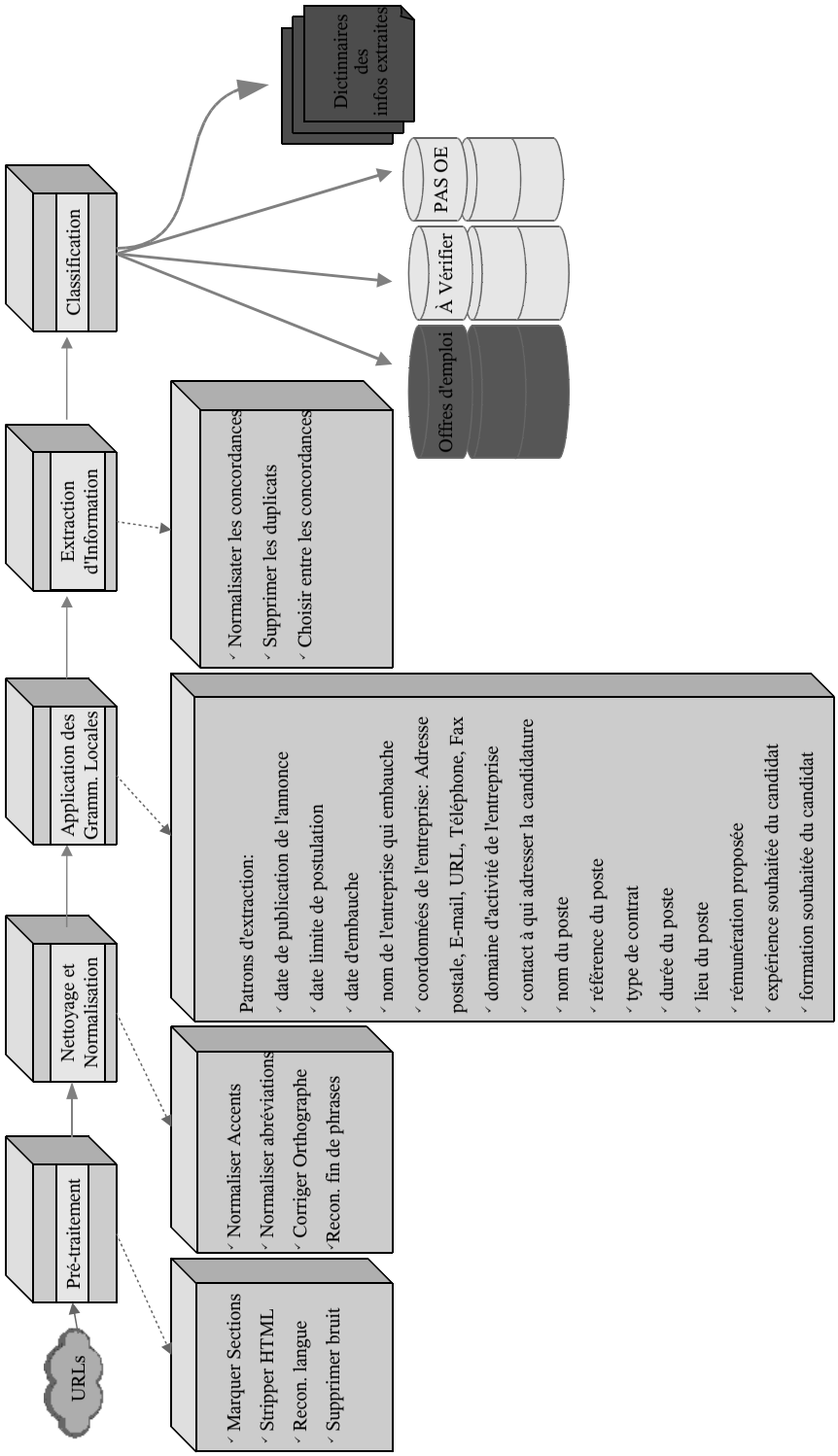


FIG. 8.1 – Le module d'extraction d'information et de transformation de l'espace de représentation : TransOE

- Le module de reconnaissance automatique de la langue peut alors être appliqué. Il s'agit d'une analyse du texte reçu en entrée avec des dictionnaires de différentes langues. Nous testons l'appartenance du texte aux trois langues : française, allemande et anglaise pour lesquelles nous disposons de dictionnaires. Ce programme nous fournit le pourcentage d'appartenance du texte à chacune des trois langues testées. Seuls les documents reconnus pour être rédigés en langue française sont retenus dans la suite de l'analyse. Les autres sont transférés dans une base de données parallèle qui nous servira ultérieurement dans la phase d'expansion de notre système aux deux autres langues citées.
- Le module de suppression du bruit que nous discernons dans la figure (fig. 8.1) est particulier à certaines entreprises. En effet nous avons élaboré lors de la phase d'apprentissage et pour un certain nombre d'entreprises, un module de pré-traitement spécial qui supprime des passages textuels inutiles ou menant à confusion lors de l'extraction automatique, nous supprimons par exemple des passages publicitaires introduits sur chaque offre à la même position ou des phrases types non porteuses d'informations intéressantes et ce en fonction du nom de domaine de la page Web étudiée.

Nettoyage et Normalisation

- La phase de nettoyage consiste à corriger les fautes d'orthographe et les accents manquants dans le texte récupéré. Nous disposons d'une liste de fautes d'orthographes que nous avons récolté lors de la phase d'apprentissage sur un corpus d'offres d'emploi de grande taille, auxquels nous faisons correspondre les séquences équivalentes correctement écrites. De même pour les mots sans accents, pour lesquels nous avons élaboré des paires de mots, les premiers sans accents et leur correspondant avec accents. Nous avons pour ceci extrait tous les mots avec accents à partir du corpus d'apprentissage et avons mis en correspondance les formes avec les accents manquants. Nous retenons à ce niveau que les mots non ambigus pour lesquels la forme sans accent ne correspond à aucune unité lexicale connue des dictionnaires de la langue française.
- la phase de normalisation permet quant à elle de restituer les abréviations dans leur formes originales (*ing.* $\mapsto$ *ingénieur, comm.* $\mapsto$ *commercial...*). A cette fin nous avons également constitué une liste d'abréviations extraites à partir du large corpus d'apprentissage auxquelles nous avons associé les termes ou les séquences

d'origine. A la fin de cette étape nous disposons d'un texte nettoyé et normalisé auquel les analyses syntaxiques et lexicales peuvent être appliquées.

Application des grammaires locales

Pour cette phase d'application des grammaires locales, nous utilisons les programmes de segmentation, d'analyse lexicale et sémantique et d'application des graphes distribués avec Unitex.

- Nous avons élaboré plusieurs dictionnaires spécialisés de mots simples et composés respectant le format DELA étudié dans un chapitre antérieur. Nous comptons 17 dictionnaires en tout à côté de ceux distribués avec la version standard d'Unitex. Les mots de nos dictionnaires ne sont pas forcement nouveau par rapport au DELAS, mais sont organisés dans des classes sémantiques plus adaptées à notre sous-langage comme la classe *«Domaine d'activité», «Service dans les entreprises», «Nom de Profession simples et composés», etc.*
- Étant donné que nous usons du logiciel *Unitex* pour l'application des patrons de reconnaissance sous forme de graphes, il est indispensable de passer les textes par la phase de pré-traitement requise par ce dernier et qui consiste aux programmes suivants :
  1. *Convert* : conversion du texte en Unicode (UTF-16LE)
  2. *Normalize* : normalisation des séparateurs dans le texte
  3. *Fst2txt en mode merge* : reconnaissance des fins de phrases
  4. *Fst2txt en mode replace* : normalisation des formes ambiguës
  5. *Tokenize* : segmentation du texte en unités lexicales
  6. *Dico* : application des dictionnaires au texte
  7. *Sort* : trie des unités lexicales reconnues des dictionnaires
- Une fois l'analyse lexicale exécutée, nous procédons à l'extraction d'information par application des grammaires locales en cascade. Pour chaque information à extraire nous avons construit plusieurs grammaires plus ou moins sûres qui sont appliquées en cascade et sous différentes conditions. Certaines grammaires locales assez ambiguës ne sont appliquées que dans le cas où les patrons d'extraction prioritaires sûrs ne délivrent aucun résultats. Cette méthode permet d'améliorer de beaucoup les performances de notre système d'extraction. La suite de ce chapitre détaille cette phase particulière de construction de grammaires locales ainsi que leur exécution en cascade.

Extraction d'information
Dans la phase que nous appelons «extraction d'information», il s'agit de

- normaliser les différentes concordances trouvées pour une information cherchée et ce en normalisant les espaces et en supprimant tous les caractères de ponctuation de la séquence.
- supprimer les duplicatas : des concordances doubles peuvent être retrouvées par différents patrons de reconnaissance où les séquences sont identiques mais l'étiquetage syntaxico-sémantique est différent.
- choisir parmi les concordances incluses les unes dans les autres. Il arrive que des concordances s'avèrent être des sous segments d'autres concordances pour la même information cherchée, dans un tel cas nous favorisons toujours les concordances les plus longues.

Classification
La phase de classification est la phase de remplissage de la base de données des offres d'emploi ou celle des offres à vérifier. Nous avons mis en place des règles qui permettent de décider en fonction de la quantité d'information trouvée s'il s'agit d'une offre d'emploi ou pas. Les documents dans lesquels plus de 50 % des informations cherchées sont trouvées sont directement introduits dans la base de données des offres alors que ceux où aucune information n'a été trouvée sont automatiquement écartés et ceux avec un taux d'information trouvée moyen sont migrés vers une base de données parallèle et doivent être vérifiés manuellement. Nous avons développé une interface Web conviviale qui nous permet d'avoir une vue d'ensemble du document et d'enrichir en quelques clics les champs du formulaire manquants ainsi que d'enrichir les dictionnaires sémantiques par les mots inconnus coloriés différemment dans l'interface.

Nous avons montré au niveau du chapitre 1 des exemples d'offres d'emploi et avons également insisté sur les différentes informations que nous tentons d'en extraire et qui font office de valeurs aux attributs du formulaire que nous avons fixé. Nous tentons pour chaque document de le structurer selon le formulaire suivant :

Structure d'une offre d'emploi à extraire	
Date de publication	22. Jan 2007
Date limite de postulation	fin février
Date d'embauche	mi- Mars
Nom du poste	ingénieur d'étude en électromécanique
Type de contrat	intérim à temps partiel : 1 à 2 jours/semaine
Durée du poste	8 mois renouvelables
Lieu du poste	sud-est de Paris
Rémunération	selon profil et expérience
Référence du poste	MOR34544/ing-21
Expérience minimum	expérience de 2 à 3 ans dans un poste similaire
Formation du candidat	de formation Bac+5 de type école d'ingénieur
Nom de l'entreprise	CGF Sarl
Coordonnées de l'entreprise	**Adresse :** 34 bis rue des Berthauds, 93110 Rosny s/s Bois **Téléphone :** 0 (+33) 1 12 34 45 67 **Fax :** 0 (+33) 1 12 34 45 68 **Email :** contact@cgf.fr **Home Page :** http :///www.cgf.fr
Contact à qui adresser la candidature	Directeur des RH, Mr. Brice
Domaine d'activité de l'entreprise	Construction électromécanique

Les catégories informationnelles recensées dans le tableau ci-dessus représentent les attributs d'indexation des offres d'emploi que nous ajoutons à notre base de document. Ces attributs représentent des informations très utiles pour améliorer les performances de recherches ultérieures en donnant la possibilité aux utilisateurs de poser des filtres sur les attributs qui correspondent ou pas à leurs volontés. Nous détaillons dans la suite de ce chapitre les grammaires locales construites afin de reconnaître ces catégories informationnelles dans n'importe quelle offre d'emploi.

8.2 Nom du poste

Le nom du poste est l'une des informations les plus importantes à extraire à partir d'une offre d'emploi, car il représente l'image de celle-ci face aux demandeurs d'emploi. Nous avons consacré le chapitre précédent à la description du dictionnaire électronique des noms de profession simples et composés que nous avons construit, nous y avons étudié la typologie des noms de profession complexes et y avons distingué les différentes catégories sémantiques des sous-composants pouvant intervenir dans ces derniers. L'utilisation des dictionnaires est certes très utile dans une tâche d'extraction d'information mais peut également occasionner des ambiguïtés s'ils sont appliqués hors contextes, du fait de la richesse de langue française en polysémie en particulier. Afin d'optimiser la reconnaissance du nom du poste dans les offres d'emploi, nous nous sommes donc intéressés à l'étude de leurs contextes d'apparition externes après en avoir étudié les contextes internes au niveau du chapitre précédent.

La reconnaissance du nom du poste est un processus tâche itératif que nous recherchons en cascade sur 8 niveaux de priorités conditionnels. Nous appliquons nos grammaires de reconnaissance les plus sûres en priorité et si aucune concordance n'a été trouvée dans le document, nous appliquons les grammaires du niveau supérieur, dans le cas contraire où une ou plusieurs concordances ont été trouvées, l'algorithme s'arrête et nous considérons la ou les concordances trouvées comme la réponse à notre besoin informationnel. Afin de choisir la concordance la plus authentique parmi plusieurs trouvées, nous avons élaboré un certain nombre de règles d'élagage assez simples comme le fait de supprimer les signes de ponctuation ou de vérifier la contenance d'une séquence dans une autre etc. Au fur et à mesure des itérations, les grammaires deviennent de moins en moins sûres et peuvent entraîner de plus en plus de bruit. Cette méthode d'itérations conditionnelles nous permet de minimiser les extractions ambiguës ainsi que le nombre de concordances trouvées possibles au niveau de chaque itération. L'application en même temps de la totalité des grammaires locales élaborées pour reconnaître le nom du poste, fournirait une liste assez longue de concordances pour lesquelles, nous devrions construire des patrons d'élagage qui peuvent s'avérer complexes.

Avant de détailler les grammaires de reconnaissance des différents

niveaux mis en oeuvre, nous présentons ici la description d'un contexte très puissant qui devrait accompagner tout nom de profession selon les règles légales de rédaction des annonces d'emploi publiées par l'ANPE et qui interdisent entre autres toute forme de discrimination fondée sur le sexe dans le texte de l'annonce ainsi que dans son titre. Pour éviter de tomber dans ce piège de la discrimination, la convention propose à tout annonceur d'accompagner le nom du poste écrit dans sa forme de base par la mention « (H/F), HF, mf » et qui précise que l'annonce s'adresse aussi bien à des candidats Hommes(H) qu'à des candidates Femmes (F). Nous présentons dans la grammaire de la figure (fig. 8.2), la typologie d'un nom de profession se terminant par le contexte droit « H/F ». Nous avons mentionné dans le chapitre précédent que nous admettons l'hypothèse que tout nom de profession complexe se doit de commencer par un nom de profession simple parmi ceux que nous avons recensé dans le dictionnaire *DELAS-ProfSimple* et auxquels nous ajoutons la liste des composés du genre « Prud'homme, second de cuisine, etc ». Nous avons éliminé de notre *DELAS-ProfSimple* les noms simples très ambiguës comme *« aide, chargé, attaché, adjoint »* qui viennent souvent augmenter un nom de profession simple mais ne représentent pas des entités reconnues pour être des noms de profession à part entière. Nous avons également enrichi les noms de profession simples ambiguës par des catégories sémantiques particulières pour éviter les confusions causées par leur caractère polysémique, citons les noms de profession simples comme *« cadre, industriel, opérateur, commercial, général, modèle... »*, que nous manipulons séparément et surtout au niveau des itérations les plus profondes où l'on essaye d'extraire le nom du poste en ignorant son contexte droit et gauche. Cette répartition nous a permis de diminuer considérablement le bruit dans la phase d'extraction du nom du poste dans les offres d'emploi.

Ainsi nous avons construit la grammaire descriptive « H/F » telle qu'elle reconnaît toute séquence commençant par un nom de profession simple ou toute combinaison comme décrite dans le sous-graphe *JD-init* présent en seconde position sur la figure (fig. 8.2) et qui permet de reconnaître des séquences commençant par :

- <aide><-> ?<NomProfessionSimple> → un Aide-comptable
- <NomProfessionSimple><-><NomProfessionSimple> → un concepteur-programmeur
- <Prefixe><-> ?<NomProfessionSimple> → un électro-

mécanicien

Dans cette grammaire, nous permettons n'importe quelle suite de caractères et de mots pouvant se trouver entre le début restreint comme ci-haut et l'une des séquences représentative de « H/F ». La liste suivante présente quelques concordances reconnues par cette grammaire dans le corpus test que nous avons construit.

- <PosteName>Commercial / Chargée de Clientèle </PosteName> (H/F)
- <PosteName>Concepteur, développer / Chef de Projet BI </PosteName> (H/F)
- <PosteName>Concepteur Logiciel STB H264 HD </PosteName> (H/F)
- <PosteName>Commerciaux terrains</PosteName> (H/F)
- <PosteName>Assistant administratif département Voyages (CDD) </PosteName> (H/F)
- <PosteName>Ingénieur Commercial Fort Potentiel </PosteName> H/F

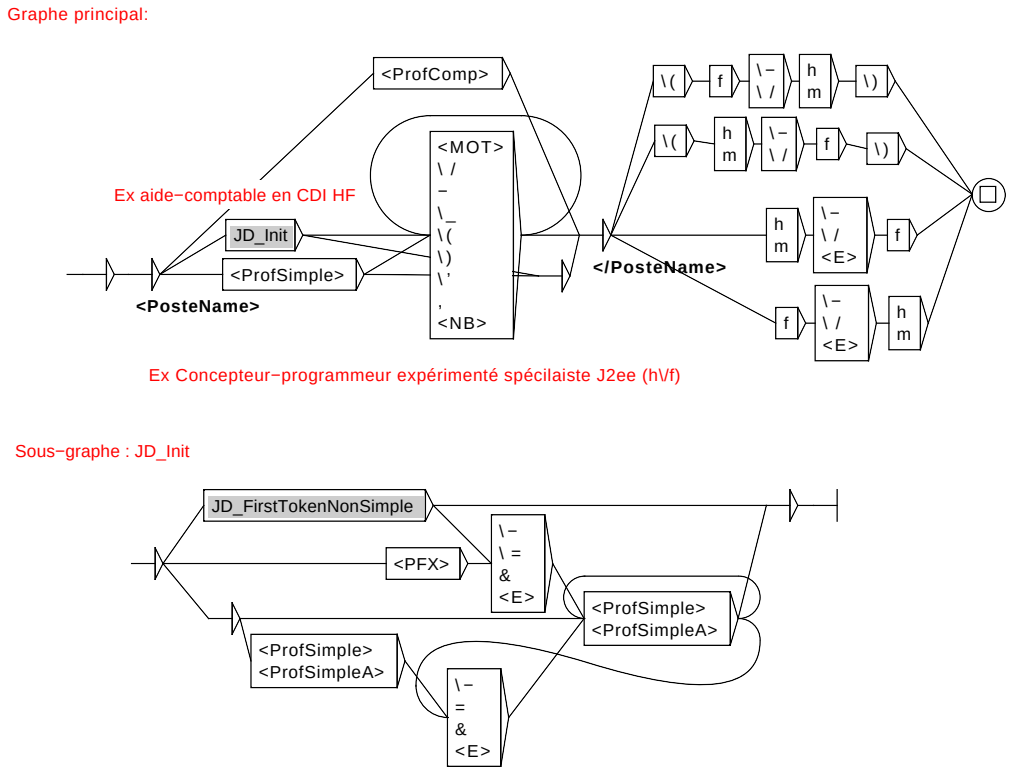


FIG. 8.2 – GL de niveau supérieur pour le *Nom du poste*

- <PosteName>Architecte Mécanique, Thermique, Propulsion satellite </PosteName> H/F
- <PosteName>Conseiller de Vente - temps partiel - Parly2 </PosteName> (H/F)
- <PosteName>Assistante de projet polyvalente trilingue/ Secrétaire </PosteName> (m/f)
- <PosteName>Intégrateur logiciel (tests et recette) </PosteName> (H/F)
- <PosteName>STAGIAIRE (hf) développement JAVA , ingénieur INFORMATIQUE </PosteName> (H/F)
- <PosteName>Comptable International Magasins Anglais néerlandais (59) </PosteName> H/F
- <PosteName>Technico-Commercial B to B </PosteName> H/F
- <PosteName>aide-conducteur de machine héliogravure </PosteName> (H/F)

Nous pouvons voir au niveau des concordances présentées que les noms des postes ne sont pas toujours des appellations d'emploi que nous pouvons recenser dans un dictionnaire. La reconnaissance du nom du poste de la 3ème concordance par exemple et qui ne serait pas accompagnée du contexte « H/F » ne serait possible que dans le cas où les termes *« STB »*, *« H264 »*, *« HD »* soient déjà classifiés dans la classe « Produit » ou dans la classe « Nom spécifique (*<NSpec>*) ». On remarque également dans l'exemple *« Architecte Mécanique, Thermique, Propulsion satellite »* qu'il nous suffit d'avoir préalablement classifié les 3 élément « mécanique », « thermique »et « propulsion satellite » dans la classe « Domaine » ou dans la classe « Connaissances » pour être capable de reconnaître tout nom de poste formé à partir d'une combinaison d'un ou plusieurs domaines d'activité comme ce dernier. Nous augmentons au fur et à mesure semi-automatiquement nos dictionnaires sémantiques par le biais d'une interface conviviale que nous avons développé. Cette grammaire « H/F » est très puissante car elle nous permet de capturer les noms des postes même si l'appellation d'emploi n'a pas encore été recensée dans nos dictionnaires ou si les composants pas encore été classifiés dans les catégories sémantiques identifiées. Elle nous permet également d'enrichir semi-automatiquement toutes ces classes. Bien que très puissante, elle peut provoquer 2 types d'erreurs en fonction que l'on choisisse l'hypothèse de la concordance la plus longue ou celle de la concordance la plus courte. Le tableau suivant montre le taux d'erreur observé pour chacun de ces 2 choix sur un corpus test de 400 offres d'emploi, recueillies à partir de sources différentes.

Erreurs observées		
	NB concordances erronées	Taux d'erreur
Les concordances les +longues	19/398	4,7%
Les concordances les +courtes	91/413	22,03%

Si l'on choisit de favoriser les concordances les plus courtes, on aura tendance à augmenter les erreurs de délimitation gauche du nom du poste comme on peut l'observer dans les exemples suivants :

- France -IDF-PARIS-CONSULTANTS <PosteN>ARCHITECTE TECHNIQUE </PosteName>H/F
- IDF-Paris-Assistant/<PosteN>Assistante Polyvalent de Langue Maternelle Anglaise </PosteName>H/F
- la recherche.S Concepteur, développeur / <PosteN>Chef de Projet BI </PosteName>H/F
- -IDF-RUEIL-MALMAISON-TECHNICIEN PHYSICO- <PosteN>CHIMISTE </PosteName>H/F
- division Ajilon Technical recherche Technico-<PosteN>Commercial - Distribution éléctrique /Domotique </PosteName>m/f
- de maisons en bois, un secrétaire <PosteN>COMMERCIAL BILINGUE ALLEMAND </PosteName> H/F
- nous amènent à intégrer : Ingénieur <PosteN>commercial fort potentiel </PosteName> H/F
- Creyf's recrute en CDI un aide - <PosteN>conducteur de machine héliogravure </PosteName>H/F
- Seine et Marne (77)-DIRECTEUR <PosteN>FINANCIER(ÈRE) RÉGIONAL (Chimie fine) </PosteName>H/F
- 15 000 entreprises clients. PEINTRE <PosteN>INDUSTRIEL </PosteName>H/F
- consultant fonctionel SAP CO pour <PosteN>INDUSTRIEL PRESTIGIEUX en CDI </PosteName>H/F
- ients. RESPONSABLE PRODUCTION BOULANGERIE <PosteN>INDUSTRIELLE </PosteName>H/F
- 55458373 STAGIAIRE développement JAVA , <PosteN>ingénieur INFORMATIQUE </PosteName>H/F

Cette reconnaissance partielle engendre certes des noms syntaxiquement corrects mais ils ne reflètent souvent plus le besoin exprimé dans l'annonce initiale. Si on recherche dans l'annonce d'origine *« un aide-cuisinier »* ou *« un chef-cuisinier »* et que nos patrons d'extraction

favorisant les concordances les plus courtes se contentent de reconnaître la partie *« cuisinier »* du nom du poste , nous invitons des candidats non ciblés initialement à consulter l'annonce et augmentons ainsi d'une part le bruit dans nos listes de résultats mais engendrons également un élargissement du champ d'application de l'offre par rapport aux désirs du recruteur qui a pris le soin de spécifier exactement son besoin en personnel. Dans la liste des concordances citées, on peut voir quelques exemples de cette transformation du besoin initial et surtout dans le dernier cas, où *un stagiaire* est recherché et que ce type d'extraction associerait plutôt « ingénieur informatique » au nom du poste de cette annonce. Bien que le stagiaire recherché se doive de suivre des études d'ingénieur informatique, le public concerné par les deux noms des postes « stagiaire développement Java, ingénieur informatique H/F » et « ingénieur informatique »est très différent.

Si l'on choisit par contre la reconnaissance sur la base des concordances les plus longues, on obtient d'autres types d'erreurs, liées le plus souvent à la mauvaise ponctuation dans les pages Web et que nous nous devons d'améliorer dans notre module d'analyse de la structure Html au niveau du prétraitement des annonces. Ces erreurs sont beaucoup moins nombreuses que dans le premier choix comme on peut le remarquer dans le tableau cumulatif présenté ci-haut, dans lequel on compte 19 reconnaissances trop longues sur un total reconnu de 398 et dont 379 reconnaissances sont correctement bornées aussi bien à gauche qu'à droite. Voici quelques exemples de ces reconnaissances erronées :

- avec 48 000 <PosteName>collaborateurs présents dans 150 pays, recherche pour sa Direction Informatique française un ingénieur concepteur </PosteName> H/F
- Dans le <PosteName>cadre de la forte croissance de la filiale française, nous souhaitons intégrer le plus rapidement possible un Attaché commercial en Home Office bilingue allemand </PosteName> h/f
- management d'une équipe <PosteName>commerciale (définir les objectifs, suivis des objectifs, former, évaluer, accompagner sur le terrain, motiver, déléguer), idéalement issu du monde de l'édition, vous êtes fortement impliqué dans la vie locale et êtes un</PosteName> H/ F
- <PosteName>Responsable d'édition (H/F) CREYFS RECRUTEMENT, référence du recrutement en CDD et CDI dans la région sud-ouest, recrute pour un de ses clients, un des leaders sur le marché de la presse gratuite en France, filiale du groupe SUD-OUEST, RESPONSABLE D'ÉDITION </PosteName> (H/F)
- <PosteName>Technicien de synthèse organique senior (H/F) en Belgique KELLY SCIENTIFIQUE, spécialiste du recrutement et de la délégation

de scientifiques, recrute pour une société de biotechnologies qui développe et commercialise une nouvelle génération de solutions pour la protection des cultures et de nouveaux concepts et molécules thérapeutiques, un Technicien de synthèse organique senior </PosteName> (H/F)

- <PosteName>PROJETEUR éléctrique H/F Adecco Bureau d'études recherche pour un de ces clients spécialisé dans le domaine électrique un dessinateur projeteur en électricité </PosteName> H/F

Dans tous les exemples présentés ci dessus, nous remarquons bien qu'il n'y a qu'une seule occurrence complètement incorrecte, tandis que les autres renferment bien le nom du poste mais la reconnaissance des bornes gauche et droite y est mal placée. Dans l'exemple 3, le contexte « H/F » a été placé avant le nom du poste et non pas à la fin, ce qui engendre l'erreur de reconnaissance du nom. Nous avons opté pour l'adoption du choix de la concordance la plus longue et ce pour tous les types d'informations à extraire. Afin de pallier ces différents problèmes cités et minimiser les ambiguïtés et les reconnaissances erronées, nous avons développé un ensemble de grammaires contextuelles que nous appliquons aux textes itérativement et sous conditions. Les itérations des premiers niveaux font appel à des patrons de reconnaissance certains pour lesquels nous n'avons pas observé d'erreurs d'extraction. Le tableau ci-dessous résume les 8 niveaux d'applications de ces grammaires triées des plus détaillées et sûres aux plus générales et ambiguës.

Les 8 niveaux de priorités pour l'extraction du Nom du Poste	
Niveau1	Locution structurelle type+gram. de contexte externe(H/F)
Niveau2	Nous recherchons+grammaire de contexte externe(H/F)
Niveau3	Locution structurelle type+Noms de profession du dictionnaire
Niveau4	Nous recherchons+Noms de profession du dictionnaire
Niveau5	Locution structurelle type+typologie des noms de profession
Niveau6	Nous recherchons+gram. typologique des noms de profession
Niveau7	Grammaire du contexte externe H/F
Niveau8	Noms de profession sans contexte

Les grammaires du niveau $n + 1$ sont appliquées dans le cas où les grammaires du niveau n ne retournent aucun résultat.

Niveau 1 :

Au niveau de la première itération , nous appliquons la grammaire locale de la figure (fig. 8.3) et qui permet d'extraire le nom du poste à partir d'offres d'emploi plus ou moins structurées, enrichies par des locutions types telles que :

Nom du poste :
Intitulé de l'offre :
Le poste :
Poste à pourvoir :
Offre proposée de :
...

Ces locutions identifiées dans une phase d'apprentissage sont très sûres, elles représentent le contexte gauche venant introduire le nom

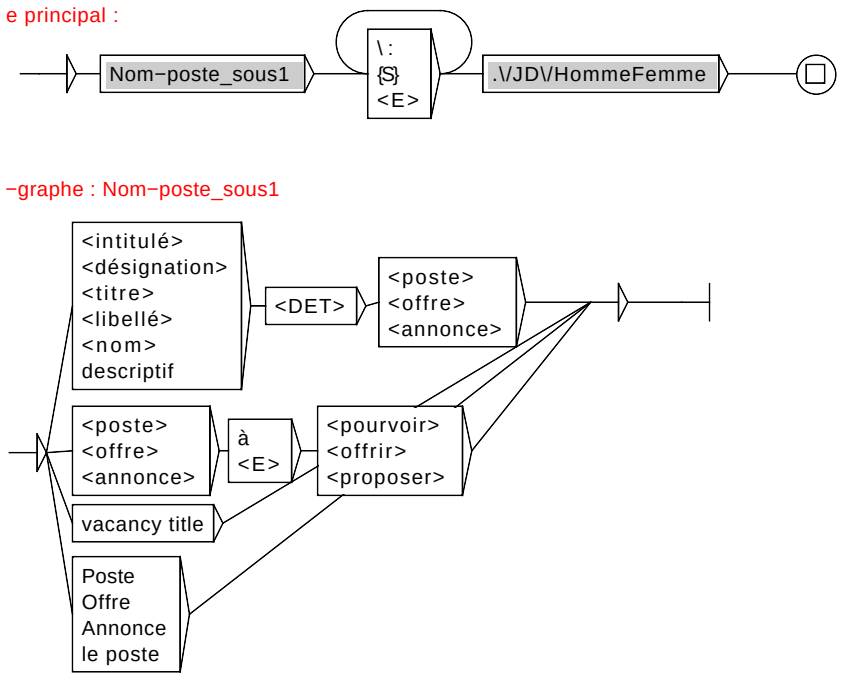


FIG. 8.3 – Introduction du nom du poste par des locutions semi-figées

du poste dans les annonces d'emploi semi-structurées.

Cette grammaire est très restrictive, elle a donc un rappel assez médiocre mais donne une précision de 100 %. Les entreprises de grande vigueur, ont tendance à structurer sémantiquement leurs annonces d'emploi, c'est pourquoi nous nous sommes consacrés dans une première phase d'apprentissage à l'identification des différentes classes structurelles sémantiques ainsi qu'au repérage des locutions usitées par les recruteurs pour les mettre en valeur. Ces classes structurelles sémantiques ainsi que des exemples de leurs contenus sont décrits au niveau du chapitre 1. Dans le cas où l'annonce est structurée, nos programmes d'extraction ne rencontrent plus de problèmes de reconnaissance de bordures pour les différentes informations recherchées, c'est pourquoi il est important dans une première itération de vérifier le niveau structurel de l'annonce, qui en fonction de son existence ou pas, nous délivre plus ou moins de concordances ambiguës. Une telle grammaire lancée sur le même corpus test cité plus haut délivre les concordances suivantes :

- Postuler à cette Offre<PosteName> CONSULTANT MOA Banque Finance/Titres </PosteName> (H/F) Partners est
- INTITULÉ DU POSTE :<PosteName>WEBMASTER </PosteName>(H/F) RATTACHÉ AU : CHEF DE PROJET
- Poste :<PosteName> CONSEILLERS EN GESTION DE PATRIMOINE </PosteName> (H/F) Type : Temps plein
- Description du poste :<PosteName>Directeur Système </PosteName> H/F En collaboration avec la Direction item propose un poste<PosteName> ANALYSTE </PosteName>(H/F) (Conception et Développement de SI) Mission : Au
- cette offre<PosteName> Responsable de Projet confirmé </PosteName> H/F entreprise Il agit de
- Contenu de annonce<PosteName>Ingénieur PRODUCTION Junior </PosteName> (H/F) WINDOW Server
- postuler à cette offre<PosteName> CHEF DE PROJETS OFFRES </PosteName> (H/F) Cette entreprise est
- VINGT POSTES :<PosteName> MENUISIER INSTALLATEUR POSEUR </PosteName> H/F Notre enseigne
- dans le cadre d'une création de poste <PosteName> GESTIONNAIRE CHARGES LOCATIVES </PosteName> H/F Auprès de
- croissance et de développement et créons le poste<PosteName>Assistant Contrôleur de Gestion </PosteName> h/f à Achenheim
- un contexte une création de poste :<PosteName> RESPONSABLE APPROVISIONNEMENT </PosteName> H/F IDF
- infos sur www.sante.com Poste<PosteName> Conseiller commercial vente de plats cuisinés mixés et hachés </PosteName> h/f

- Poste :<PosteName> AUDITEUR TECHNIQUE TOULOUSE </PosteName> (H/F) Type : Temps
- 2 postes<PosteName> ingénieur Broadcast </PosteName> (H/F), IDF 92000 Géant mondial
- très motivés pour le poste<PosteName>Ingénieur support client réseau NSS </PosteName> (H/F) Le candidat

Une seule concordance erronée a été observée, elle est dûe à une mauvaise manipulation des règles typographique. Dans cette concordance trop longue :

> postuler à cette Offre <PosteName> Fabriquant d'instruments de mesure et de systèmes numériques industriels, basé en région parisienne (92) recherche des technico-commerciaux </PosteName> (H/F)

on remarque bien que deux facteurs entrent en jeux dans son déclenchement : le premier déclencheur *Offre* est mis faussement en majuscule et le second, le terme « Fabriquant » appartient entre autres à la classe des noms de profession simples.
Si cette première itération ne nous délivre aucun résultat, l'algorithme passe à la seconde itération en appliquant les grammaires du 2ème niveau.

Niveau 2 :

Cette seconde grammaire de la figure (fig. 8.4) joint des grammaires descriptives des contextes gauches à la grammaire du contexte droit « H/F » étudiée plus haut. Nous avons élaboré ces grammaires par *Bootstarping* en observant les contextes gauches et droits des verbes *« rechercher », « recruter », « chercher »* qui viennent très souvent

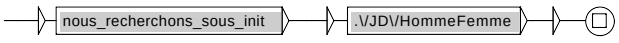


FIG. 8.4 – Introduction du nom d'emploi par des verbes prédicatifs

introduire le nom du poste à pourvoir dans un texte d'une annonce d'emploi. Nous présentons dans les figures (fig. 8.5, 8.6, 8.7) les sous-grammaires les plus importantes appelées dans la grammaire principale de la figure (fig. 8.4).

Les concordances reconnues par la grammaire du second niveau sont :

- recherche pour sa filiale française : un<PosteName>Expert </PosteName>H/F
- nous recrutons un :<PosteName> Responsable Régional CHR </PosteName> H/F
- <ORG>BRIDGESTONE FRANCE </ORG> recrute : 1<PosteName>CHARGE DE CLIENTELE </PosteName> (H/F)
- Nous recherchons pour notre client basé à <Lieu> Montreuil </Lieu>, un<PosteName>aide comptable </PosteName>H/F
- recherche pour sa filiale française basé à<Lieu> Quimper </Lieu> un :<PosteName> Responsable Financier</PosteName> H/F

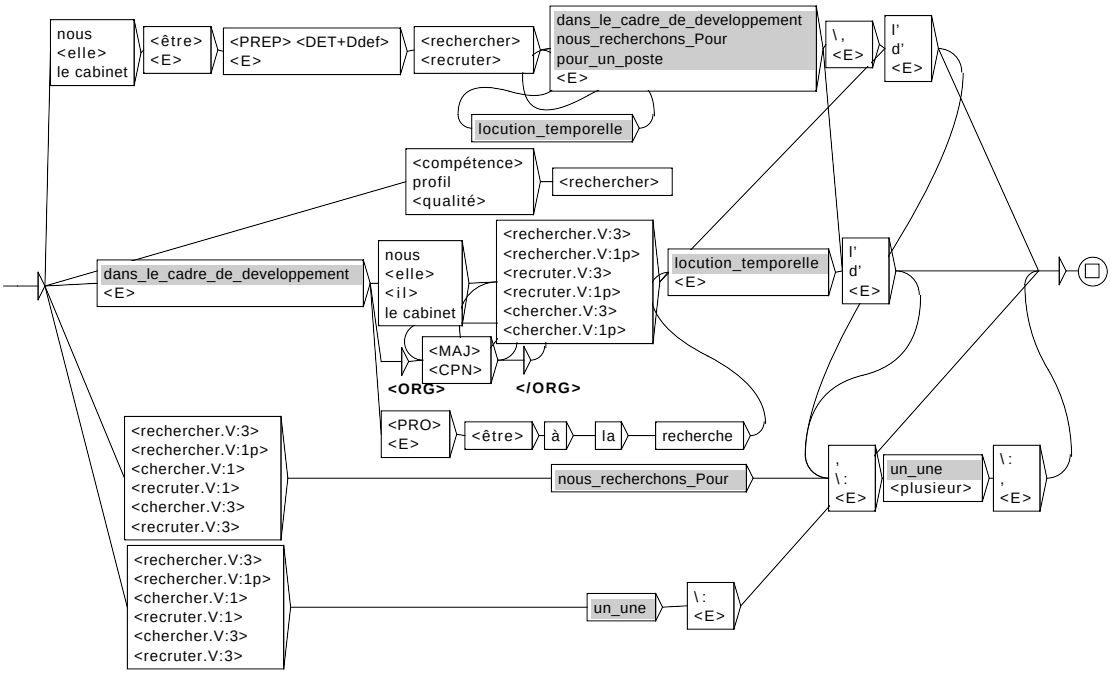


FIG. 8.5 – Sous-grammaire : *Nous_ recherchons_ HF_ sous1*

- Nous recherchons un(e)<PosteName> Technicien Méthodes </PosteName> (H/F)
- recherche, pour accompagner son développement, des : <PosteName> Commerciaux Sédentaires </PosteName> H/F
- rechercheur<Duree> plusieurs mois </Duree> un <PosteName> Assistant export trilingue Anglais/Allemand</PosteName> H/F
- Nous sommes à la recherche d'un<PosteName> Assistant Paie et Gestion Sociale </PosteName> H/F
- nous recherchons pour notre siège européen<Lieu> de Paris </Lieu> un : <PosteName> Responsable financier </PosteName> H/F
- RH Facilities recrute pour l'un de ces clients : un<PosteName> collaborateur d'agence </PosteName> (H/F) </PosteName>
- recherche pour l'un de ses clients spécialisé dans le domaine industriel un :<PosteName> ASSISTANT</PosteName>H/F
- recherche pour son service comptable un(e) <PosteName> secrétaire comptable </PosteName> (H/F)
- recherche pour l'un de ses clients<Lieu> de la métropole lilloise </Lieu> un <PosteName> COMPTABLE </PosteName>H/F

Nous avons observé certaines erreurs en appliquant ces grammaires

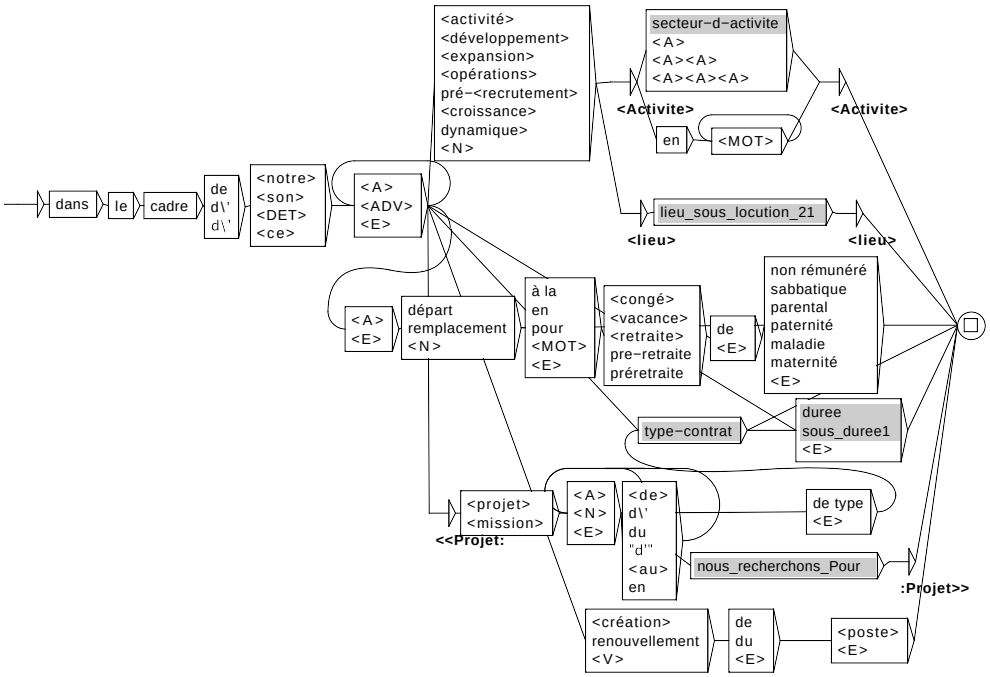


FIG. 8.6 – Sous-grammaire : *nous_ recherchons_ dans_ le_ cadre_ sous1*

du second niveau sur le corpus test, il s'agit néanmoins d'erreurs d'étiquetage, ou d'étiquetages multiples comme on peut le voir dans les exemples qui suivent extraits à partir de la liste des concordances reconnues.

- <ORG>GROUPAMA PARIS VAL DE LOIRE </ORG> recherche des :<PosteName> Managers Commerciaux </PosteName> H/F
- commercialisation, nous recherchons pour notre siège <Lieu> européen de Paris </Lieu> un :<PosteName> Responsable financier </PosteName> H/F

Une seconde erreur que nous avons pu constater est dûe au fait qu'il arrive que les rédacteurs des annonces usent de la marque « H/F » à la suite du nom de profession simple au lieu de le disposer à la fin du nom du poste complexe. Nous avons pu constater ceci dans la concordance suivante par exemple :

- nous sommes à la recherche de notre <PosteName> Assistant </PosteName> (H/F) de Direction.

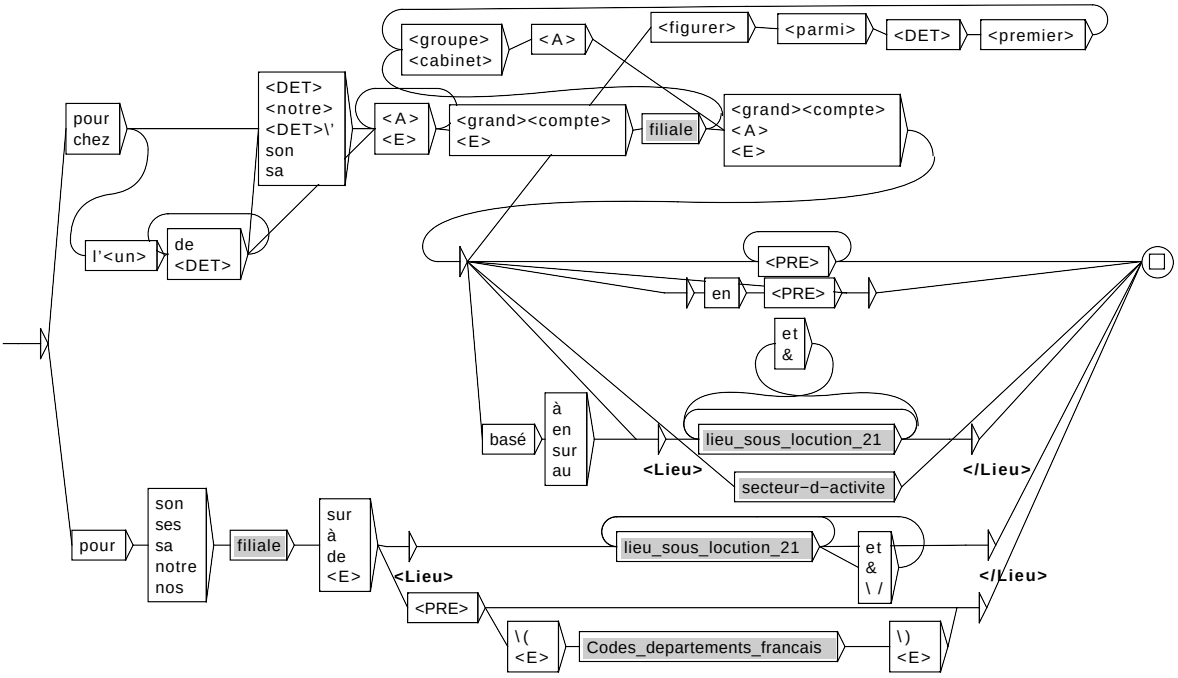


FIG. 8.7 – Sous-grammaire : *nous_ recherchons_ Pour*

Cependant et toujours en se référant à notre corpus test construit à partir de sources différentes, ces deux erreurs sont statistiquement distribuées comme suit :

Erreurs observées sur le corpus Test1	
Nombre de concordances trouvées	43
Nombre de concordances pertinentes	45
Nombre de concordances incorrectes du type 1	1
Nombre de concordances incorrectes du type 2	2
Nombre de concordances non trouvées	2
Precision	40/43= 0,93 %
Rappel	40/45 = 0,88 %

Niveau 3 et 4 :

Au niveau de la troisième et de la quatrième itération, nous avons construit des grammaires qui fixent les contextes gauches comme décrit dans les deux premiers niveaux et reconnaissons comme « nom du poste », les noms recensés dans notre dictionnaire des noms de profession composés ou des combinaisons entre plusieurs noms de profession composés du dictionnaire DELAC-Profession. La figure (fig. 8.2) présente les deux graphes principaux des deux niveaux de l'itération 3 et de l'itération 4 qui nous le rappelons sont des itérations atteignables que dans le cas où le nom du poste n'a pu être trouvé au niveau des itérations précédentes.

Nous présentons dans la liste suivante des exemples d'extraction du nom du poste par la grammaire du niveau 3 :

- Formation : BAC + 5 Désignation du poste : <PosteName> Contrôleur de gestion Production Expérimenté </PosteName>
- financière Libellé du Poste : <PosteName> Responsable Comptable Crédit-Bail Contexte </PosteName>
- nombre de sites Poste :<PosteName> Chef de projet Déploiement </PosteName> - Secteur Santé
- Intitulé du poste :<PosteName> Développeur JAVA/WEBLOGIC </Pos-

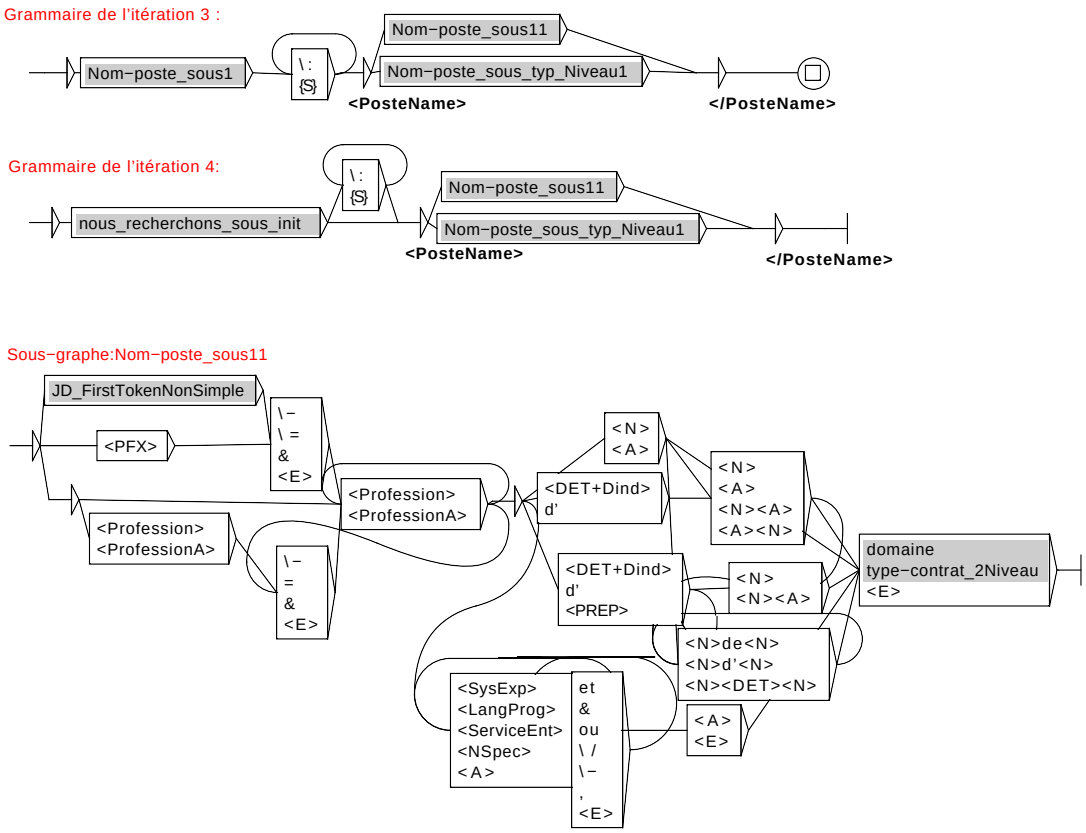


FIG. 8.8 – Grammaires des niveaux 3 et 4

teName> Date de démarrage : ASAP
- Poste : <PosteName> Analyste décisionnel INFORMATICA Senior </PosteName> Intégration au projets
- nous recrutons pour notre agence de Toulouse un(e) <PosteName> CHEF DE MISSION </PosteName>
- nous recherchons un <PosteName> Ingénieur Commercial expérimenté </PosteName>
- recrute un <PosteName> Ingénieur Etudes et Développement PACBASE </PosteName>
- Nous recherchons pour un poste en <TypeContrat> CDI </TypeContrat>, Un <PosteName> Responsable Normalisation </PosteName>
- nous recherchons un <PosteName> ingénieur en microélectronique numérique </PosteName>
- recrute son/sa <PosteName> Responsable Marketing & Communication </PosteName>

Les erreurs majeures rencontrées au niveau de ces grammaires sont des erreurs de reconnaissance des bordures droites. Le nom du poste est souvent reconnu en partie seulement comme dans les concordances suivantes :

- Direction : DRH Libellé du Poste :<PosteName> Assistant Gestion des Temps </PosteName> / Paie - CDD
- Poste :<PosteName> Analyste </PosteName> DATASTAGE Technique
- l'europe. Poste :<PosteName> Analyste Business </PosteName> Objects
- recrute des<PosteName> Dessinateurs & Projeteurs en Tuyauterie </PosteName> et/ou Installation générale.

Niveau 5 et 6 :

Pour ce qui est des grammaires de l'itération 5 et 6 , elles décrivent les typologies internes des noms des postes. Nous fixons les contextes d'apparition gauches par les différentes preuves que nous avons pû récolter pendant la phase d'apprentissage et que nous décrivons en détail dans les deux premières itérations. Nous présentons dans les exemples suivants une liste de concordances extraite par ces grammaires.

Ces typologies font intervenir les multiples classes sémantiques que

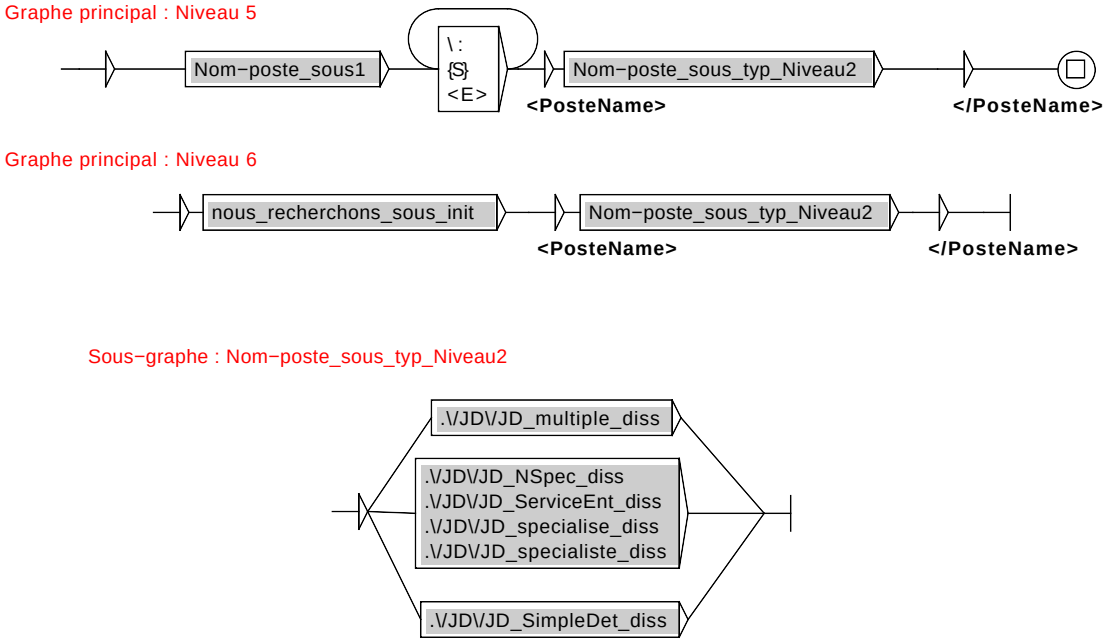


FIG. 8.9 – Grammaires du niveau 5 et du niveau 6

nous avons construites dont les objets viennent souvent composer les noms des postes, comme la classe : « Domaine d'activité », « Services dans les organisations », « Spécialité des professions », « Adjectifs spéciaux », « Nom spéciaux », « Produits », « Marques », « Connaissances » et autres.

- Vacancy Title <PosteName> Technico- Commercial Tivoli - IT Spécialiste </PosteName> "Automation" Contact
- conseil. Poste :<PosteName> Consultant commercial junior ou senior </PosteName>
- de 1,5 M€ de CA. Le poste proposé<PosteName> Assistante administrative et Comptable </PosteName>
- PRESENTATION DU POSTE<PosteName> Interlocuteur principal des clients </PosteName>
- Poste<PosteName> Conseiller commercial vente de plats cuisinés </PosteName>
- INTITULÉ DU POSTE :<PosteName> ASSISTANT DRH </PosteName>

- recherche une <PosteName> secrétaire de livraison VO à particuliers </PosteName> avec de l'expérience
- 500 collaborateurs) recherche des<PosteName> TECHNICIENS DE MAINTENANCE ITINERANTS imprimantes </PosteName>, MFP, copieurs
- nous recherchons pour l'un de nos clients, un(e) <PosteName> Responsable Maintenance / Travaux neufs </PosteName>

A l'opposé des grammaires du niveau précédent, les erreurs observées ici sont des erreurs de reconnaissance de bordures gauches, qui sont souvent trop longues :

- industrielle. Poste<PosteName> généraliste nécessitant une forte polyvalence </PosteName>, autonomie, capacité de travail
- Poste :<PosteName> Contrôleur de Gestion Filiale Au sein </PosteName> de l'une de nos
- recherche un<PosteName> administrateur SAP BC pour le renforcement de son équipe </PosteName>
- Nous recherchons des<PosteName> ingénieurs connaissant les outils </PosteName> et les méthodologies appliquées à la conception de System On Chip
- <ORG>Tyco </ORG> recherche un (e)<PosteName> Chargé d'Affaires pour son activité </PosteName> « Traffic & Transportation »

Niveau 7 et 8 :

Si le nom du poste n'a toujours pas été trouvé avec les grammaires contextuelles des 6 premières itérations, nous appliquons des grammaires descriptives des contextes internes en ignorant les contextes d'apparition externes. En effet, le nom du poste à pourvoir est souvent mis en évidence par des formatages *HTML* et est donc présenté hors contexte, il est donc important d'appliquer les grammaires typologiques des noms des postes dans les itérations profondes pour capturer la séquence répondant à notre besoin informationnel. Dans l'itération 7 et 8 nous appliquons d'abord les grammaires de contexte « H/F » que nous avons décrites au début de cette section et qui présentent comme nous l'avons montré des résultats de reconnaissance assez performants, puis et dans le cas où aucune concordance n'est reconnue, nous recherchons la concordance la plus longue formée à partir d'une combinaison ou d'un nom de profession composés disponibles dans nos dictionnaires

des noms de profession composés. Dans un dernier recours, nous recherchons des correspondances avec nos dictionnaires de noms de profession simples non ambigus.

Dans un premier temps, avant de séparer les noms de profession simples en deux classes, nous obtenions en appliquant les grammaires typologiques hors contextes des erreurs de concordances comme les suivantes :

- Une <PosteName> bonne prestance alliée</PosteName> à un sens du travail en équipe
- Nous recrutons des<PosteName> candidats à fort potentiel </PosteName>,
- des <PosteName>éditeurs de logiciel</PosteName>
- des éléments <PosteName>financiers du projet</PosteName> (coûts, facturation, prise de revenu)
- Expérience <PosteName>professionnelle dans des environnements internationaux </PosteName>
- <PosteName>acteur Européen du conseil en management</PosteName> et des services informatiques
- qui va du <PosteName>conseil en stratégie <PosteName> jusqu'à l'externalisation

Ces erreurs se sont minimisées par la dissociation des noms de profession simples à caractère polysémique.

8.3 Type de contrat

Afin d'améliorer les résultats de recherche d'information dans le domaine des offres d'emploi, nous nous intéressons à indexer celles-ci par l'attribut « Type de contrat », car il représente un critère important dans le choix d'un poste par un candidat. Ce que nous appelons ici « type de contrat »englobe le type du contrat proposé (*CDD, intérim, indépendant...*) mais aussi ce que certains appelleront »type d'emploi »et qui représente l'information sur le nombre d'heures de travail par exemple comme *plein temps, temps partiel...* . Nous groupons ces deux informations car l'une ou l'autre est souvent omise et que nous considérons qu'elles représentent ensemble l'information que nous vé-

hiculons par l'attribut « type de contrat ».

Afin de recenser les indices contextuels pour repérer cette information, nous avons procédé comme pour les autres grammaires par *Bootstraping* après avoir arrêté la terminologie interne pouvant représenter le *type du contrat*. Après avoir enrichi nos grammaires des contextes gauches associés à cette information, nous avons été en mesure d'augmenter également l'index des preuves internes reconnues dans notre large corpus d'apprentissage. La terminologie représentative de ce type d'information joue également le rôle de preuves internes, il s'agit de termes comme « intérim, indépendant, CDI, contrat à durée déterminée... ». Une correspondance directe de l'index des preuves internes sans tenir compte des contextes externes ne suffit pas à l'extraction performante du type de contrat dans une annonce de poste vacant. Nous avons pour cela développé des grammaires locales intégrant des contextes gauches en particuliers et qui sont organisés dans différentes grammaires en fonction de leur taux d'ambiguïté à engendrer et que nous appliquons en cascade et sous conditions.

La première grammaire appliquée et présente à la figure (fig. 8.3) résume certains contextes externes gauches que nous trouvons souvent dans les annonces structurées et qui représentent une preuve externe fiable. Nous présentons ci-joint certaines concordances reconnues par cette dernière.

- Contrat : <TypeContrat>CDD 4 mois</TypeContrat>
- Statut : <TypeContrat>Temps plein, CDI </TypeContrat>
- Type de poste : <TypeContrat>Plein temps, CDI </TypeContrat>
- Type de contrat : <TypeContrat> CDD de 6 à 8 mois</TypeContrat>
- Type de contrat : <TypeContrat> Intérim/CDD/Mission</TypeContrat>
- Emploi proposé : <TypeContrat>CDD 5 à 6 mois transformable en CDI </TypeContrat>
- Poste à pourvoir en <TypeContrat> CDD (1 an) avec perspectives de CDI </TypeContrat>
- Poste à pourvoir en <TypeContrat> Contrat à Durée Indéterminée </TypeContrat>
- Poste à pourvoir <TypeContrat> en vue d'un CDI </TypeContrat>
- Poste à <TypeContrat> mi temps uniquement (25 heures hebdomadaire) </TypeContrat>
- Contrat <TypeContrat> intérim à temps partiel : 1 à 2 jours/semaine </TypeContrat>
- embauche en <TypeContrat> CDI </TypeContrat>

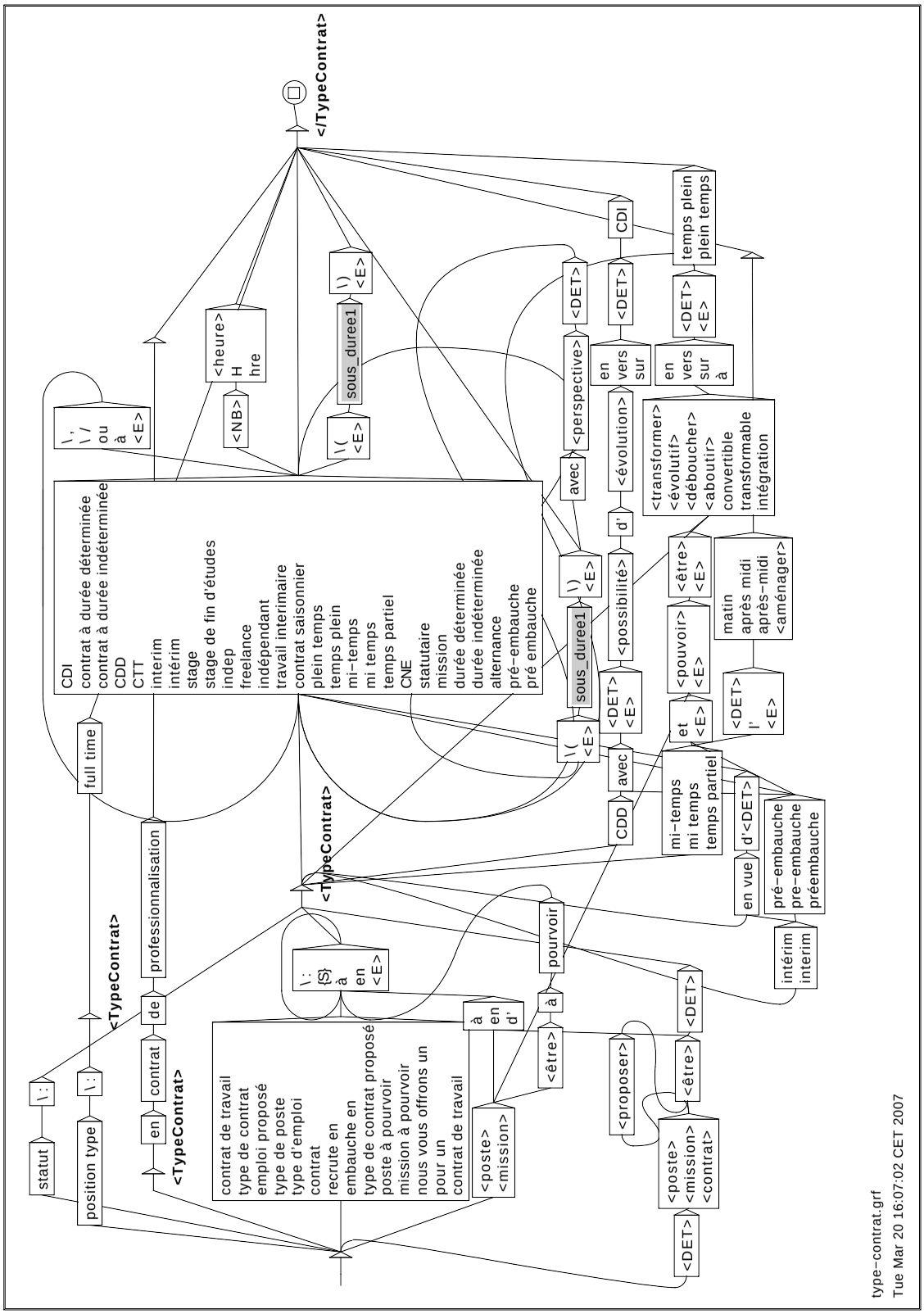


FIG. 8.10 – GL de niveau supérieur pour le *type de contrat*

– Le contrat proposé est un <TypeContrat> CDI </TypeContrat>
– Le poste est un <TypeContrat> CDI (35 heures/semaine) </TypeContrat>
– mission d'interim de 3 mois renouvelable </TypeContrat>
– Poste à <TypeContrat> temps partiel 17.50h/semaine </TypeContrat>

Si cette grammaire de premier niveau ne permet pas de reconnaître le type du contrat proposé, nous appliquons la seconde grammaire de la figure (fig. 8.11) et qui représente des contextes externes gauche pouvant mener à des ambiguïtés. Cette grammaire permet de reconnaître des séquences telles que :

– Comptable à temps partiel (H/F)
– chef de secteur gms H/F (en CDD de 2 mois transformable en CDI)
– pour une mission d'intérim de 6 mois sur le
– la recherche d'un Teamleader pour une période à durée déterminée
– nous recherchons dans le cadre d'un CDI, un Ingénieur
– recherche d'un(e) : bilingue allemand en CNE
– des bouchers expérimentés en vue d'un CDI

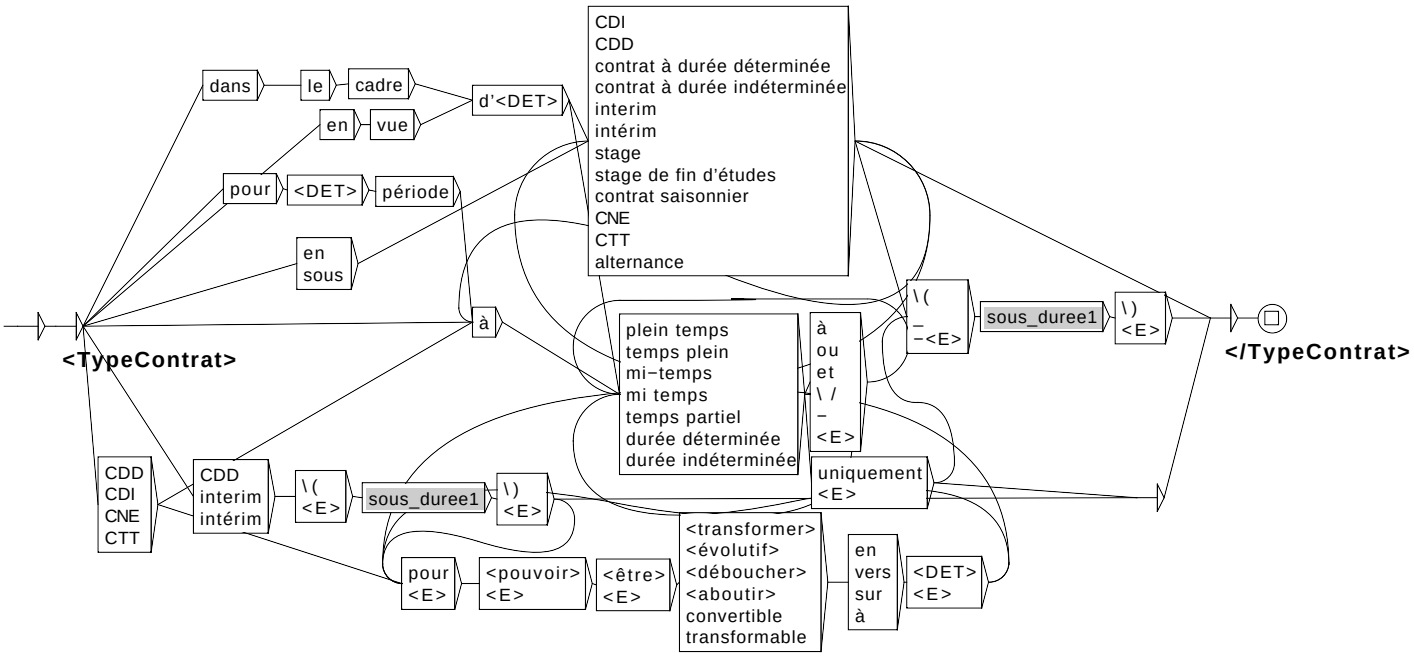


FIG. 8.11 – GL de niveau 2 pour le *type de contrat*

– au sein de cette SSII sous contrat à durée indéterminée (CDI)

Dans le cas où cette grammaire de second niveau ne livre toujours pas de concordances dans le texte, nous appliquons alors la grammaire la plus simple (fig. 8.12) qui met en jeu uniquement les preuves internes en ignorant le contexte d'apparition. Ce conditionnement permet d'augmenter les performances de notre système car le type de contrat particulier représente une information souvent mentionnée sans contextes externes clairs comme c'est le cas dans les exemples suivants :

– concepteur C++ pour Projets de Conseil (92/CDI) (H/F)
– Secrétaire Commerciale-CNE à Lille
– Création de poste (CDI) à pourvoir le plus rapidement possible.
– Sud Ouest CDD / 30000 Euros
– Secrétaire H/F Service Export CDI
– dans laquelle vous voulez évoluer ! CDI, Salaire motivant
– poste basé à Chaponnay (Sud-Est de Lyon) CDI temps plein.

Cette grammaire engendre des cas d'erreurs du type :
– Le groupe SBC Interim (où *Interim* fait partie du nom de la

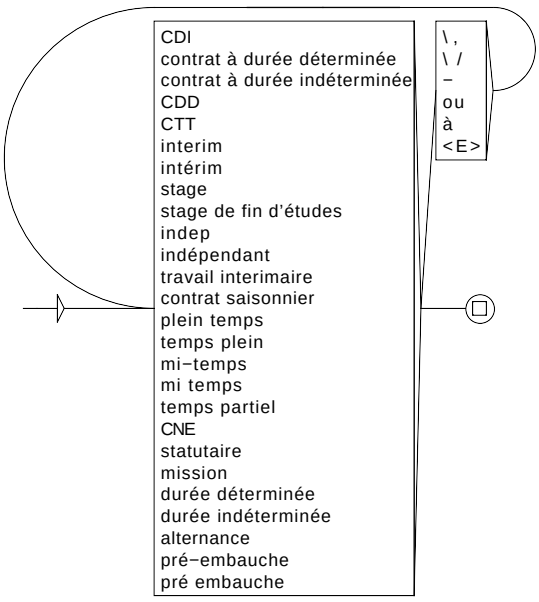


FIG. 8.12 – GL de niveau 3 pour le *type de contrat*

société)
- premier groupe indépendant d'agences (où *indépendant* est un adjectif associé à la société)

8.4 Durée du poste

Pour ce qui est de l'attribut *« Durée du poste »*, il n'est pertinent que s'il s'agit d'un poste à durée déterminée ou une mission en intérim, information reconnue à partir de la phase d'extraction précédente concernant l'attribut *type de contrat*. En effet, un contrat en CDI est illimité, l'information sur la durée y est implicite. Nous avons observé au niveau de l'étape précédente, que la durée de la mission est fréquemment exprimée conjointement avec le type du contrat ; ceci car la première représente un contexte gauche très puissant de la seconde. Mais à côté du type de contrat, nous avons pu répertorier et décrire plusieurs contextes externes gauche qui permettent d'introduire cette information dans les offres d'emploi. La grammaire de la figure (fig. 8.4), permet par exemple de reconnaître les séquences suivantes :

- Durée : <Duree> 3 à 4 mois</Duree>
- durée de <Duree> 6 mois renouvelables</Duree>
- Durée de la mission : <Duree> 3 mois renouvelables</Duree>
- La durée de mission sera de <Duree> 2 mois </Duree>
- Durée du Stage : <Duree> Idéalement UN an </Duree>
- stage prévu jusqu'à <Duree> début septembre </Duree>
- mission d'intérim de <Duree> longue durée </Duree>
- la mission aura une durée approximative de <Duree> 6 mois</Duree>
- pour une mission d'environ <Duree> 3 mois</Duree>
- pour une période minimum de <Duree> 6 MOIS </Duree>

On montre à travers ces exemples, que l'expression d'un laps de temps peut prendre différentes formes, nous avons nécessité afin de mener à bien l'extraction d'une telle information de développer plusieurs grammaires qui décrivent la typologie interne de l'expression d'une période dans le temps. Une de ces grammaires est présentée dans la figure (fig. 8.14), elle permet de reconnaître des périodes exprimées comme en gras dans les exemples ci-dessus.

- 3 à 4 jours/semaine sur 4 mois

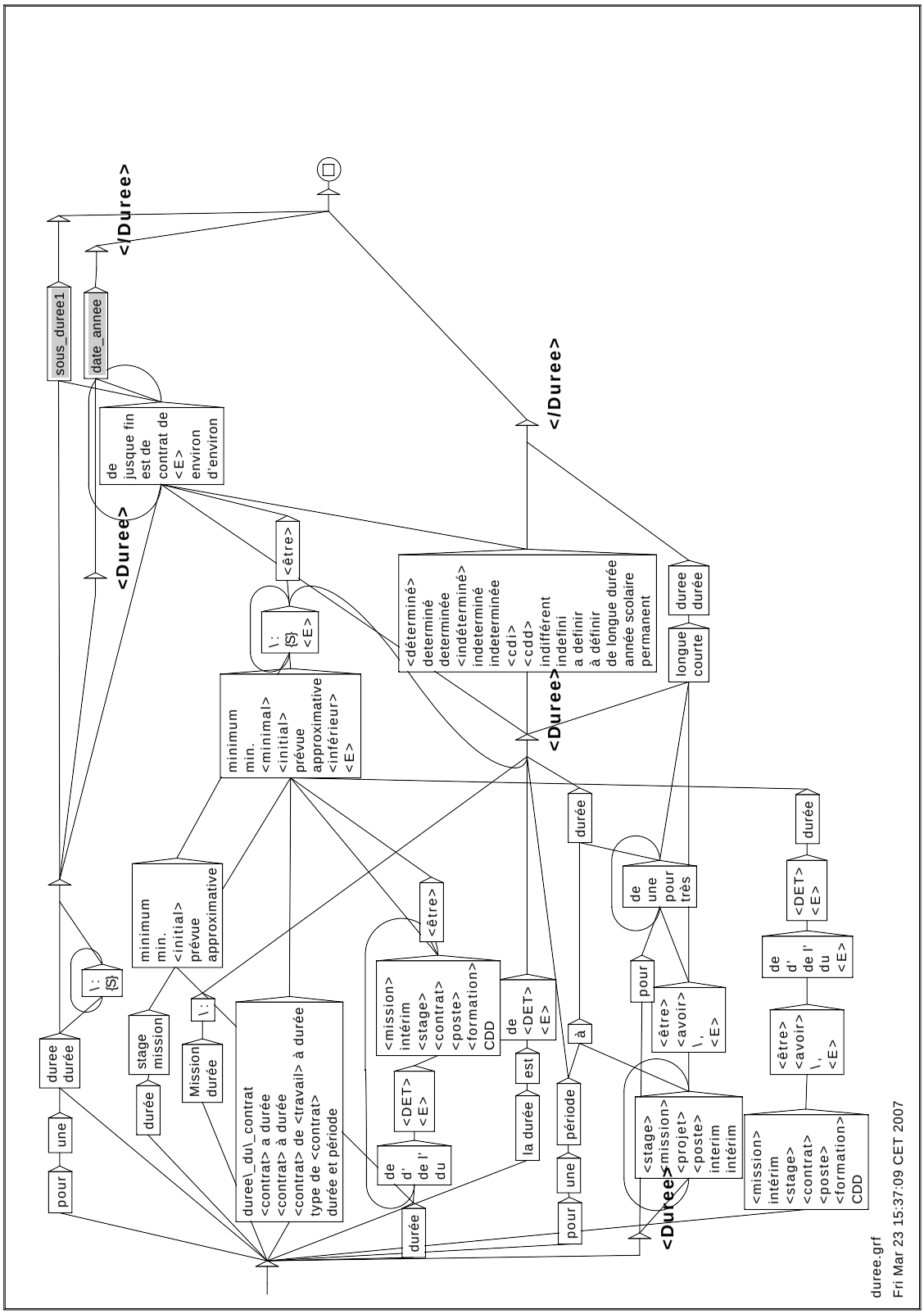


FIG. 8.13 – GL de niveau 1 pour la *durée du poste*

– 5 semaines renouvelables
– 2 mois
– un an
– début septembre
– longue durée
– 6 mois prolongeables sur 9
– 1 à 2 mois minimum.
– du 1. avril 2006 jusqu'à la fin août

8.5 Rémunération

La rémunération proposée pour un poste est une information obligatoire, mais est très souvent omise ou exprimée par des locutions non informatives du genre *« selon profil, à négocier, suivant expérience... »*, car les recruteurs se montrent généralement flexibles face à des candidats intéressants mais non correspondant aux critères initialement souhaités. Ils acceptent par exemple des débutants pour des postes où

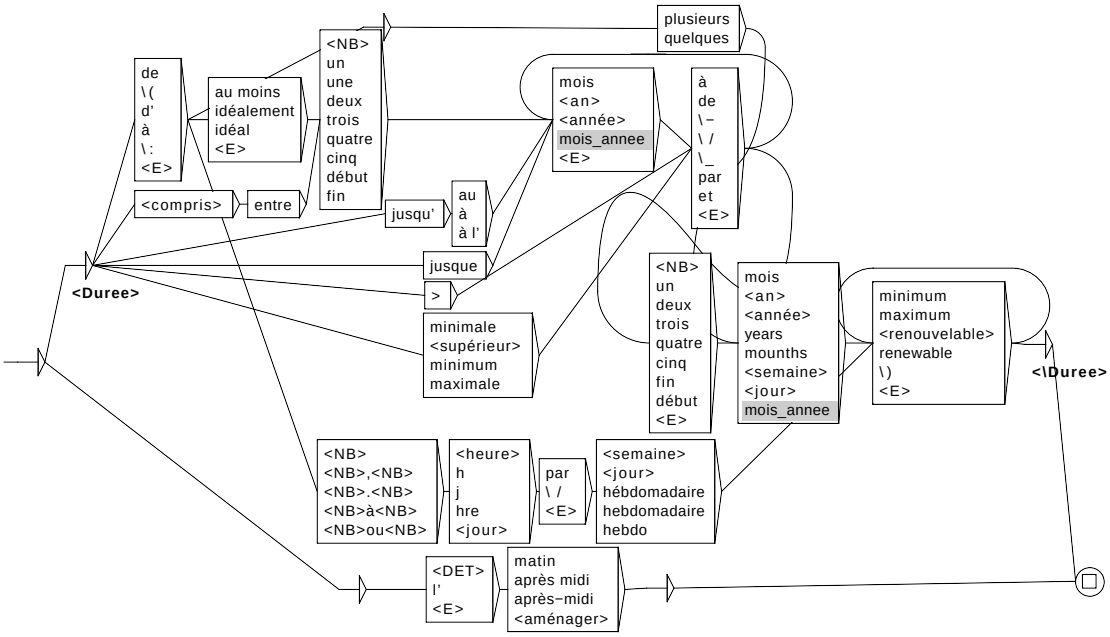


FIG. 8.14 – Typologie interne d'une *Période*

ils ont insisté sur un nombre minimum d'années d'expériences. C'est pourquoi ils préfèrent ne pas fixer une somme pour se laisser la liberté de décider en fonction des désirs du candidat ou de son dernier salaire.

La grammaire 8.15 résume les contextes externes gauche autour de la rémunération, elle reconnaît des concordances du type :

- Indemnités de 500 € / mois
- Rémunération : Selon expérience
- Salaire : Fixe + variables+ frais
- Salaire motivant
- rémunération à négocier en fonction de l'expérience
- Salaire selon profil et expérience entre 28 et 32 KE
- Salaire annuel brut (k€) : 30 - 38 K€
- Salaire annuel entre 40000.00 et 45000.00 € sur 14 mois selon profil
- Salaire entre 23 à 34 k€ bruts annuels selon profil et expérience.
- Salaire : 8,20 - 8,50 EUR /heure + tickets restaurant
- Salaire : 1 500,00 EUR à 3 500,00 EUR par mois Fixe négociable selon profil
- Rémunération : 60 000 / 70 000 € + primes

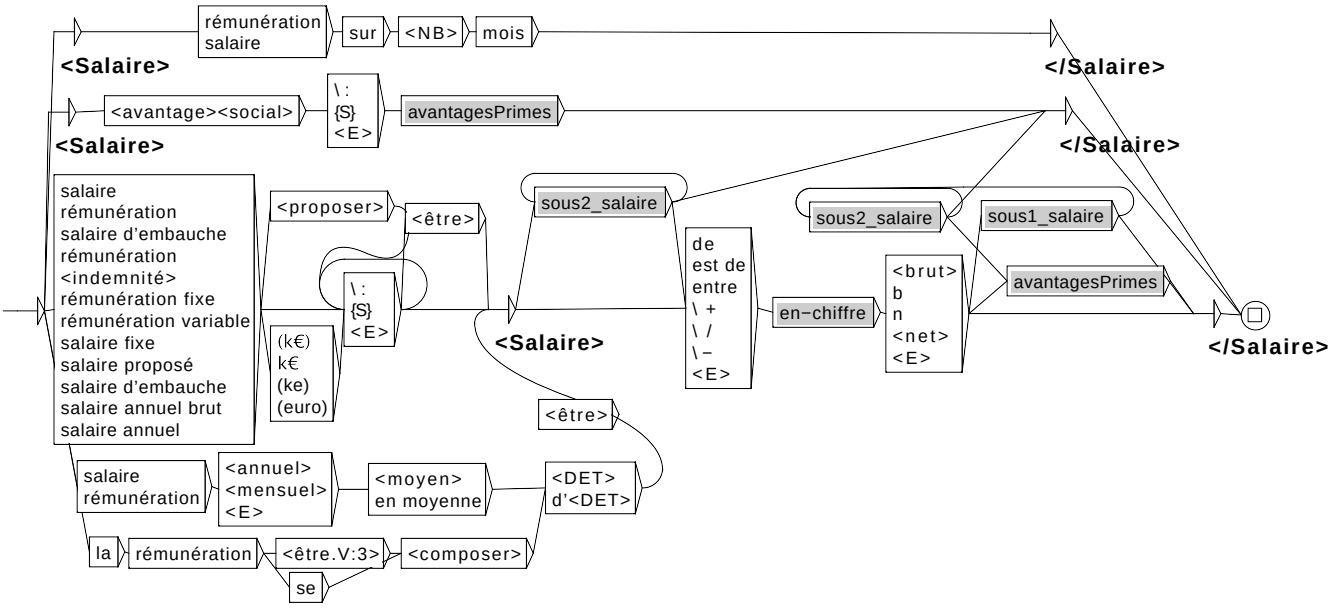


FIG. 8.15 – GL de niveau 1 pour la *rémunération proposée*

Ces exemples montrent la typologie interne variée permettant d'exprimer l'information sur le salaire proposé et pour lesquels nous avons construit des grammaires locales très précises.

8.6 Les Dates

Une des informations primordiales à extraire à partir d'une offre d'emploi est bien évidemment la date de lancement du poste à pourvoir. Ceci pour plusieurs raisons apparentes. La première est de pouvoir gérer les offres dépassées et ainsi les supprimer de la base de données des offres valides, la seconde est de donner la possibilité aux candidats de lancer leurs requêtes de recherche en tenant compte de leurs dates de disponibilité et par conséquent ne recevoir en réponse que les offres commençant à partir de la date désirée.

Une offre d'emploi, peut renfermer plusieurs dates différentes : date de publication de l'annonce, date de disponibilité du poste, date limite de publication. Nous détaillons les contextes internes et externes que nous avons pu recensé tout au long de la phase d'apprentissage des trois dates disponibles dans les annonces d'emploi et montrons pourquoi leur extraction est indispensable au bon fonctionnement de notre système.

Date de prise de fonction

La date de prise de fonction est la date la plus importante pour les candidats, elle est exprimée dans plusieurs contextes différents que nous avons essayé de capturer pendant la phase d'apprentissage et que nous présentons au niveau du graphe (fig. 8.16). Nous montrons dans les exemples suivants certaines concordances reconnues par cette même grammaire sur le corpus de test.

- Date de début : 26 septembre 2007
- Ce poste à pourvoir à partir de septembre 2007
- Poste libre pour la mi-octobre

- Date de prise de fonction immédiate
- Date d'embauche dès que possible
- Ce poste est à pourvoir dans un délai maximum d'un mois

Afin de reconnaître et d'extraire ces dates, nous avons récolté l'ensemble des contextes externes [14] utilisés à cet usage dans les offres d'emploi que nous avons ensuite modélisé sous forme de graphe Unitex (fig. 8.16).

Remarque 1 *Dans ce graphe, nous avons aussi introduit certains contextes externes contenant des fautes d'orthographe, et ce car ce sont des fautes très répandues dans les offres d'emploi et qui ne sont pas corrigées au niveau de la phase de nettoyage.*

Après une analyse d'un certain nombre d'offres d'emploi, nous avons remarqué que la date de lancement du poste est souvent exprimée par une locution temporelle ne précisant pas vraiment la date exacte comme dans les trois derniers exemples cités ci-haut, où *« immédiat, dans un délai maximum d'un mois et dès que possible »* sont des expressions dépendantes d'une information supplémentaire qui est relative à la date de publication de l'offre en question, ce qui signifie qu'elles ne sont compréhensibles que conjointement avec cette seconde information. Dans de tels cas il faut en plus de l'extraction de la date d'embauche, extraire une autre date qui est la date de publication de l'offre. Si cette date n'existe pas alors nous n'avons aucun moyen de savoir si l'offre est actuelle ou pas. Nous avons pu constater que ce problème est courant dans les annonces postées dans les sites carrière des entreprises.

Date de publication de l'offre d'emploi

Comme précisé dans la section précédente, il est nécessaire d'extraire la date de parution de l'offre, pour le cas où la date de prise de fonction n'est pas explicite ou que celle-ci est une locution temporelle faisant référence implicite à la date de parution. Afin d'extraire ces dates, nous avons développé une grammaire locale recensant les contextes externes assemblés à partir de notre corpus d'offres d'emploi

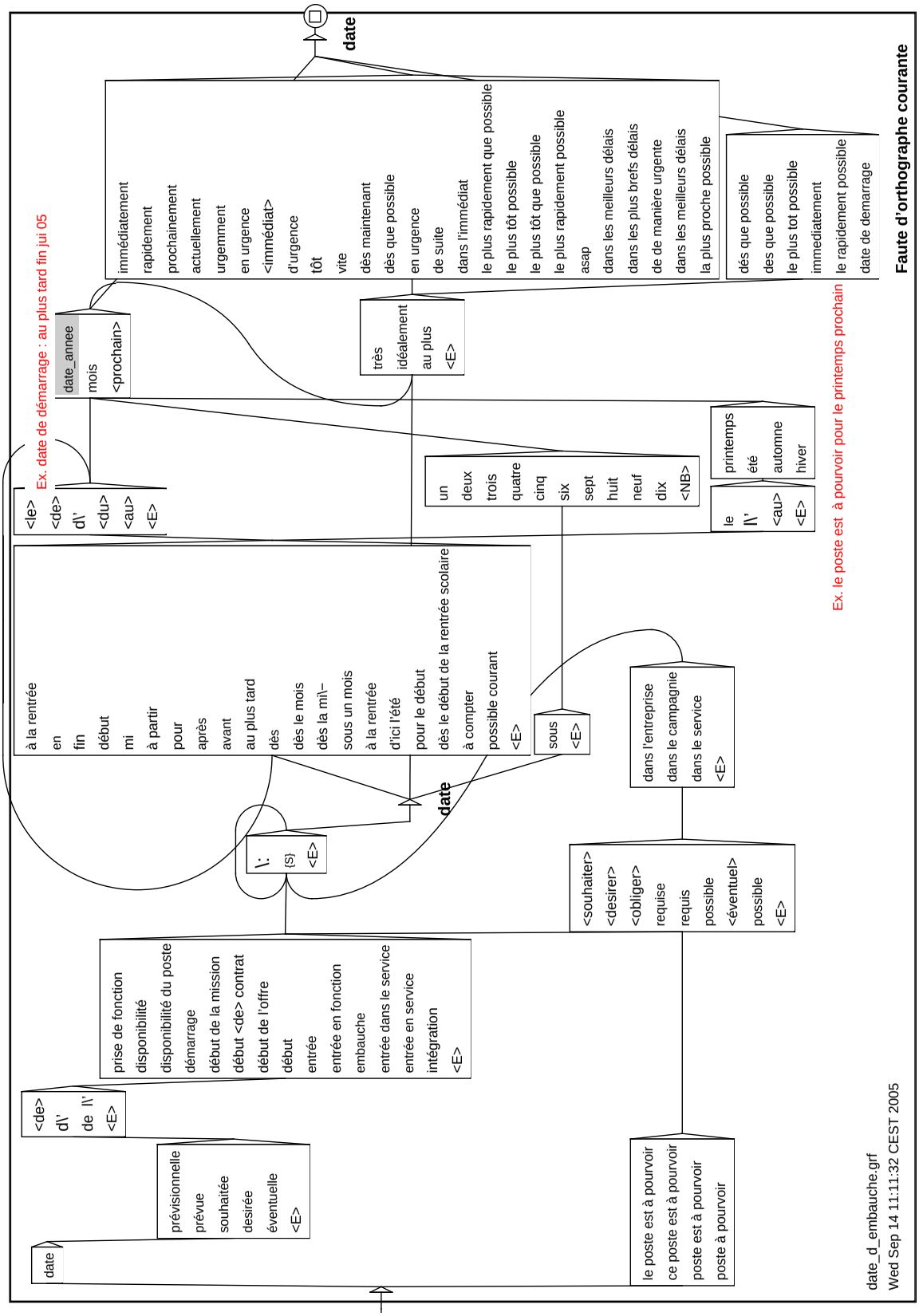


FIG. 8.16 – Date de prise de fonction

et spécifiques à la date de mise en ligne de l'offre. Nous pouvons voir cette grammaire au niveau de la figure (fig. 8.17).

Les concordances associées à cette grammaire sont :

- date de publication de l'annonce :2007-03-24
- date d'émission de l'offre d'emploi est lun 3/02
- mise en ligne : 15-Avril-07
- offre modifiée le mardi 14 Jan. 07
- annonce du 12 02/06
- date de dernière modification : 12.10.2006

Date limite de candidature

Dans certaines annonces de vacation de postes, on exprime également la date limite de dépôt des candidatures, nous avons construit à cet effet une grammaire locale que l'on peut voir au niveau de la figure (fig. 8.18) et qui donne des concordances du type :

- date limite de dépôt de candidature :s 29-Avril-07
- la date finale de réception de dossiers est fin juin
- la date limite pour postuler est le 10.03

Les trois grammaires contextuelles sûres, décrites ci-dessus, sont appliquées en cascade. Si après l'exécution de ces dernières aucune des

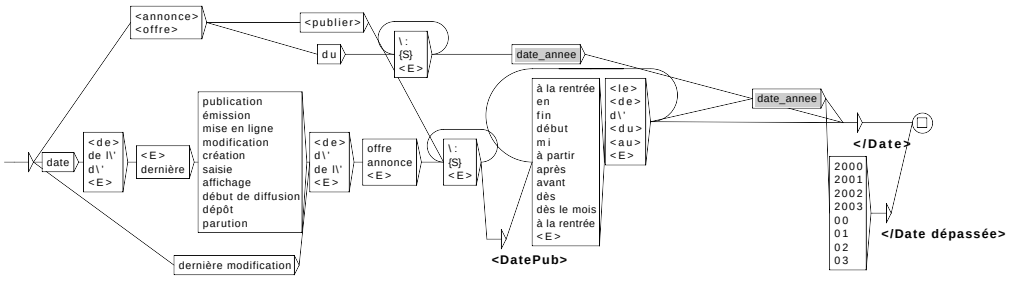


FIG. 8.17 – Date de mise en ligne de l'annonce

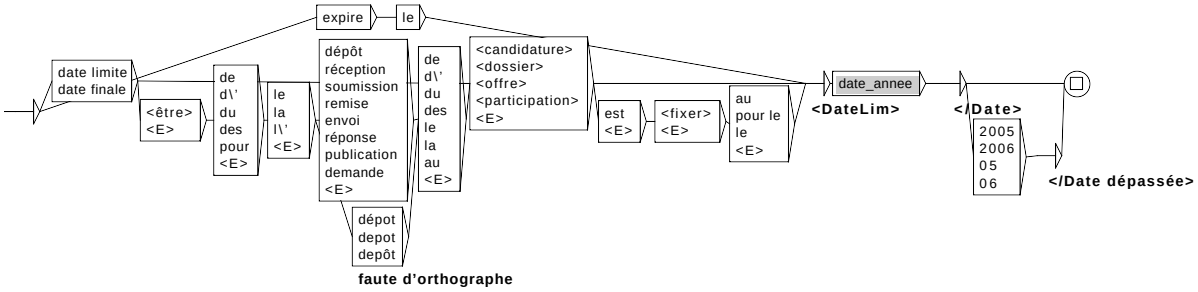


FIG. 8.18 – Date limite de réception des candidatures

dates recherchées n'a pu être trouvée, nous appliquons les grammaires de niveau 2, qui sont des grammaires de reconnaissance de dates hors contextes, décrivant les preuves internes d'une date. Elles permettent de repérer les dates hors contextes ou celles pour lesquelles les grammaires ne couvrent pas encore les contextes droits ou gauches. Dans ce cas, nous interprétons la date reconnue comme étant la« date de mise en ligne de l'annonce ». Nous pouvons voir le graphe correspondant à la description interne d'une date au niveau de la figure (fig. 8.19).

Après observation d'un certain nombre d'offres d'emploi, il s'est avéré courant de rencontrer des offres sur lesquelles aucune date n'est

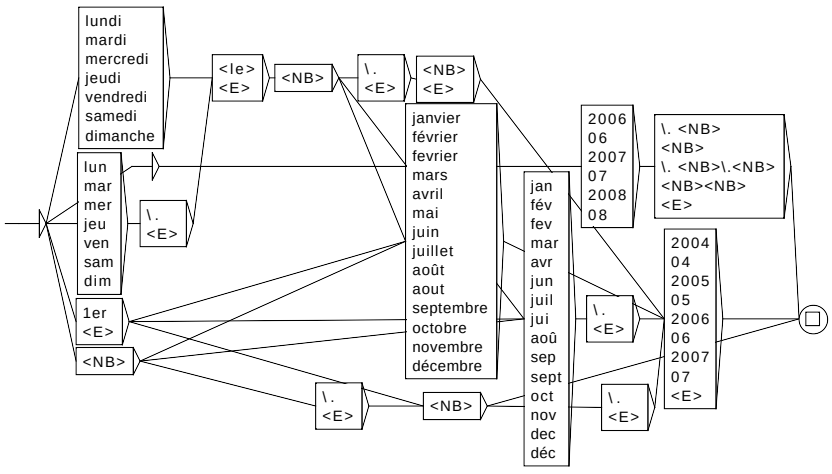


FIG. 8.19 – Description du contexte interne d'une Date

précisée clairement. Ce manque d'information est très désavantageux, aussi bien pour le moteur de recherche qui n'est pas capable de supprimer ces dernières de sa base de données, car n'ayant aucune information explicite sur son actualité, mais aussi pour les recruteurs qui risquent de recevoir des dossiers de candidature longtemps après avoir satisfait leur besoin en ce poste et qui est coûteux en terme de traitement de dossiers.

8.7 Lieu du Poste à pourvoir

Il est primordiale d'extraire, à coté de la date de lancement du poste, le lieu de travail, ceci pour la simple raison que les demandeurs d'emploi sont plus strictes avec le lieu qu'avec l'activité qu'ils souhaitent exercer. En effet les individus à la recherche d'un emploi préfèrent souvent changer la branche que le lieu d'habitation. L'extraction de cette information est à son tour assez complexe, car même si on possède un dictionnaire de toponymes, on ne peut pas simplement extraire ceux-ci sans tenir compte de leurs contextes d'apparition externes, ceci car les toponymes retrouvés dans l'offre d'emploi pourraient aussi bien référer à l'adresse de l'agence de recrutement qu'à l'adresse du siège de l'entreprise, ou encore à l'adresse d'une filiale future et pas nécessairement au lieu du poste à pourvoir. C'est pour cette raison que nous nous sommes concentrés en premier lieu sur l'identification des preuves externes droites et gauches associées à cette information.

Les contextes externes liés à l'entité : lieu de travail

Nous procédons itérativement afin d'extraire l'entité *« lieu de travail »*. Ceci car les offres d'emploi ne sont pas toutes de la même qualité d'une part et ne contiennent pas toutes une multitude d'informations d'autre part. En effet certaines offres sont écrites sous forme de paragraphes homogènes alors que d'autres sont plutôt sous forme de suite de phrases nominales. Ainsi nous avons construit des grammaires locales qui sont lancées en cascade. Chaque itération dépendra du résultat de la précédente.

Il y a dans un premier temps deux sortes de contextes externes à distinguer, les locutions précises comme dans les deux premiers exemples de la liste des concordances qui suit et qui font partie des locutions sémantiques structurelles que l'on trouve dans les annonces semi-structurées et les patrons cachés comme dans les quatre derniers exemples de cette même liste.

- lieu de travail : région parisienne
- lieu dans le nord de paris
- le poste est à pourvoir sur la métropole lilloise
- Le poste en CDI est basé au siège social de McDonald's France à Guy Moquet
- Dans le cadre de notre expansion, nous recrutons sur LOZERE (48) pour notre filiale DELMAS SA
- nous recrutons au sein de notre Agence AUDE ROUSSILLON pour son site de Paris 14ème

Comme nous pouvons bien le remarquer au niveau des derniers exemples, il est indispensable de décrire le contexte comme un tout même si le vrai indicateur dans le cas du 4ème exemple par exemple est *« est basé au »*. Nous ne pouvons pas nous contenter de construire un patron d'extraction dans la 1ère itération tel que :

est basé à <Toponyme> ou <Locution de lieu>

car nous n'aurons pas, dans ce cas, la garantie d'extraire le lieu du poste à pourvoir, comme dans l'exemple 8.1 où ce patron permet de reconnaître le lieu du siège de l'entreprise qui embauche et non pas le lieu du poste lui même.

(8.1) dont le siège est basé à Zurich.

Nous avons ainsi développé des grammaires locales de niveaux différents que nous décrivons dans la suite.

Dans la grammaire de la figure (fig. 8.20) nous avons collecté les

différentes expressions figées ou semi-figées employées dans notre sous langage d'offres d'emploi et faisant référence au lieu du poste.

Ville :
Poste basé à
Poste à pourvoir à
Lieu du poste :
Lieu de travail
Mission située dans
...

Ce graphe sera lancé en premier dans le processus itératif d'extraction. Dans le cas où aucun lieu n'est reconnu, nous passons à la seconde itération avec les grammaires locales des cas spéciaux de la figure (fig. 8.21).

- Localisation : <Lieu> PAU 64 </Lieu>
- poste basé près de <Lieu> Gent </Lieu>
- Région : <Lieu>France -île de France </Lieu>
- poste est basé à <Lieu> Lens </Lieu>
- Lieu : <Lieu> France -Lorraine -Luxembourg </Lieu>
- Poste basé à <Lieu> Vitry (94) </Lieu>
- poste est à pourvoir sur <Lieu> Boulogne Billancourt </Lieu>
- Poste à pourvoir sur <Lieu> la région lyonnaise </Lieu>
- Région : <Lieu> 94 </Lieu>
- poste étant basé en <Lieu> Allemagne </Lieu>
- Poste à pourvoir dans le <Lieu>42 - FEURS </Lieu>

Nous observons à ce niveau des erreurs de reconnaissance tel que :

- Poste basé <Lieu>à Ivry </Lieu> S/Seine (5 mn de Paris par RER C)
- Région : <Lieu> Cergy </Lieu>, IDF 95800
- Postes basé <Lieu> en France </Lieu>, (Paris-La Défense (Headquarter))
- Ville de <Lieu> Garges </Lieu> les Gonesse
- Région :<Lieu> France -IDF </Lieu>-LE VESINET, CHATOU, LE PECQ

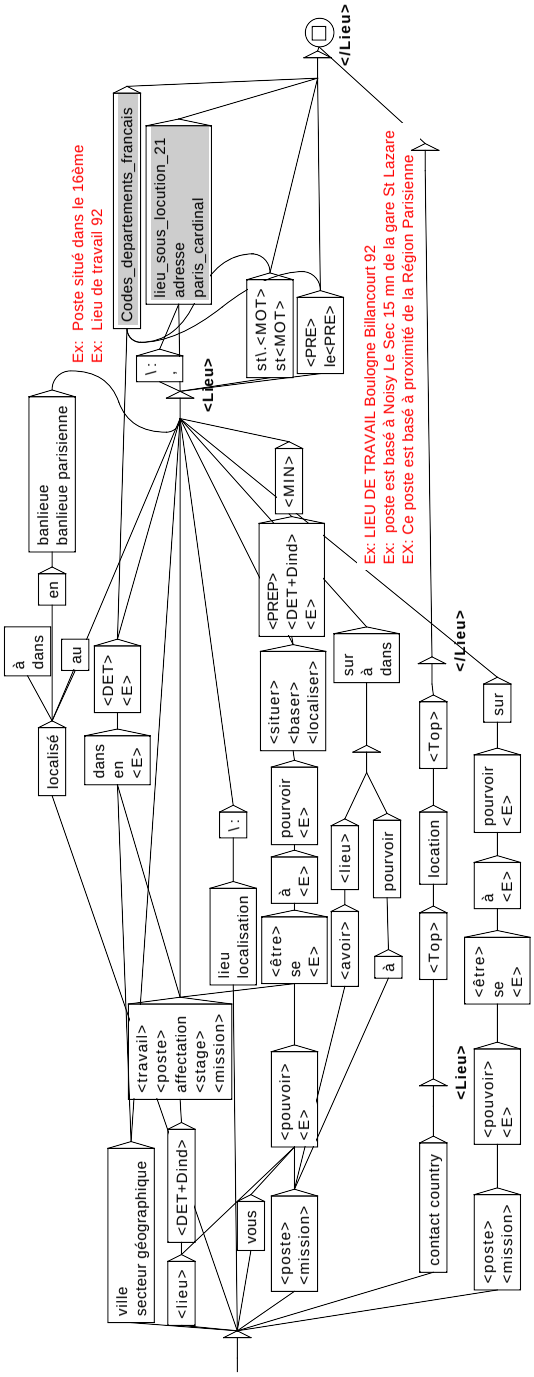


FIG. 8.20 – Grammaire de niveau 1 : locutions locatives

- Poste basé <Lieu> à Villefranche </Lieu> s/Saône (69)
- Poste basé sur notre <Lieu>agence de Vitrolles </Lieu>
- recherche pour l'un de ses clients aéronautique basé à <Lieu> Bouguenais un PEINTRE </Lieu>

Les Contextes internes liés à l'entité : lieu

Pour ce qui est des contextes internes liés au lieu du poste, nous avons décrit, au niveau des graphes (fig. 8.22, 8.7) les différentes grammaires locales possibles employées dans un sous langage comme le notre. Nous énumérons ci-dessous un ensemble d'exemples de ces contextes internes.

- Lieu de la mission 69006 LYON
- Poste basé en région parisienne nord 95
- Poste est basé à Noisy Le Sec 15 mn de la gare St Lazare
- Lieu de travail PARIS 1ER ARRONDISSEMENT
- Poste basé dans le Nord 50 KM au nord de Lille
- Poste basé sur la métropole lilloise
- Lieu : Clichy (92)

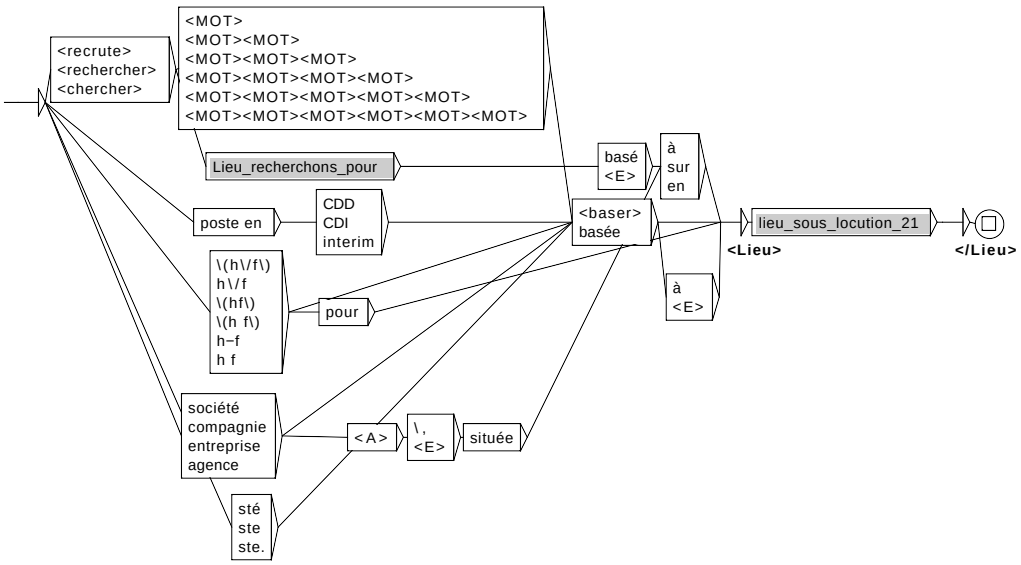


FIG. 8.21 – Grammaire de niveau2 : Locutions locatives

– Region 27100
– poste basé dans le sud est de l'ile de France

Nous disposons du dictionnaire des toponymes développé dans le projet Prolex [1] ; cependant cela ne semble pas être suffisant dans un contexte comme le notre où les écrits comprennent très souvent des fautes d'orthographe ou alors des Noms propres écrits en minuscule. Une des règles d'écriture en langue française impose que la première lettre des Noms Propres soit écrite en majuscule, ce qui est d'ailleurs repris dans le dictionnaire des Toponymes. D'autre part ce dictionnaire bien qu'il soit d'une taille considérable, n'est en aucun cas complet, c'est pourquoi nous retrouvons dans nos grammaires locales, des chemins comme :

– *<poste><être><baser><PREP><MOT>*
ou alors
– *<mission> <baser> <à> (st/st./saint)<Mot>*

[1] http ://tln.li.univ-tours.fr/Tln_Unitex.html

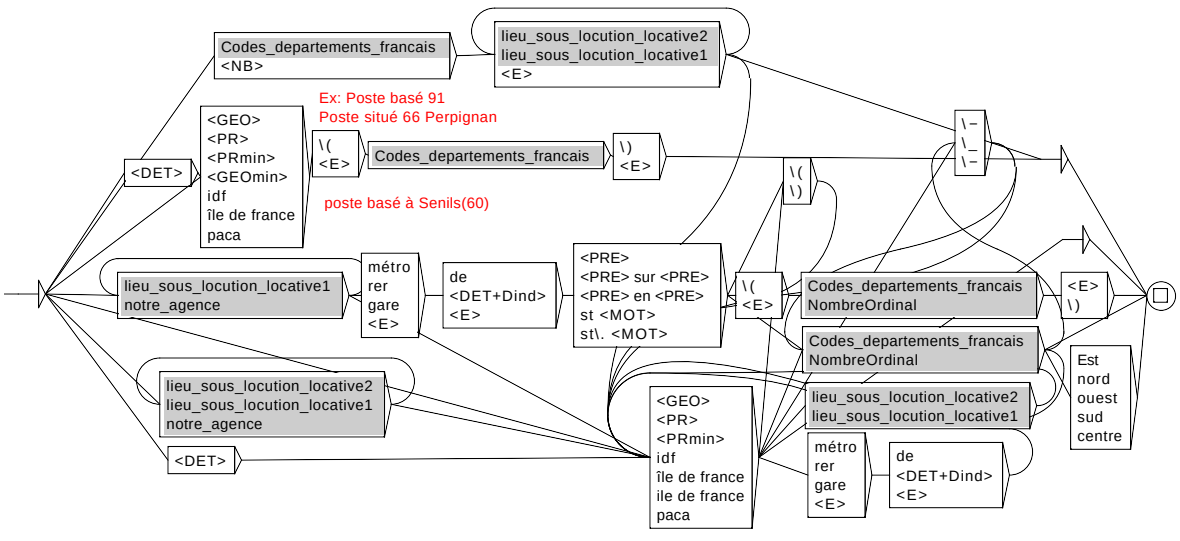
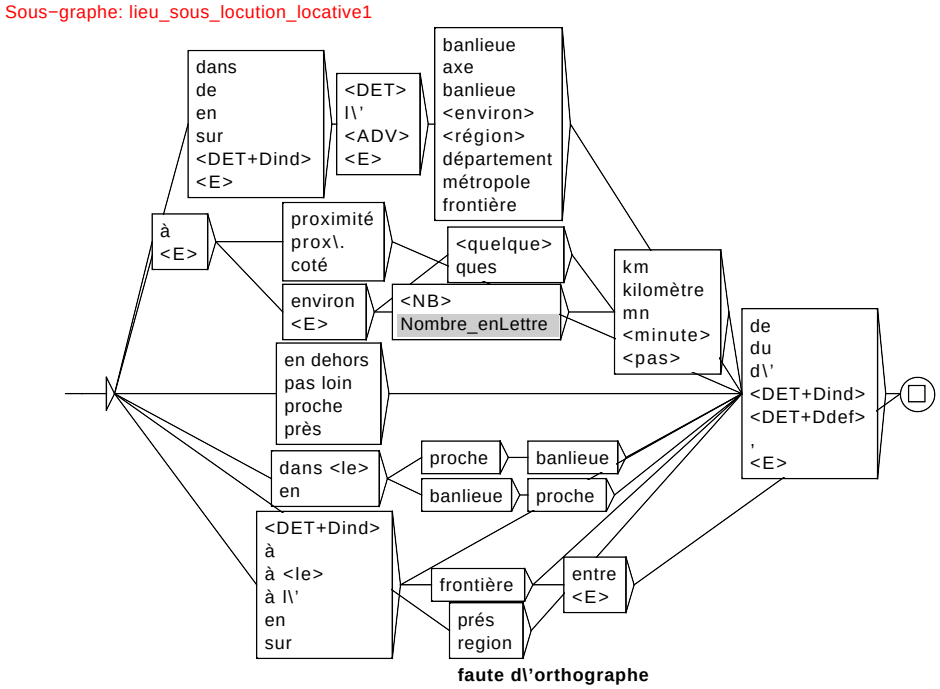


FIG. 8.22 – Grammaire locale des Contextes internes du lieu



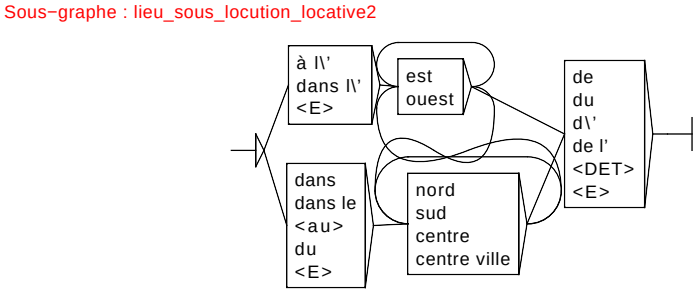


FIG. 8.23 – Sous-graphes : Lieu_sous_locution_locative 1 et 2

Ces deux cas provoquent parfois des reconnaissances partielles du lieu du travail, mais dans ce cas nous avons opté pour le choix des concordances les plus courtes car engendre un taux d'erreur moins important que le choix de la concordance la plus longue. Une seconde remarque est à signaler pour le cas où la locution se termine par *« Est »* comme dans *« paris Est »*, nous imposons alors que ce lexème soit en majuscule, la perte d'information est minimale par rapport au bruit que cela engendrerait de l'admettant en minuscule, où la variation morphologique du verbe être est alors le plus souvent reconnue.

Dans le paragraphe suivant nous décrivons les grammaires locales d'une adresse postale. Ce choix est lié au fait que très souvent le contexte interne du lieu de travail est représenté par l'adresse complète de l'entreprise qui propose le poste.

Les erreurs observées

- ville <Lieu>1mn de la gare</Lieu> Qualifications.
- Hôtel de Ville rue Jean Jaurès BP 63 59264 Onnaing VILLE DE CELLES SUR BELLE.

Description grammaticale des Adresses

Le lieu du poste est souvent exprimé par l'adrese complète de l'agence ou de l'entreprise en besoin de personnel. Pour ce cas nous avons développé des patrons d'extraction d'une adresse postale dans les offres d'emploi et qui vient dans un premier temps accompagner les grammaires contextuelles gauches décrites ci-dessus. Cependant la reconnaissance de l'adresse postale est une information qui peut nous servir pour répondre à d'autres besoins informationnels comme :

- si aucun lieu n'a été trouvé, l'adresse de l'entreprise est considérée comme répondant directement à ce besoin
- elle peut représenter l'endroit auquel les candidats doivent envoyer leurs dossiers de candidatures

Une Adresse française peut s'écrire sous différents formats comme on peut le voir au niveau des exemples suivants :

- 45, rue des bertauds, 93110 Rosny s/s Bois, France
- Route de Bellegarde 74330 SILLINGY
- BP 30, 41451 marne la vallee cedex 15
- 1ère avenue, 9ème rue BP 121 06513 Carros Cedex.

La première étape dans la construction des grammaires locales des adresses fut de rassembler tous les odonymes. En français, une adresse est certainement introduite par un odonyme comme *« rue »*, *« Place »* ou encore *« Boulevard »*. Nous en avons récolté une centaine en y incluant également les abréviations souvent rencontrées dans notre corpus. Cette liste est disponible à l'annexe. Après avoir obtenu une liste assez significative, nous avons procédé par la méthode de *Bootstraping* énoncée par [36] et utilisée dans les différentes étapes de nos travaux.

Cette méthode s'est avérée très puissante pour accroître et perfectionner itérativement les grammaires locales. Il suffit dans cette méthode de prendre le Mot qui nous intéresse et autour duquel nous voulons construire une grammaire locale, d'observer ensuite les contextes gauches respectivement droits dans lesquels il évolue et reconnaître ainsi les diverses preuves externes ainsi que les preuves internes. Afin d'être sûr de la complétude de nos graphes nous avons aussi essayé de commencer les itérations par d'autres Mots significatifs comme des contextes externes sûr que nous avons déjà observé.

Au cours du processus d'extraction de l'adresse nous avons aussi formé un dictionnaire des noms des rues que nous augmentons au fur et à mesure de nos extractions et vérifications.

La grammaire locale d'extraction des adresses postales est décomposée en deux sous-grammaires indépendantes. Nous pouvons les observer dans les deux figures (fig. 8.24) et (fig. 8.25).

Une adresse peut être écrite de différentes manières :

1. Commençant par le numéro de la rue
   <NB>-<NB>,<odonyme><NRue>,<NB><NVille><NPays>
2. Commençant par un odonyme comme *Place*
3. Commençant par BP.
4. un mélange des deux formats précédents

Voici quelques concordances extraites du corpus test décrit plus haut.

- <Adresse> Immeuble Péricentre - Rue<Rue> Van Gogh Villeneuve d'Ascq Cedex 59658 </Adresse>
- <Adresse> Immeuble Ordinal -<Rue> Rue des Chauffours Cergy-Pontoise France 95002 </Adresse>
- <Adresse> 67 boulevard<Rue> Malesherbes Paris 75008 </Adresse>

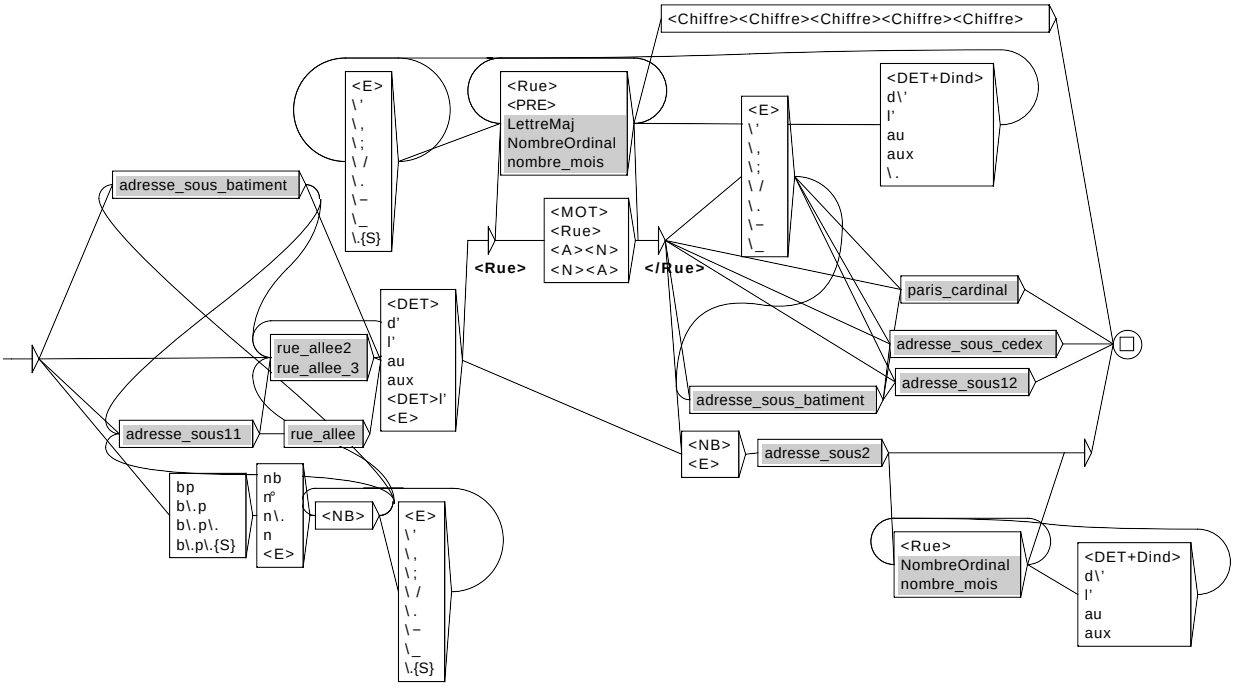


FIG. 8.24 – Grammaire Locale d'une adresse postale (1er type)

- <Adresse> 2-6 rue des<Rue> Bourets</Rue> SURESNES FRANCE 92150 </Adresse>
- <Adresse>BP 76209 45062 Orléans Cédex 2 </Adresse>
- <Adresse> 15, avenue de l'<Rue>Europe- BP</Rue>60016 Schiltigheim Strasbourg cedex France </Adresse>
- <Adresse> 46/52 Rue<Rue> Albert</Rue> Paris 75013 </Adresse>

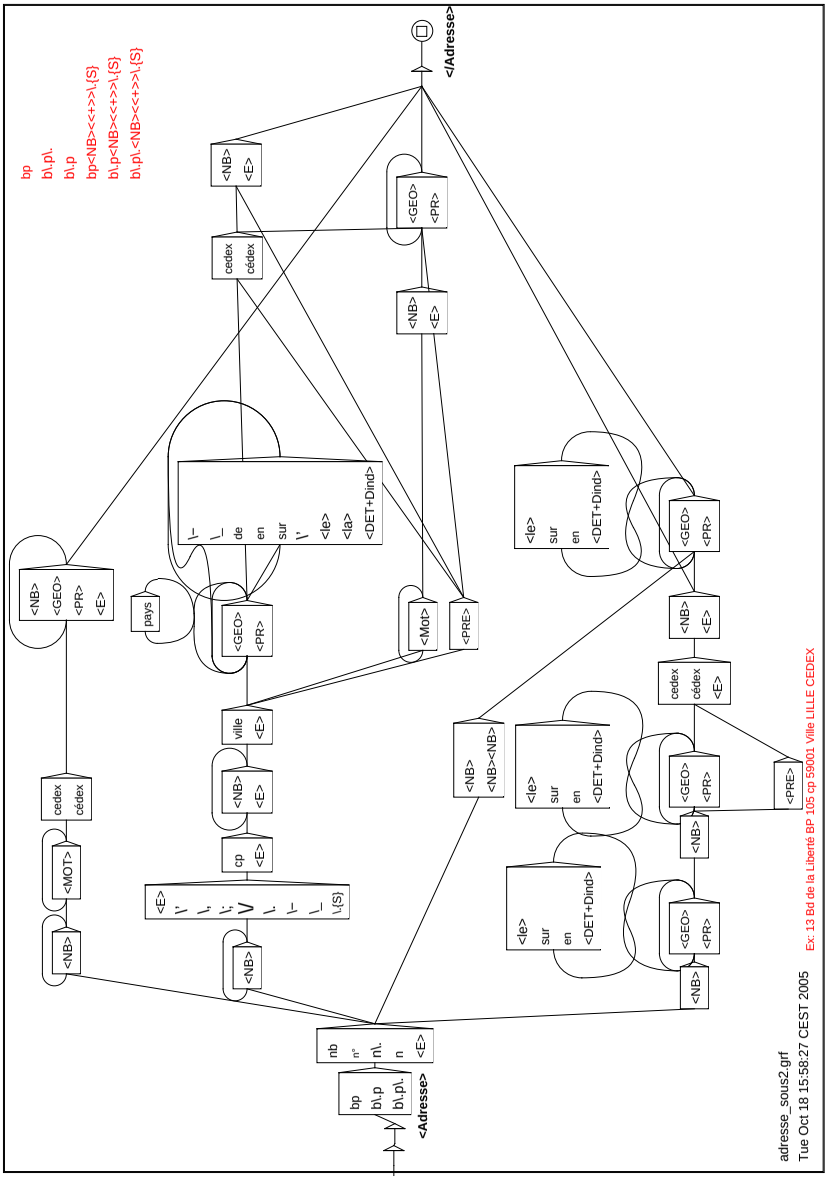


FIG. 8.25 – Grammaire Locale d'une adresse postale(2ème type)

- <Adresse> 62 bis avenue<Rue> André Morizet Boulogne Billancourt 92100 </Adresse>
- <Adresse>28 ROUTE DE<Rue> BAYONNE BILLERE,FRANCE 64140 </Adresse>
- <Adresse>37, boulevard de<Rue> Montmorency</Rue> - 75781 Paris Cedex 16 </Adresse>
- <Adresse>110/114 rue<Rue> Victor Hugo Levallois</Rue> Perret92300 </Adresse>
- <Adresse> Zone industrielle de<Rue> Reyrieux Reyrieux</Rue> France 01600 </Adresse>
- <Adresse> Z.I. de l'<Rue>Abbaye PONT EVEQUE,RHA 38200 </Adresse>

8.8 Les domaines d'activité

Le domaine d'activité est une information difficile à délimiter car peut s'exprimer de différentes manières et qu'il est impossible de recenser tous les noms de domaine dans un dictionnaire. Nous avons alors essayé d'extraire manuellement les locutions types rencontrées dans les corpus d'apprentissage et avons généré à partir de ceux-ci les automates du texte correspondants. A partir de là, nous avons généré des grammaires locales précises dans le but de restreindre ultérieurement la recherche à cet ensemble de grammaires et non pas récupérer, par exemple, tous les mots en majuscule qui suivraient le mot *« domaine »*. Un tel patron engendrerait un taux d'erreurs très élevé, de plus on ne peut pas se fier à l'utilisation correcte des règles typographiques dans un sous-langage comme le notre bien qu'il s'agisse de textes reflétant directement l'image de l'entreprise face aux futurs employés. Celles-ci rédigent souvent leurs annonces d'emploi sans porter de l'attention aux fautes d'orthographes et au respect des règles typographiques, tandis que les mots ordinaires voulant être mis en relief et en évidence sont écrits en majuscule, les noms propres qui doivent être en majuscule ne le sont souvent pas.

Comme pour le nom du poste, le domaine d'activité peut être introduit par des locutions types rendant compte du degré de structuration de l'annonce. Au vue de cette observation, nous recherchons, dans une première itération, à reconnaître les domaines exprimés dans les offres plus ou moins structurées (fig. 8.26), pour lesquelles le taux de recon-

naissances erronées ou partielles est minimal.

- Catégorie de l'offre : <Domaine> Ressources Humaines </Domaine>
- Catégorie du poste : <Domaine> Informatique-Software/PAO-CAO/Editeurs </Domaine>
- Catégorie de la mission : <Domaine> Marketing/Communication/Publicité/RP </Domaine>
- Filière : <Domaine> Finances/Contrôle de gestion </Domaine>
- Domaine du poste : <Domaine>Chimie </Domaine>
- Domaine du poste : <Domaine>Commercial - Ventes </Domaine>
- Secteur du poste : <Domaine>production de produits cosmétiques </Domaine>
- Secteur de l'offre : <Domaine>Assurance/Banque </Domaine>

Dans une seconde itération nous nous concentrons à l'étude des évidences externes qui viennent introduire les domaines d'activité dans les textes des annonces d'emploi. Nous avons pu recenser des séquences comme celles disponibles dans la liste suivante et dont les grammaires locales sont résumées dans la figure (fig. 8.27).

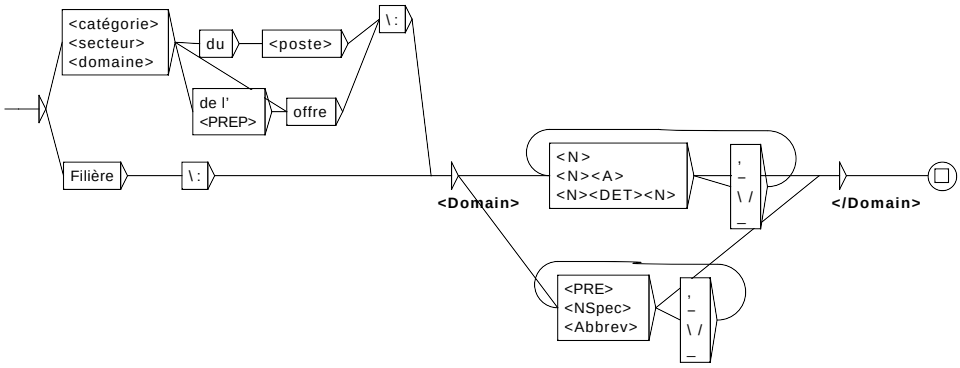


FIG. 8.26 – Locutions types dans les offres semi-structurées

dans <le> <domaine>
spécialisé en
spécialiste <du>
société du domaine
société du secteur
dans le cadre de nos activités en
sur le secteur
nous sommes spécialiste
leader <PREP>
au sein de l'équipe
....

Il est important d'insister sur le fait que ces grammaires locales peuvent répondre à trois besoins informationnels distincts : ① le domaine d'excellence du candidat souhaité, ② le domaine d'activité de l'entreprise en besoin de personnels et ③ le secteur d'activité du poste à pourvoir. En ce qui nous concerne, nous classifions toutes les concordances reconnues à travers les grammaires de la seconde itération comme représentant le domaine d'activité de l'entreprise, bien que ce

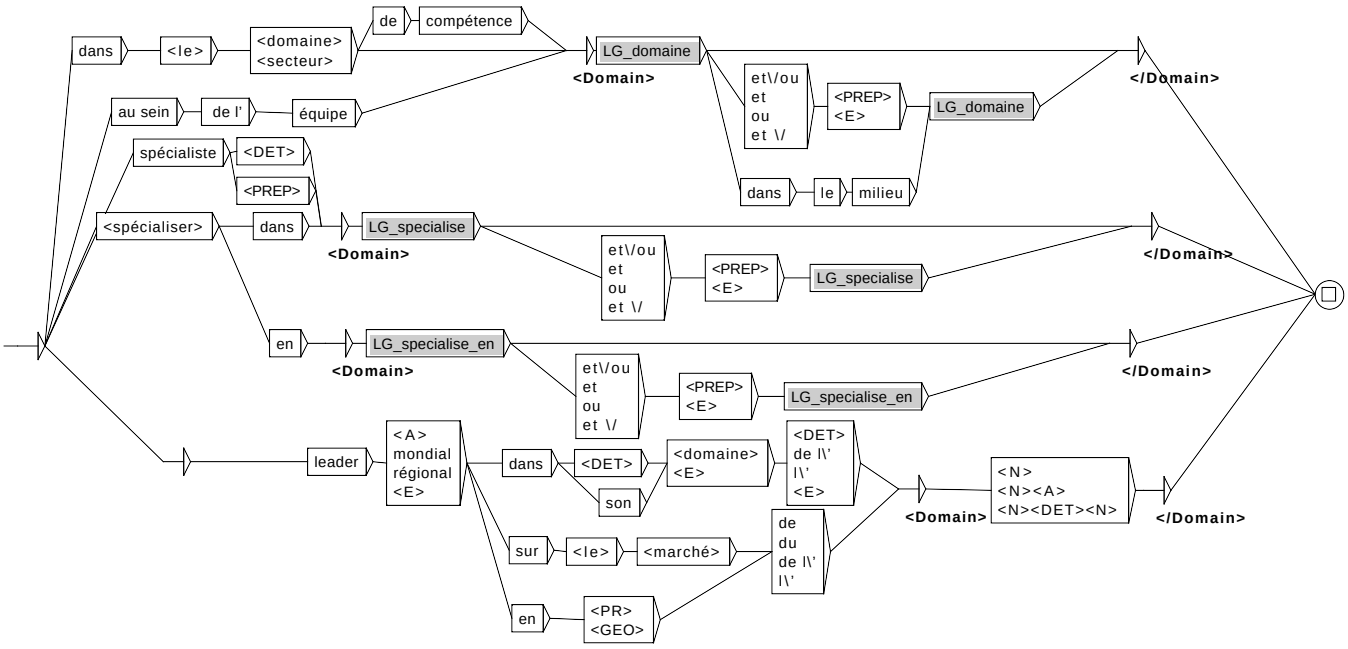


FIG. 8.27 – Patrons d'extraction des domaines d'activité : Itération 2

ne soit pas une heuristique complètement fiable. Prenons le cas d'une entreprise de publicité qui recherche un « cuisinier pour la cantine interne », dans ce cas le domaine d'activité de l'entreprise ne joue pas un rôle très important par rapport à celui du poste qui peut s'exercer en dehors des boîtes de publicité.

Nous avons procédé par *« Bootstarping »* afin d'identifier les différentes typologies internes des domaines d'activité et qui représentent les sous-graphes appelés dans la grammaire de la figure (8.27). Nous présentons dans la liste suivante un certain nombre d'exemples construit à travers la phase d'apprentissage sur un corpus de plus de 20.000 offres d'emploi et en rapport avec le contexte externe gauche : *« dans le domaine »* :

Dans le domaine ...
- <A>
  Ex : dans le domaine aéronautique
- <DET+Dind><N>
  Ex : dans le domaine de l'aéronautique
- <DET+Dind><N><PFX><A>
  Ex :dans le domaine de l'hygiène bucco dentaire
- <DET+Dind><N><A>/<A>
  Ex :dans le domaine de la chimie organique / analytique
- <DET+Dind><N><DET+Dind><N>
  Ex :dans le domaine de l'industrie des transports
  <DET+Dind><N><PREP><N>
  Ex :dans le domaine de la maintenance en agroalimentaire
  <DET+Dind><N><DET+Dind><N><PFX>
  Ex :dans le domaine de la prévention des maladies cardio
- <DET+Dind><N><DET+Dind><N><N>
  Ex :dans le domaine automobile et/ou en conception document

De même pour chacune des fonctions présentent plus haut, nous avons construit des grammaires locales nous permettant d'extraire les concordances suivantes :

- Leader
  leader dans le domaine des <Domain> ustensiles de cuisine </Domain>
  leader mondial dans les domaines de l'<Domain> aéronautique </Do-

main>
leader sur les marchés du <Domain>chauffage </Domain>

– spécialisé
spécialisé en <Domain> nouvelles technologies </Domain>
spécialisée en <Domain> comptabilité et en finance </Domain>
spécialisée dans <Domain> la conception microélectronique </Domain>
spécialisée dans <Domain> le financement locatif </Domain>

– domaine
dans les domaines <Domain> de la chromatographie et de la spectroscopie </Domain>
dans le domaine <Domain> des Télécommunications </Domain>
dans le domaine <Domain> audiovisuel </Domain>

– secteur
dans les secteurs <Domain> civil et parapublic </Domain>
dans le secteur <Domain> des vêtements de travail </Domain>
dans le secteur <Domain> électronique grand public </Domain>

– équipe
Au sein de l'équipe <Domain>Back office </Domain>
Au sein de l'équipe <Domain> R&D </Domain>

Nous avons également construit un dictionnaire de noms de domaines, que nous augmentons à fur et à mesure de nos extractions. Ce dernier est malheureusement loin d'être exhaustif et son utilisation directe hors contexte génère un taux d'erreurs très élevé. Ce taux élevé de concordances erronées ou partielles s'explique par le fait que les mots simples et composés classifiés dans la catégorie sémantique « Domaine » sont également présents dans d'autres classes sémantiques et syntaxiques et peuvent répondre à un autre besoin informationnel quand ils apparaissent dans des contextes différents. Prenons l'exemple de l'entrée *« Télécom »* dans le dictionnaire des noms de domaines.

1. vous êtes dédié à un compte client majeur du domaine télécom
2. Les technologies télécom n'ont plus de secret pour vous
3. les services aux entreprises (marketing, info-télécom, services généraux, RH)
4. Dans le respect du cahier des charges France Télécom

Il est très important pour nous au vue de ces exemples de reconnaître puis de décrire tous les contextes permettant d'introduire le domaine d'activité dans les offres d'emploi pour ainsi améliorer les performances de notre système d'extraction. Dans les exemples suivants, nous montrons certains contextes observés dans des offres d'emploi et non encore introduits dans nos grammaires. Dans le premier exemple, il s'agit d'un contexte très particulier, qui n'est pas très récurrent, mais il est indispensable d'ajouter cette entrée dans nos grammaires si l'on tend à augmenter le taux du rappel de notre système.

- notre société doit sa forte notoriété à la reconnaissance de ses produits par les professionnels de l'automobile et de la moto
- Dans le cadre d'une mission pour un de nos clients à Luxembourg (Grande Banque de la place)
- Bull Services & Solutions Régions propose une offre complète de services pour accompagner les entreprises dans la réalisation de leurs projets

8.9 Les Coordonnées de l'entreprise

Une entreprise en besoin de personnels, rédige une offre d'emploi, il lui reste ensuite le choix de la publier directement sur son site carrière ou sur l'un des espaces réservés à cet effet ou alors de déléguer la tâche de publication et de recherche du candidat idéal à une entreprise de communication RH qui se chargera alors de toutes les étapes du recrutement entre la première phase de publication de l'annonce jusqu'au choix final du futur employé. Dans le second cas, le nom et les coordonnées du recruteur sont cachés, les candidats sont amenés à postuler en passant par l'agence de recrutement en ignorant l'identité du futur employeur, tandis que dans le premier cas, les recruteurs prennent le soin de décrire leur entreprise, son activité principale, sa position sur le marché, ses objectifs pour attirer les candidats. En effet, l'image de l'entreprise joue un rôle très important dans le choix d'un emploi. Les candidats donnent beaucoup d'importance à la relation de leur futur employeur avec ses employés, sa notoriété, son bon état financier, sa situation prépondérante ou pas sur le marché. Le choix d'un poste se fait très souvent en premier lieu en fonction des avantages offerts par

l'entreprise avant même de s'intéresser aux tâches qui y seront déléguées.

Nous nous sommes donc concentrés sur la tâche d'extraction des coordonnées de l'entreprise à partir de chaque offre d'emploi. Nous remplissons ainsi une base de données d'entreprises en renseignant les attributs : Nom, adresse postale, numéro de téléphone, numéro de fax, page d'accueil et E-mail. Une telle base de données peut s'avérer très utile pour élargir la gamme des services offerts par notre système et qui pourrait être très utile à des candidats voulant se renseigner sur certaines entreprises dans lesquelles il y aurait des postes intéressants.

Malheureusement, les coordonnées des entreprises ne sont pas toujours explicitées dans les offres d'emploi, soit car le poste est publié par une agence de recrutement ou alors car il est posté directement sur le site carrière de l'entreprise même. Les sites carrières étant conçus pour être lus par des utilisateurs humains, capable d'analyse et de raisonnement par analogie, les recruteurs dès lors ne se soucient pas de préciser leurs coordonnées, considérant que toute personne peut déduire que celles-ci sont disponibles sur la page de la classe « Contact »du même site Web.

Nous présentons dans la suite de cette section, le détail de reconnaissance des 7 informations représentatives des coordonnées d'une entreprise que l'on peut trouver dans le texte d'une annonce de vacation de poste, à savoir :

- Nom de l'entreprise
- Adresse postale
- Téléphone
- Fax
- URL de la page d'accueil
- Email

et auxquels nous ajoutons le nom du contact à qui il est nécessaire d'adresser les dossiers de candidature.

8.9.1 Les Noms d'organisations

L'extraction des noms d'organisations est indispensable dans notre système d'offres d'emplois. Elle est presque aussi importante que l'extraction de la fonction à pourvoir, et ce, car l'organisation choisie pour postuler est aussi décisive que le poste lui même. En effet, les candidats attachent beaucoup d'importance à l'image de l'entreprise, sa taille, sa relation avec ses employés et sa relation avec l'environnement dans leur prise de décision.

La reconnaissance automatique des noms d'organisations n'est pas une application récente, elle est sous-tâche du domaine de l'extraction automatique d'entités nommées, méthode décrite au chapitre (4). Tous les systèmes décrits ont cependant un aspect commun : le corpus de travail est dans la plupart des cas une collection de dépêches journalistique et plus particulièrement la rubrique économique de ces dernières. Dans notre cas, de textes spécialisés, les auteurs usent d'autres preuves externes et d'autres locutions pour introduire les noms propres et les noms d'organisations en particulier.

F. Schmidt [61] se concentre dans ses travaux de thèse en 2004 sur la reconnaissance automatique des noms d'organisations dans le sous-langage particulier de l'actualité des affaires en langue anglaise. Elle obtient alors des taux de F-mesure supérieurs à 95% sur 5 corpus collectés à partir de 5 journaux différents (New York Times, Wall Street Journal, Reuters, ...). F. Schmidt avait construit 113 grammaires locales au total pour la reconnaissance de la structure interne des noms d'organisations.

N.Friburger a consacré aussi une partie de sa thèse à l'extraction des noms d'organisations, elle constituait en 2002, 35 grammaires locales pour l'identification de ces dernières dans un corpus journalistique en langue française.

D'après cette dernière dans [25] « 50% des noms d'organisations sont accompagnés d'une preuve interne, car sont, pour la plupart, des noms propres à base descriptive. »

Cette dernière rend compte aussi que certains mots sont à la fois

des preuves internes et des preuves externes, comme on peut l'observer dans les exemples suivants :

- la banque **Société** Générale
- la société Parisbas
- la **banque** du sud

Pour notre part, à côté de la constitution d'un dictionnaire de noms de compagnies collectés à partir de listes disponibles sur le Net et à partir des multiples extractions occasionnées tout au long de la phase d'apprentissage. Nous avons également construit des patrons d'extraction mettant en oeuvre les différents descripteurs, preuves internes et preuves externes que nous avons rencontrées dans nos divers corpus d'apprentissage. Ces grammaires permettent d'augmenter le dictionnaire *CPN-dic* au fur et à mesure des extractions sur de nouveaux textes d'annonces d'emploi. Un échantillon du dictionnaire *CPN.dic* serait :

ADITO RH,.CPN
ADLER & ALLAN LTD,.CPN
ADMBassereau,.CPN
Aenix,.CPN
AM'TECH médical,.CPN
Château Jacques Blanc SCEA;.CPN
Christian Boulet et Compagnie SA,.CPN
Pêcheries les Brisants SARL,.CPN
Pemex,.CPN
Pierre Bourquin Imprimerie,.CPN
Robert Bosch France SA,.CPN
Scierie Bartel et Fils,.CPN
Sécuritas Domen,.CPN
ZOLPAN Innovation Performance Expertise,.CPN

Les preuves comme *« société »*, *« groupe »*, *« institut »*,... qui jouent le rôle de preuves internes aussi bien que celui de preuves externes n'ont été retenues dans le dictionnaire que si elles font partie intégrante du nom comme dans *« Société Générale »*.

Les preuves internes

Nous avons commencé la construction de nos patrons d'extraction par le recensement des descripteurs juridiques des entreprises. Ceux-ci ont déjà été décrit au niveau du chapitre 6, où il fût question de reconnaître le nom de la compagnie sur la page d'acueil de celle-ci. Les descripteurs juridiques sont les formes juridiques que peut prendre toute organisation à but lucratif, en France on connait par exemple, les SARL (Société à Responsabilités Limitées), ou les SA (Société Anonyme), etc. Une liste exhaustive de ces descripteurs est disponible sur le Web à l'adresse *http ://www.corporateinformation.com/defext.asp*. Nous avons transformé ces unités en un dictionnaire électronique respectant le format DELAC, dans lequel les formes étendues des descripteurs juridiques représentent les formes de base et leurs acronymes les formes dérivées, comme ceci :

Société Anonyme,.FJ
SA,Société Anonyme.FJ
Société A Responsabilités Limitées,.FJ
SARL,Société A Responsabilités Limitées.FJ
...
...

Dans une seconde phase, nous nous sommes concentrés sur le recensement des preuves internes et à la description des grammaires locales associées. Nous citons à cet effet les preuves :

- agence
- atelier
- cabinet
- centre
- division
- école
- établissement
- entreprise
- groupe/group
- institut
- industrie
- laboratoire
- magasin
- office

- société
- université
- usine

Nous montrons dans la figure (fig. 8.28), les grammaires descriptives de la structure interne des noms d'organisations introduits par les preuves : groupe ou group, laboratoire et office et que l'on retrouve dans des noms comme :

– Groupe Air Fraice
– Degetel Group
– Groupe Fayat
– Cabinet Arevarecrutement
– Cabinet Selescope
– Laboratoire de physique chimie de Marseille
– Laboratoire de recherche de Bobigny
– Institut Supérieur de Gestion
– Ecole de physique générale et de technologie
– cabinet de recrutement BEAVER IT
– Société SEDITEC
– ...

Les contextes externes

Nous avons rassemblé, dans la phase d'apprentissage, les verbes introducteurs d'une organisation comme textit« compter », « recruter », « proposer », « rejoindre », qui interviennent aussi bien dans les contextes externes gauches que dans les contextes externes droits. Nous montrons dans la figure (fig. 8.29), la grammaire locale d'extraction du nom de la compagnie avec le patron *« <CPN> compte <NB> collaborateurs/Clients »* et la grammaire locale autour du patron *« <rejoindre> <CPN> »* qui donnent les concordances suivantes :

<rejoindre>
– Rejoignez<CPN> Sagem Communication</CPN>
– En **rejoignant** les 1700 collaborateurs de <CPN>Ricoh France</CPN>
– Rejoindre<CPN> ALTEN</CPN>

<compter>
– <CPN>RANDSTAD</CPN> compte 750 collaborateurs
– <org>Pizza Paï</org> compte aujourd'hui 2500 collabora-

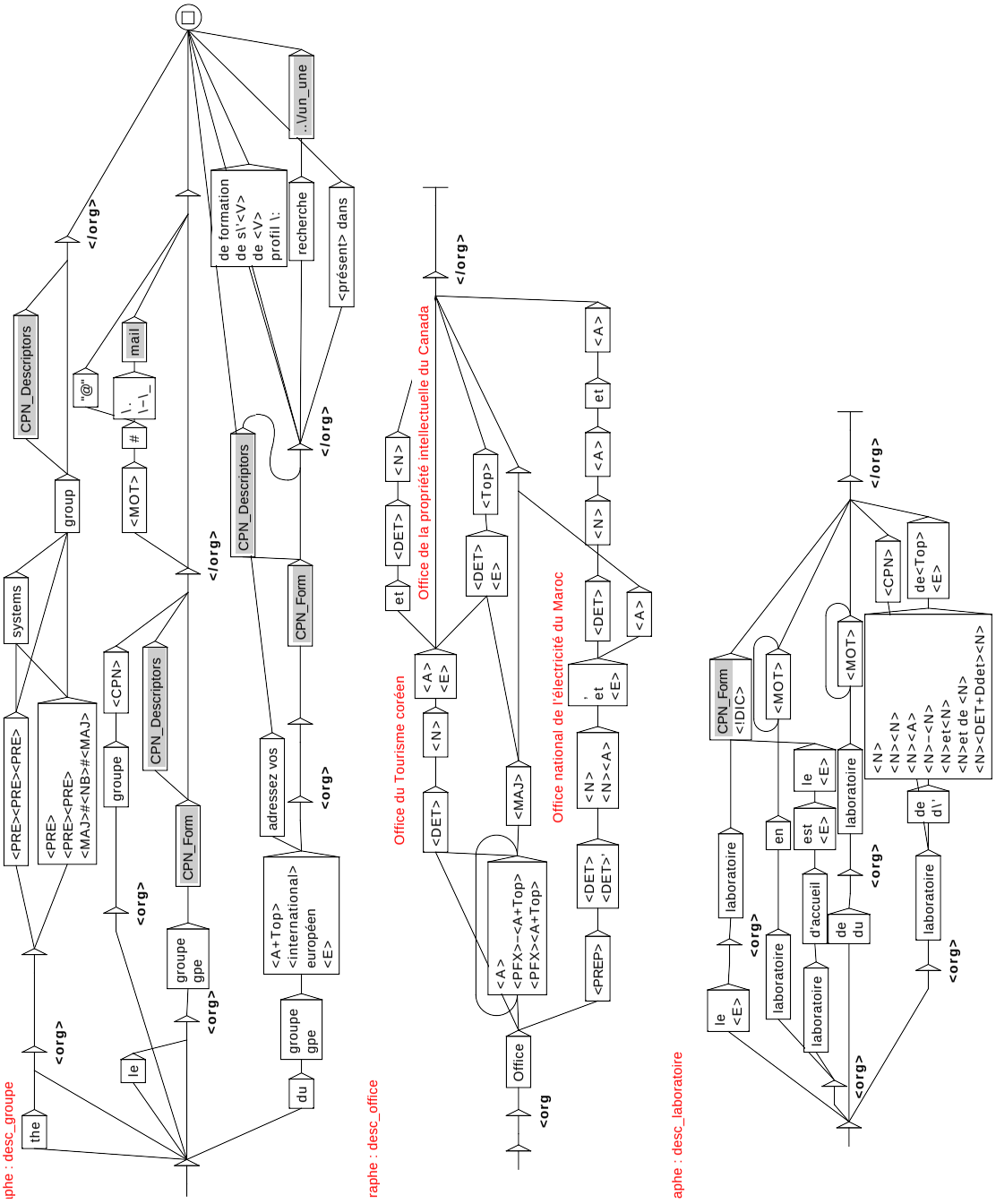


FIG. 8.28 – Description des contextes des noms d'organisations

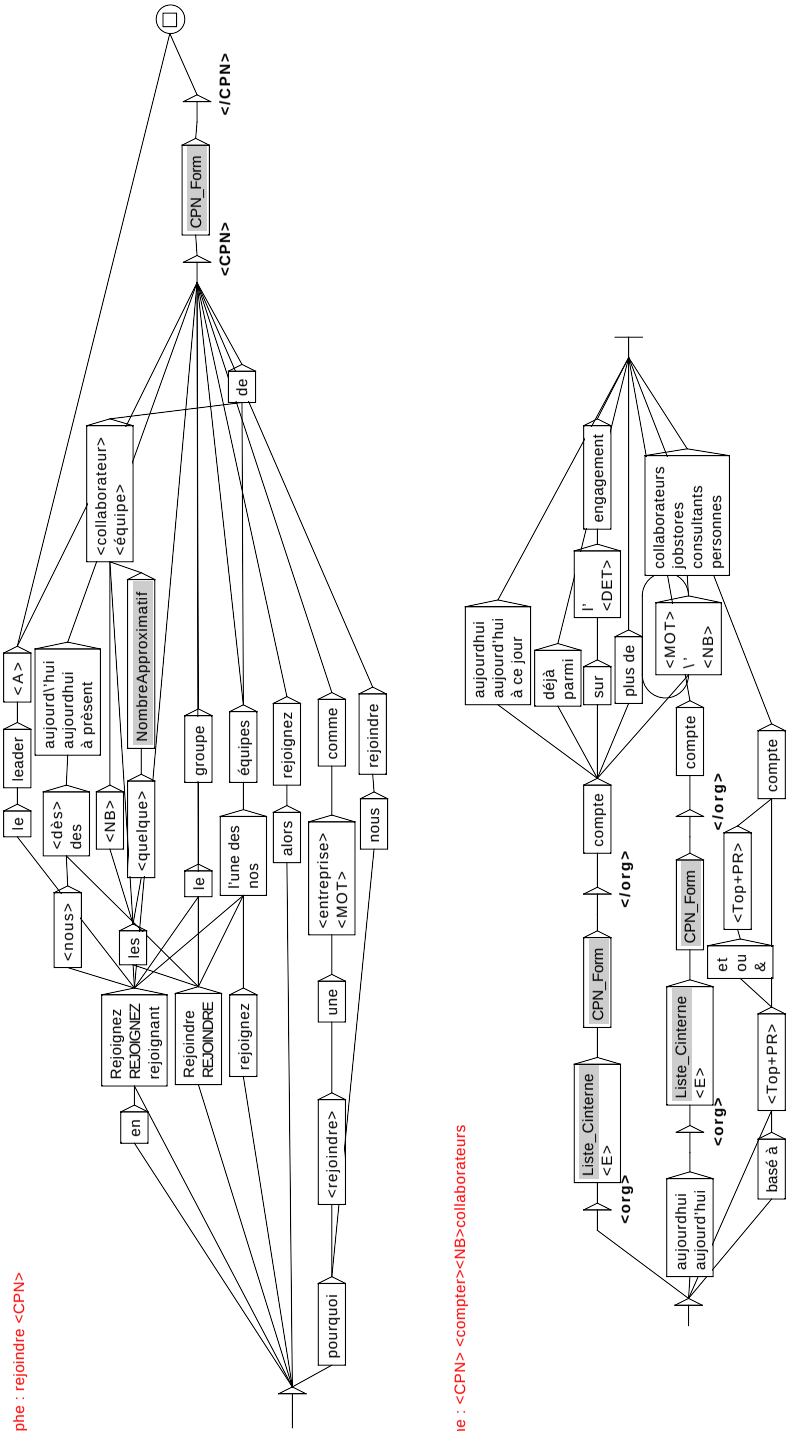


FIG. 8.29 – LG autour des verbes : compter et rejoindre

teurs
- <org>Adia</org> compte aujourd'hui 460 Jobstores

<recruter>
- <org>ALTEN Techno</org> recrute
- <org>S&H</org> recrute
- <org>CRIT AGENCE TERTIAIRE ET CADRES DE VALENCIENNES</org> RECRUTE

<CPN> filiale de <CPN>
- <org>F Dalkia</org>, filiale de<org> VEOLIA ENVIRONNEMENT</org>
- filiale d'<org>EDF</org>
- <org>Management Support</org>, filiale du Groupe <org> USG PEOPLE </org>

Certaines concordances erronées sont à observer à ce niveau. Il s'agit, néanmoins dans la plupart des cas d'erreurs dûes à une mauvaise utilisation des règles typographiques et surtout celle des majuscules. Nous présentons dans la liste ci-dessous quelques exemples de concordances erronées.

- Filiale<org : >Au </org>
- <org>F CDI Adecco </org> recrute
- <org : >France</org>, filiale du <org>groupe SUD-OUEST </org></org>
- <org> Loire</org> recrute

Nous avons développé 35 grammaires au total que nous appliquons en tenant compte de certaines combinaisons et en cascade tout en favorisant les grammaires les plus fiables. Nous avons organisé ces derniers sous forme de 5 itérations conditionnelles, chaque itération est dépendante des résultats obtenus à la précédente. Les figures (fig. 8.30) et (fig. 8.31) présentent les grammaires des 2 premières itérations.

Les 5 niveaux de priorités pour l'extraction du Nom de l'organisation	
Niveau1	Locutions types + descriptions des typologies internes
Niveau2	Combinaisons de Contact + Nom d'organisation + Adresse
Niveau3	Verbes introductifs + typologies internes et exetrnes
Niveau4	Preuves externes
Niveau5	Dictionnaire des noms d'organisations hors contextes

Nous présentons dans le tableau suivant les performances, Rappel et Précision obtenues pour la tâche d'extraction du nom de la compagnie sur un ensemble d'offres d'emploi récupéré aléatoirement sur le Net. Nous avons observé 3 types d'erreurs que nous comptons séparément dans le tableau récapitulatif suivant. Le premier est la reconnaissance trop longue d'une séquence, la seconde est la reconnaissance trop courte d'une séquence par rapport au nom exacte de l'entreprise et le troisième type d'erreur est celui de la reconnaissance d'une séquence comme étant le nom d'une entreprise lorsqu'il n'en est pas.

Ces performances sont calculées sur des offres d'emploi extraites aléatoirement à partir du Web.

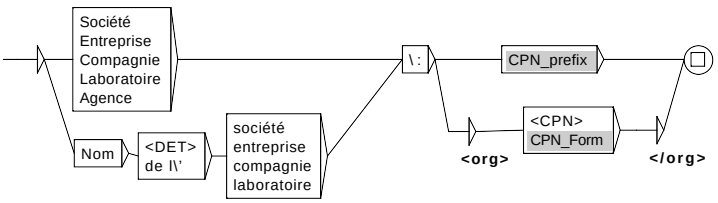


FIG. 8.30 – GL de la 1ère itération dans l'extraction du nom de la compagnie

Performances d'extraction du nom de l'organisation dans 100 offres d'emploi	
Nombre d'offres analysées	50
Nombre de concordances correctes	38
Nombre de concordances trop longues	5
Nombre de concordances trop courtes	2
Nombre de concordances incorrectes	3
Nombre de concordances non trouvées	2

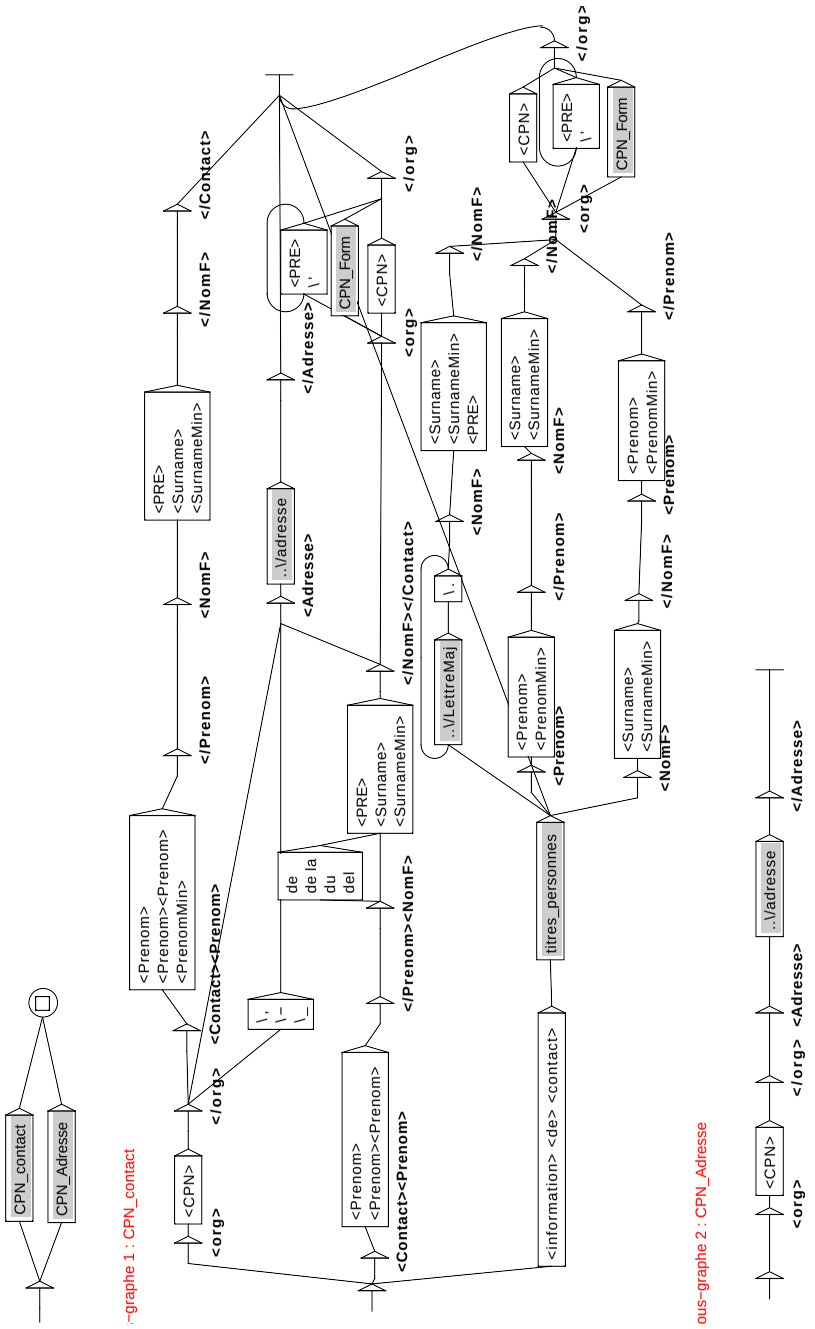


FIG. 8.31 – GL de la 2nde itération dans l'extraction du nom de la compagnie

8.9.2 L'adresse Postale de l'entreprise

Nous avons décrit au niveau de l'extraction du lieu du travail, les grammaires descriptives des adresses postales dans les offres d'emploi, nous ne reviendrons donc pas dessus. Cette information sera recherchée par ces mêmes patrons décrits plus haut, dans le cas où l'attribut « Adresse » n'a pas encore été rempli et que celle-ci n'a pas été capturée lors de l'extraction du lieu de travail. L'adresse de l'entreprise est une information très importante, car elle renseigne sur la localité du siège mais elle est également importante pour l'envoi des dossiers de candidature si le moyen d'envoi n'a pas été explicitement mentionné.

8.9.3 Téléphone, Fax, Email et Site Web

Ces quatre attributs sont des informations dont la structure interne est très discriminatoire. Le numéro de téléphone et le numéro de fax sont par exemple des suites de chiffre avec possibilité de parenthèses et dont la longueur est compris entre 7 et 14 digits. Pour ce qui est de l'adresse e-mail, la structure autour de l'arobase permet sa reconnaissance sans ambiguïté. Pareillement pour l'URL de l'entreprise dont la structure interne est précise et sûre.

Au niveau de l'extraction des coordonnées de l'entreprise, nous nous trouvons confronté au problème que les informations trouvées peuvent ne pas appartenir à l'entreprise directement en besoin de personnel, mais ils peuvent correspondre au cabinet de recrutement délégué. En effet, une offre d'emploi postée par un cabinet de recrutement prenant le soin de cacher l'identité du recruteur final, mentionne l'URL, l'E-mail, le numéro de téléphone et de fax de sa propre agence et non celle du recruteur direct. Ainsi un candidat désireux d'en savoir plus sur l'entreprise dont le poste l'intéresse s'attend à parcourir le site de cette dernière et non pas celui de l'intermédiaire du marché de l'emploi. Ceci est certes, contrariant, mais nous assurons par ailleurs du moins une cohérence d'association des coordonnées avec le nom de l'entreprise

trouvé et même s'il ne s'agit pas du recruteur final, l'offre est postée par l'entreprise dont nous disposons des coordonnées. Nous proposons dans la figure (fig. 8.32) et (fig. 8.33) les grammaires principales pour l'extraction du numéro de téléphone, du numéro de fax, de l'adresse e-mail et du site Web dans une offre d'emploi, suivi par une liste de concordances extraite à partir de notre corpus test.

- Ph : <Tel> (+33).1.42.99.83.33 </Tel>
- contactez nous par téléphone au <Tel> 01.43.90.48.40 </Tel>
- Fax : <Fax> 0155302829 </Fax>
- par fax au <Fax> 04-79-81-17-02 </Fax>
- mail : <Email> service -personnel.sandvik-SAS@sandvik.com </Email>
- <Email>cdossantos@telelangue.com </Email>
- site : <URL> http ://www.epigone.fr </URL>
- <URL>www.etseurope.org </URL>

Nous introduisons une information supplémentaire dans les coor-

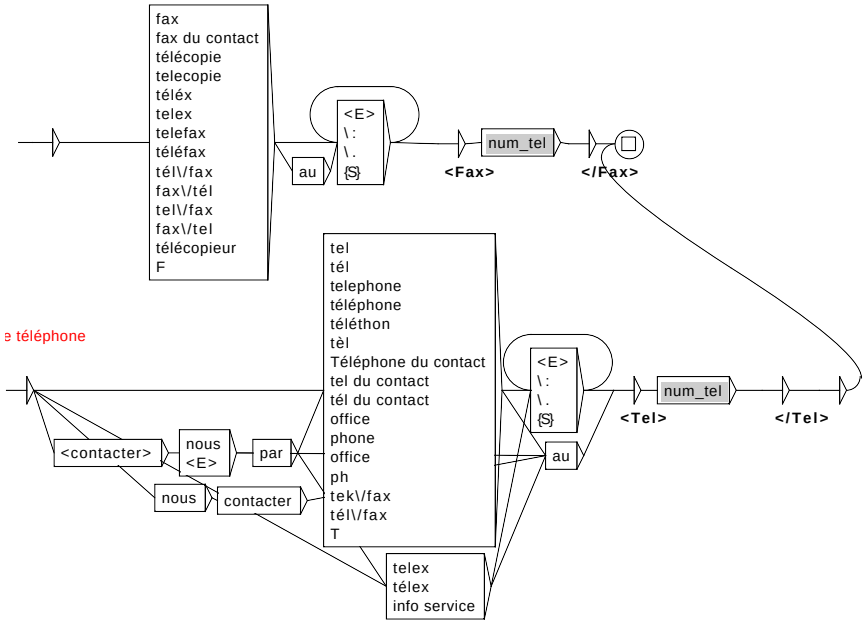


FIG. 8.32 – GL pour la reconnaissance du numéro de téléphone et du numéro de fax

données de l'entreprise, à savoir « le Contact » chargé de recevoir les dossiers de candidature. Cette information est recherchée en cascade au niveau de l'extraction du nom du poste,. Si cet attribut n'a pas été trouvé tout au long des extractions précédentes sur le nom de la compagnie, l'adresse, le numéro de téléphone etc, nous appliquons dans un premier niveau la grammaire de la figure (fig. 8.34), qui permet de délimiter le nom, prénom et fonction de la personne à contacter s'ils sont présents et introduits par les contextes externes sûrs, recensés dans la phase d'apprentissage.

Nous présentons ci-joint une liste de concordances extraites à partir des grammaires de premier niveau sur le corpus test 1.

- Contact : <Contact> <Prenom> Cyndie</Prenom> <Nom>Herrador Huntress</Nom> </Contact>
- Informations du contact<Contact> <Nom>Emmanuelle</Nom> <Pre-

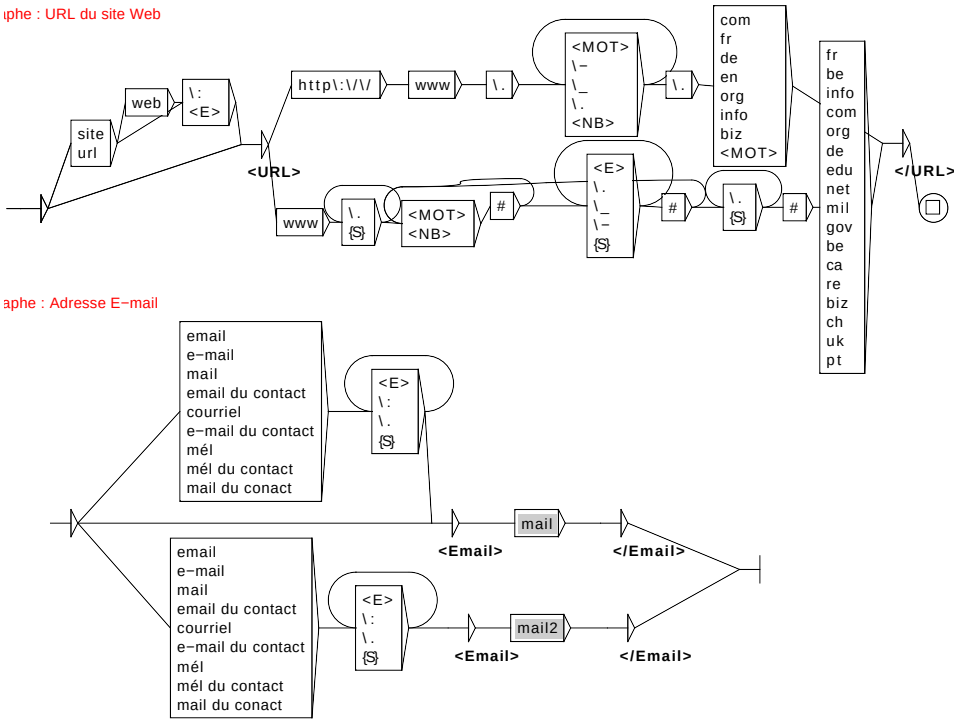


FIG. 8.33 – GL pour la reconnaissance de l'URL et de l'adresse Email

nom>HUBERT</Prenom> </Contact>
- adressez vos candidature à <Contact> <Prenom>Mathieu </Prenom> <Nom> Vetter</Nom> </Contact> <org> Computer Futures Solutions</org>
- Veuillez adresser votre candidature sous réf <Ref> NDSKIDFM07492</Ref> à <Contact> <Prenom>Emilie</Prenom> <Nom>LIENARD</Nom> </Contact>, <CPN>Cabinet SURF</CPN> <Adresse> 2-6 rue des Bourets 92150 SURESNES </adresse>
- Coordonnées du contact<Contact> <Prenom>Alia</Prenom> <Nom> KHATTAB</Nom> </Contact> <Email> a.khattab@elitis.com </Email>
- envoyer votre candidature à : <Contact> <Prenom>Anne</Prenom> <Nom>LOCHARD</Nom> </Contact> - <ServiceEnt>Chargée de recrutement</ServiceEnt>
- Contact Information<Contact> <Prenom>PIERRE-HENRI </Prenom> <Nom> LEVERD</Nom> </Contact> <org> Huxley Associates</org>
- Coordonnées du contact<Contact> <Prenom> Viktoria</Prenom> <Nom> Minya</Nom> </Contact> <Email> europerecruitment@etseurope.org </Email>

Ces grammaires ont reconnu quelques concordances erronées que l'on peut observer dans la liste suivante :

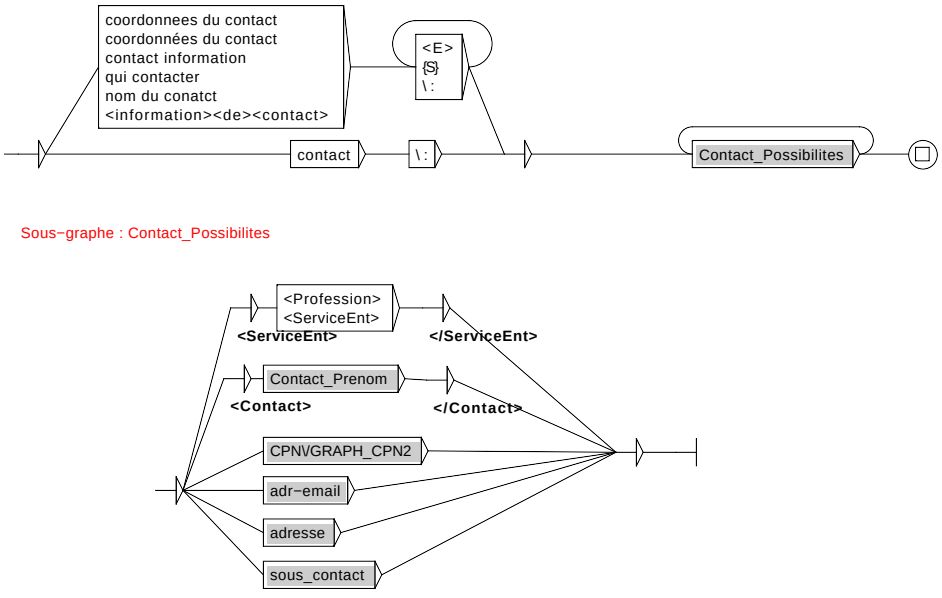


FIG. 8.34 – GL de niveau 1 pour l'extraction du nom du contact

- Informations de contact<Contact> <Prenom>Emilie LIENARD</Prenom><Nom> Cabinet Surf</Nom></Contact> <Adresse> 2-6 rue des<Rue> Bourets</Rue> SURESNES FRANCE 92150 </Adresse>
- Informations de contact<Contact><Prenom> Page</Prenom> <Nom> Personnel Postuler </Nom> </Contact>
- Coordonnées du contact<Email> inforeims@moreno-international.fr </Email> <Contact> <Prenom> Moreno</Prenom> <Nom> International</Nom> </Contact> <Adresse> 17 rue<Rue> Courmeaux REIMS 51100 </Adresse>
- Informations de Contact :<Contact> MME<Contact> <Prenom> LEROY</Prenom><Nom> Adia</Nom></Contact>

Dans cette dernière concordance erronée, par exemple, nous avions omis dans la grammaire, l'entrée reconnaissant le patron : « *<titre> <Nom> <CPN>* », que nous avons ajouté en remarquant l'erreur pour obtenir la concordance suivante :

- Informations de Contact :<Contact> MME<Nom> LEROY </Nom> </Contact> <org> Adia </org>

8.10 Expérience et formation souhaitée du candidat

Il est important d'indexer les offres d'emploi par les deux informations sur l'expérience et la formation du candidat souhaité, d'une part car les recruteurs le mentionne dans la plupart de leurs annonces et qu'il s'agit d'autre part d'un élagage très important dans la liste des documents pertinents répondant à une requête. Prenons le cas d'un jeune diplômé en informatique à la recherche d'un poste de « *Concepteur-programmeur* », il sera confronté à une liste de résultats très longue dans un moteur de recherche classique, dans laquelle il serait très pertinent de supprimer les offres où l'on recherche des candidats avec plusieurs années d'expériences ou sortant d'une école d'ingénieur et non pas ayant suivi un cursus universitaire classique. Nous avons construit à cet effet des grammaires locales permettant de reconnaître ces infor-

mations dans les textes des offres d'emploi si elles sont précisées. Les séquences présentées dans la liste qui suit ont été reconnues par les grammaires de la figure (fig. 8.35) et la figure (fig. 8.36).

- Expérience<ExperienceMin> de 3 à 4 ans </ExperienceMin>
- <ExperienceMin>expérience de 2 à 3 ans en Contrôle de Gestion industriel au sein d'un Groupe à dimension internationale </ExperienceMin>
- <ExperienceMin>vous avez une expérience de 1 à 2 ans dans une fonction similaire </ExperienceMin>
- Niveau de poste :<ExperienceMin>Débutant </ExperienceMin>
- Expérience<ExperienceMin> en esthétique </ExperienceMin>
- expérience commerciale réussie<ExperienceMin> (3 à 7 ans) </ExperienceMin>
- <ExperienceMin> vous avez une expérience réussie de 5 ans minimum dans la vente de produits grands public </ExperienceMin>
- <ExperienceMin>ayant acquis une expérience réussie en vente </ExperienceMin>
- <Formation> De formation supérieure Bac+4 </Formation>
- <Formation> De formation Bac+2 BTS ou DUT informatique </Formation>
- <Formation> Niveau d'études : DESS, DEA, Grandes écoles , Bac + 5 </Formation>
- <Formation> Vous êtes ingénieur de formation (Bac+5)</Formation>
- <Formation> Vous êtes de formation ingénieur en automatismes </Formation>

Nous avons neanmoins observé un taux d'erreurs assez élevé par rapport aux autres informations extraites dans les offres. Il s'agit dans la plupart des cas d'erreurs dûes à une reconnaissance trop longue des séquences à extraire, comme on peut le voir dans les exemples suivants :

- <ExperienceMin> expérience dans le domaine du textile est un plus Salaire </ExperienceMin>
- <ExperienceMin> expérience de 5 ans en Contrôle de Gestion dans un milieu d'Engineering est </ExperienceMin>

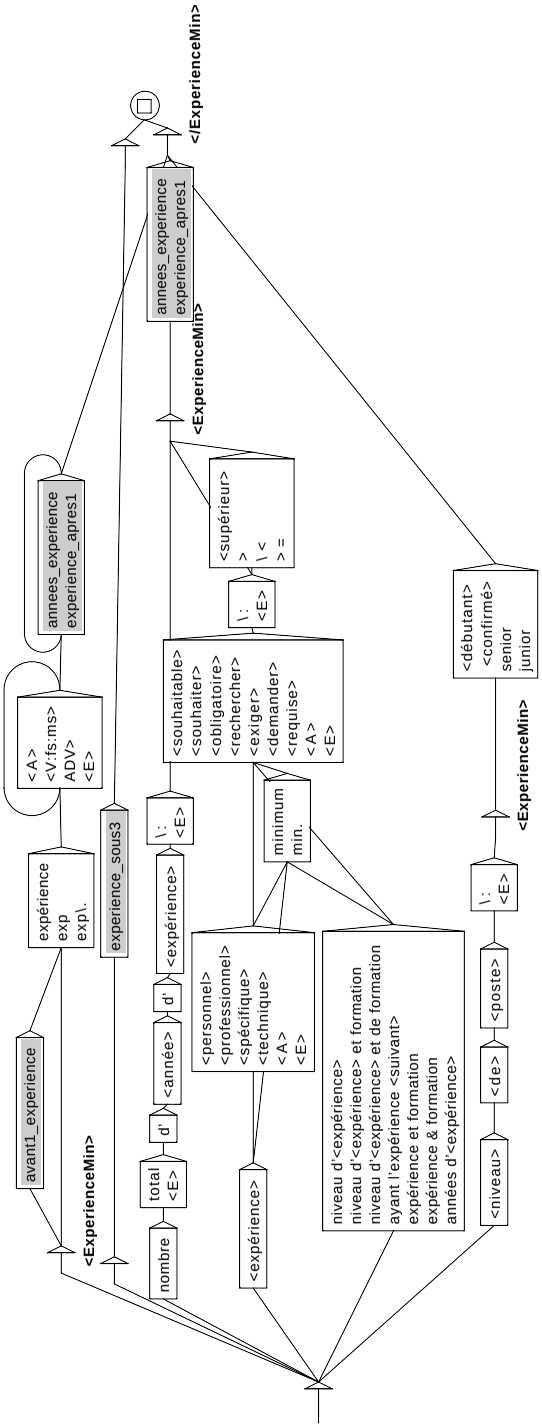


FIG. 8.35 – GL de niveau 1 pour l'extraction du niveau d'expérience exigé

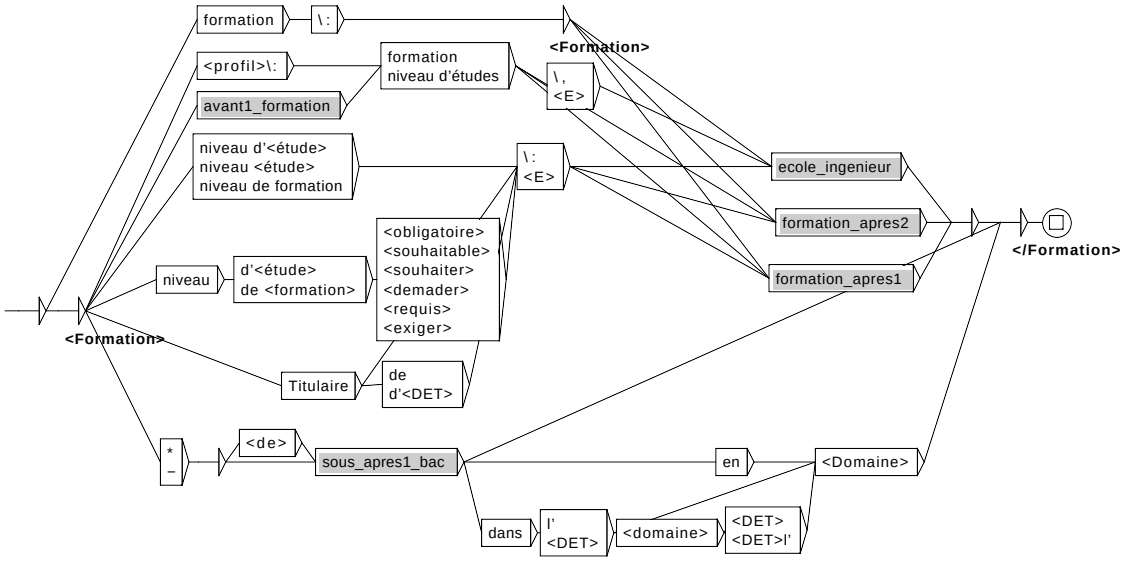


FIG. 8.36 – GL de niveau 1 pour l'extraction du niveau de formation souhaité

8.11 Exemple applicatif

Nous proposons de montrer dans cette section, la transformation de représentation décrite tout au long de ce document et ce sur un exemple concrêt d'une offre d'emploi .

En appliquant le processus de traitement décrit au début de ce chapitre et illustré par la figure (fig. 8.1) à la page (171), on obtient un étiquetage automatique d'une offre d'emploi comme celui de la figure (fig. 8.11). On y distingue les étiquetages structurels, à travers les indications en vert (*[TAGCONTACT], [TAGSALAIRE], [TAGEXP], [TAGPOSTE], [TAGCPN] ,...*) et qui segmentent l'offre par rapport à ses sections sémantiques à travers les locutions types classifiées dans les classes structurelles sémantiques décrites ci-dessus. On y distingue ensuite les étiquettes en rouge des informations recherchées et trouvées pour le remplissage du formulaire correspondant à la structure interne de la base de données. Une première erreur de reconnaissance est observée au niveau du *lieu du poste*, où la séquence reconnue *« de Paris »* ne correspond pas directement au lieu réel du poste qui devrait contenir le terme *extérieur*. En effet le lieu de la mission exprimé dans

l'annonce est : à l'*«extérieur de Paris»*, ce qui rend cette extraction incorrecte et infidèle à l'annonce initiale, car la région de l'île de France se compose de 8 départements différents et de 1282 communes avec plus de 12.072 km^2 de superficie contre une superficie de Paris de $105km^2$. Nous avons marqué explicitement en jaune(fig. 8.11) les séquences répondant aux informations cherchées mais pas reconnues. Dans l'état actuel des graphes de reconnaissance des dates de lancement du poste, nous n'avions pas décrit les 2 patrons de reconnaissance : *« qui pourrait démarrer une mission très rapidement »* et *« vous devez impérativement être disponible sous 1 à 4 semaines »*, que nous n'avions pas rencontré lors de la phase d'apprentissage et pour lesquelles, il nous a suffit de quelques minutes pour les ajouter dans les grammaires appropriées.

Nous avons élaboré une interface Web de contrôle, nous permettant de visualiser les extractions faites sur des offres d'emploi aléatoires et par la même occasion de vérifier si certaines informations existantes n'ont pas été reconnues bien qu'elles devraient l'être. Dans l'interface de la figure (fig. 8.11), nous présentons une impression d'écran dans laquelle on peut distinguer quatre cadres :

Cadre de recherche

Le cadre en haut à gauche, met un champ de saisie et un bouton de recherche à disposition de l'utilisateur. Si celui-ci entre un ou plusieurs mots clés en relation avec un emploi quelconque [2] et clic ensuite sur le bouton « Rechercher », le programme lance une requête parallèle sur plusieurs Jobboards avec le ou les termes introduits dans le champ de saisie.

Cadre de la liste des résultats

Le cadre en haut à droite propose une liste de liens correspondants aux résultats de la requête lancée sur les Jobboards. À ce niveau l'utilisateur est prié de cliquer sur une URL s'il désire lancer le processus d'extraction de l'information.

[2]Sur la figure le mot clé de recherche est « cuisinier »

URGENT ! <PosteName> DÉVELOPPEUR PERL - 94 - FREELANCE </PosteName>(H/F)
FR-IDF-ILE DE FRANCE

Descriptif :
Mon client, un éditeur de logiciel international, recherche de façon urgente un Développeur Perl.

[TAGCPN] La Société :
Mon client est un acteur majeur sur son marché travaillant avec les plus grands comptes internationaux. Suite à une surcharge importante, ils sont actuellement à la recherche d'un développeur Perl qui pourrait démarrer une mission très rapidement.
Mission située à l'extérieur <Location> de Paris </Location> (très facile d'accès par les transports en commun) pour laquelle vous devez impérativement être disponible sous 1 à 4 semaines.

[TAGPOSTE]Description de poste :
Vous devrez tout d'abord analyser plusieurs sites Web ainsi que leurs fichiers attachés puis vous aurez à charge de leur développement sous la derniere version de Perl.
Votre expertise technique, votre implication et votre motivation vous permettront d'évoluer au sein d'une équipe dynamique, pour un client qui apportera une forte valeur ajoutée à votre parcours.
Excellente opportunité de rejoindre une société très demandée, sur une mission de<Duree> 3 mois </Duree> avec de fortes possibilités de renouvellement.

[TAGEXP] Description des Candidats :
- Perl : 2 ans minimum
- Anglais est un plus
- XML : 1 an
– Html : 2 ans

[TAGSALAIRE] Tarif :
<Salaire> 290 à 330€/jour selon expérience <Salaire> .

[TAGCONTACT] Contact :
Si vous avez les compétences nécessaires, merci de me contacter très rapidement afin que je vous organise un entretien avec mon client.

<CPN> Computer Futures Solutions </CPN> est un acteur majeur sur le marché du recrutement et de la prestation de services au niveau Européen dans le domaine des <DomainOrg> technologies de l'information </DomainOrg> avec un chiffre d'affaires de plus de 220 Millions d'euros. Nous sommes présents dans les plus grandes capitales (Paris, Londres, Amsterdam, Bruxelles …).

Additional Information
Negotiable
Position Type:<TypeContrat> Full Time </TypeContrat>, Temporary / Contract / Project
<Reference> Ref Code: 391289 </Reference>

[TAGCONTACT] Contact Information
<Contact> <Prenom> Rudy </Prenom> <NomF> Nabet </NomF> </Contact>
<CPN> Computer Futures Solutions </CPN> - Paris
<Addresses> 33 RUE DE LA BOETIE, PARIS 75008 </Addresses>
Ph:<TEL> + 33 1 42 99 83 33 </TEL>
Fax:<FAX> + 33 1 42 99 83 00 </FAX>

FIG. 8.37 – Exemple complêt d'extraction et d'étiquetage d'une offre d'emploi

Cadre des informations recherchées

Le cadre à gauche en bas récapitule toutes les informations recherchées par le processus de traitement de l'offre lancé, et rempli pour chaque information recherchée la case correspondante du formulaire.

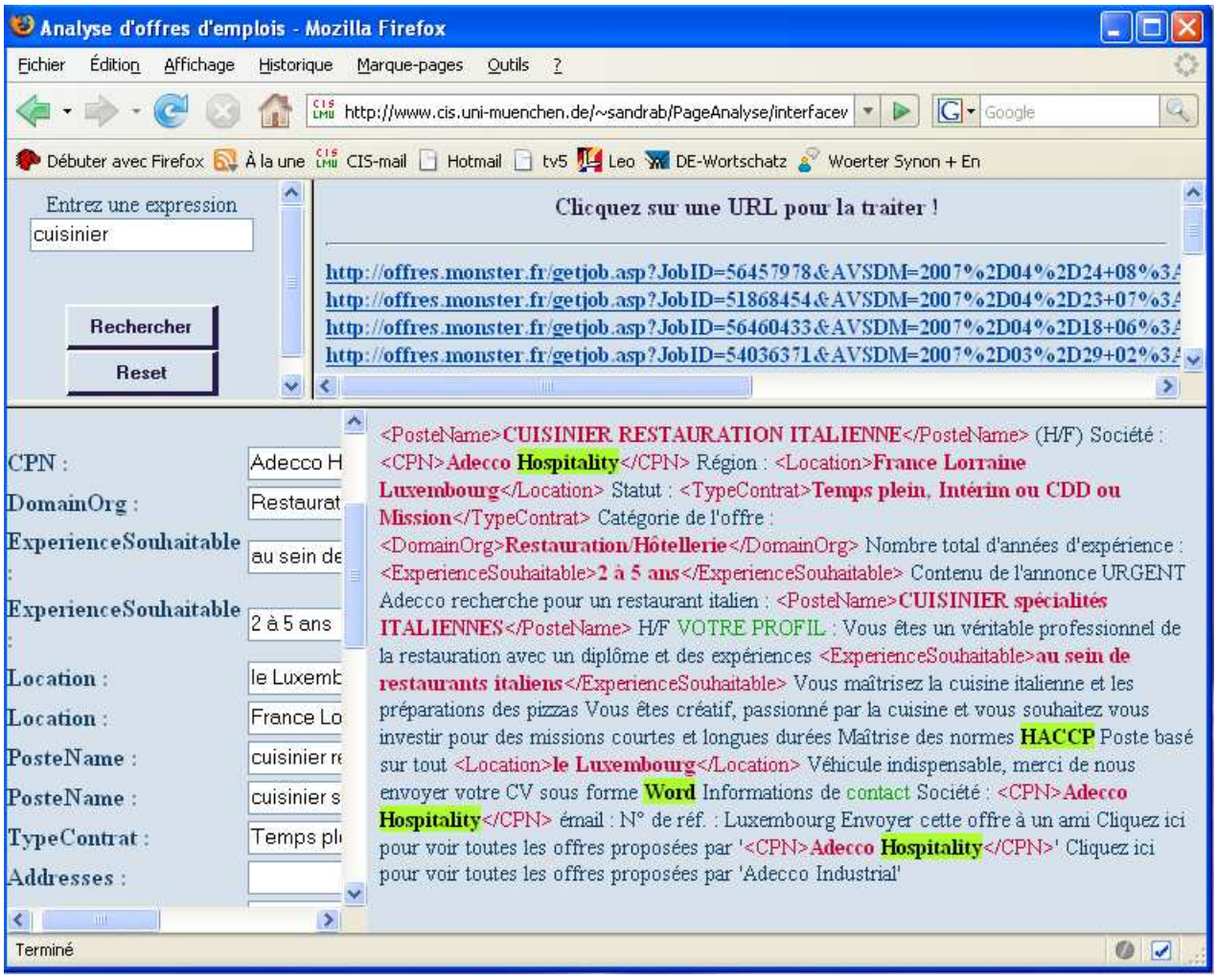


FIG. 8.38 – Capture d'écran de l'interface de contrôle

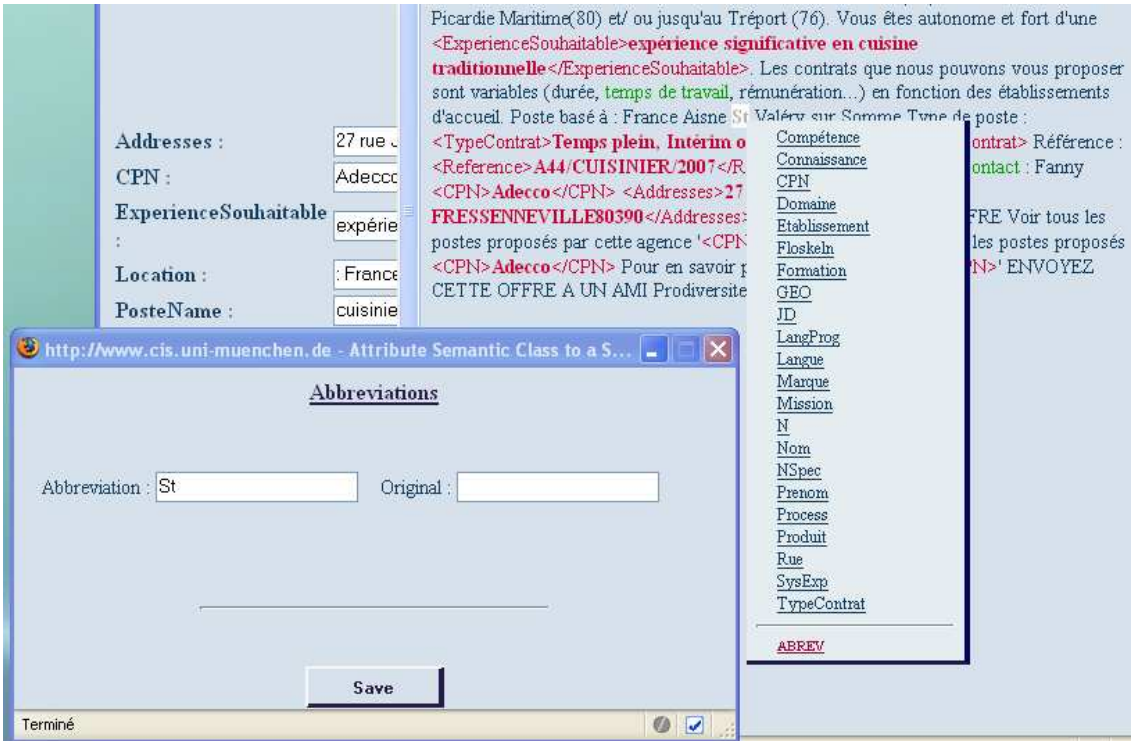


FIG. 8.39 – Capture d'écran pour l'ajout d'une abréviation dans le dictionnaire *Abrev.dic*

Cadre du texte étiqueté

Le cadre en bas à droite propose le texte étiqueté de l'URL choisie et traitée par le programme *ExtractInfo*. On y décèle également les mots inconnus de nos dictionnaires mis en relief par la couleur de fond verte.

Cette interface à été élaborée dans un but de contrôle, elle nous permet de visualiser à temps réel et sur des offres inconnues la performance de nos extractions. Elle nous a été d'une grande utilité dans la phase d'apprentissage pour améliorer les grammaire locales d'extraction, un clic gauche sur la souris, liste toutes les classes sémantiques que nous avons crée tout au long de nos travaux et qui permettent d'ajouter directement le mot ou les mots sélectionnés dans le dictionnaire choisi (voir fig. 8.11). De la sorte nous sommes toujours en mesure, de vérifier

pour une offre donnée, si la reconnaissance n'a pas été possible à cause que le terme n'est pas encore classifié, ou bien qu'il est nécessaire de mettre la grammaires correspondante à jour.

Conclusion

Nous avons présenté dans ce chapitre les différentes grammaires locales construites pour l'extraction de 16 informations à partir d'une offre d'emploi reçue en entrée. Nous avons montré aussi bien les concordances positives que les concordances négatives occasionnées par les grammaires appliquées sur des corpus de test. En fonction de l'information à extraire, celles-ci sont appliquées sous forme de suites d'itérations conditionnelles permettant de favoriser les grammaires sûres et de n'appliquer les grammaires ambiguës que si les premières ne retournent pas de résultats.

Conclusion

L'objectif de cette thèse est de générer automatiquement une plateforme centralisée mise à jour en temps réel regroupant toutes les offres d'emploi disponibles sur Internet et répandues, à ce jour, sur une multitude de vecteurs de diffusions parallèles. Nous y décrivons les étapes de recensement, d'analyse et de transformation des offres d'emploi à travers une analyse linguistique basée sur la terminologie du sous langage étudié et de la notion particulière de grammaires locales. Après avoir donné un aperçu des approches existantes dans les domaines respectifs de la recherche d'information et de l'extraction automatique d'information ainsi qu'une description du format DELA des dictionnaires électroniques des noms composés, trois domaines en rapport direct avec nos travaux, nous avons présenté les différents modules élaborés pour répondre à nos besoins. Nous avons construit dans un premier temps un système de reconnaissance automatique de pages d'accueil d'entreprises, afin de récupérer les annonces d'emploi à la source et ne pas dépendre de facteurs subjectifs de choix, comme c'est le cas des systèmes actuels. Ce premier système fournit en résultat un annuaire d'entreprises accompagnés de leurs coordonnées, mis à jour régulièrement et dont l'utilité n'est pas exclusive à nos besoins, il peut, en effet, être adapté à d'autres services comme par exemple un moteur de recherche de biens et de services géographiquement ciblé. Nous avons ensuite montré les étapes de construction de notre dictionnaire électronique des noms de profession composés auquel nous avons recourt dans la

phase d'extraction d'information, mais dont l'emploi peut également améliorer la qualité de nombreuses applications connexes du TALN.

La plus grande partie de nos travaux a cependant été consacrée à la construction de grammaires locales pour l'extraction d'information dans le sous-langage des annonces d'emploi. Informations qui permettent une transformation de l'espace de représentation des données initialement écrites en langage naturel dans un format sémantiquement structuré et de remplir automatiquement la base de données, dont la structure correspond exactement aux différentes informations recherchées. De par cette transformation automatique du format des données, nous nous distinguons des systèmes similaires dans lesquels, les offres sont introduites dans la base à travers le remplissage manuel d'un formulaire correspondant à la structure interne souhaitée. De telles grammaires bien que très puissantes, doivent être augmentées en permanence d'ou l'avantage d'avoir utilisé les grammaires locales sous forme de graphes qui rendent les patrons visuellement lisibles et facilitent par conséquent la maintenance et la mise à jour.

Parallèlement à la construction de grammaires locales pour les verbes supports représentant les tâches à accomplir dans le poste à pourvoir et les qualifications à apporter par les candidats, nous souhaitons faire évoluer notre dictionnaire électronique des noms de profession vers une ontologie des appellations d'emploi qui nous permette de capturer les relations d'hyponymie, de synonymie et d'hyperonymie entre les noms des postes et améliorer ainsi les résultats de recherche par l'extension automatique des requêtes utilisateurs par les termes associés. Une telle ontologie nous permettrait d'augmenter le nombre de documents pertinents s'il y en a peu et inversement de diminuer le nombre de documents s'il y en a beaucoup qui soient pertinents et ce à travers un système interactif d'extension de la requête avec les termes plus généraux ou plus spécifiques dépendamment du cas.

BIBLIOGRAPHIE

[1] J. Anscombre. Parole proverbiale et structures métriques. *Langues, n139*, pages 6–26, 2000.

[2] A. P. Asirvatham and K. K. Ravi. Web page classification based on document structure, 1999.

[3] M. G. B. Courtois, G Gross. Dictionnaire électronique des noms composés delac : les composants na et nn. Technical report, Université Paris 7, Rapport Technique LADL 55, 1997.

[4] Benveniste. Fondements syntaxiques de la composition nominale. *Problèmes de linguistique générale 2*, pages 145–176, 1974.

[5] D. M. Bikel, R. L. Schwartz, and R. M. Weischedel. An algorithm that learns what's in a name. *Machine Learning*, 34(1-3) :211–231, 1999.

[6] W. Black, F. Rinaldi, and D. Mowatt. Facile : Description of the ne system used for muc-7, 1998.

[7] A. Borthwick, J. Sterling, E. Agichtein, and R. Grishman. Nyu : Description of the mene named entity system as used in muc-7. *Seventh Message Understanding Conference (MUC-7).*, 1998.

[8] K.-J. Chen and S.-H. Liu. Word identification for mandarin chinese sentences. In *Proceedings of COLING-92*, pages 101–107, 1992.

[9] C.-H. Chi, C. Ding, and A. Lim. Word segmentation and recognition for web document framework. In *Proceedings of the 1999*

ACM CIKM International Conference on Information and Knowledge Management, Kansas City, Missouri, USA, November 2-6, 1999, pages 458–465. ACM, 1999.

[10] H. L. Chieu and H. T. Ng. Named entity recognition with a maximum entropy approach. In W. Daelemans and M. Osborne, editors, *Proceedings of CoNLL-2003*, pages 160–163. Edmonton, Canada, 2003.

[11] B. Courtois and Co. Dictionnaire électronique delac : les noms composés binaires. Technical report, Université Paris 7, Rapport Technique LADL 56, 1997.

[12] e. a. Courtois, Dubois. Dictionnaires éléctroniques du français. Programme de recherches coordonnées, Laboratoire d'Automatqiue Documentaire et linguistique, Université Paris7, 1989.

[13] M. M.-c. D. Lepesant. Introduction aux classes d'objets. *Langages, (131)*, pages 6–33, 1998.

[14] D. M. Donald. Internal and external evidence in the identification and semantic categorization of proper names, 1996.

[15] J. Dubois. *Dictionnaire de linguistique et des sciences du langage.* Larousse, Paris, 1994.

[16] F. Duclaye. *Apprentissage automatique de relations d'équivalence sémantique à partir du web.* PhD thesis, École Nationale Supérieure des Télécommunications, Paris, 18 Novembre 2003.

[17] T. Emerson. Segmentation of chinese text.various approaches to the problems of separating the components of a sentence, 2001.

[18] N. F. et Denis Maurel. Elaboration d'une cascade de transducteurs pour l'extraction des noms de personnes dans les textes. *Actes de la conférence Traitement Automatique du Langage Naturel (TALN'2001), Tour*, 2001.

[19] D. L. et P. Pantel. Discovery of inference rules for question-answering, 2001.

[20] F. Y. F. Duclaye, O. Collin. Apprentissage automatique de paraphrases pour l'amélioration d'un système de questions-réponses, 2003.

[21] C. Fan and W. Tsai. *Automatic word identification in chinese sentences by the relaxation technique*, volume 4, pages 35–56. Computer Processing of Chinese and Oriental Languages, 1988.

[22] A. Filzmeyer. Identifikation und analyse von firmen-homepages. Technical report, CIS, http ://www.cis.uni-muenchen.de, 2005.

[23] R. Florian, A. Ittycheriah, H. Jing, and T. Zhang. Named entity recognition through classifier combination. In W. Daelemans and M. Osborne, editors, *Proceedings of CoNLL-2003*, pages 168–171. Edmonton, Canada, 2003.

[24] Y. Fondeur and C. Tuchszirer. Internet et les intermédiaire du marché du travail. Technical report, IRES - Institut de Recherches Economiques et Sociales, 2005.

[25] N. Friburger. *Reconnaissance automatique des noms propres. Application à la classification automatique des textes journalistiques.* PhD thesis, Université François Rabelais, Tour, 2 December 2002.

[26] C. Fuchs. La paraphrase, 1982.

[27] M. Garrigues. Prepositions and the names of countries and islands : a local grammar for the automatic analysis of texts. *Language Research 31 :2, Séoul*, 1995.

[28] M. Gloeggler. *Suchmaschinen im Internet.* Springer-Verlag, Berlin, 2003.

[29] R. Grishman. Information extraction : Techniques and challenges, 1997.

[30] G. Gross. Typologie des noms composés : le lexique électronique des noms composés du français. Rapport atp, CNRS, Université Paris 13, 1986.

[31] G. Gross. Degré de figement des noms composés. *Langue 90*, pages 57–72, 1988.

[32] G. Gross. Définition des noms composés dans un lexique-grammaire. *Langue Française 87*, pages 57–72, 1990.

[33] G. Gross. Classes d'objets et description des verbes. *Langages, (115)*, pages 15–30, 1994.

[34] G. Gross. *Les expressions figées en francais. Noms composés et autres locutions.* Ophrys, Paris, 1996.

[35] M. Gross. The construction of local grammars. *Finite State Language Processing, The MIT Press, Cambridge, MA*, pages 329–352, 1997.

[36] M. Gross. A bootstrap method for constructing local grammars. In *Contemporary Mathematics : Proceedings of the Symposium, University of Belgrad*, pages 229–250, Belgrad, 1999.

[37] Z. S. Harris. *Language and Information.* New : York : Columbia University Press, 1988.

[38] E. S. H. J. Moore. Web page categorization and feature selection using association rule and principal component clustering, 1997.

[39] D. Lemire. Théorie de l'information, 2006.

[40] B. Levrat and T. Amghar. Paraphrase et reformulation : une présentation. Technical report, Université Paris-Nord. Contrat C.N.E.T. Stratégie de reformulation, 1995.

[41] W. S. M. Stock. Klassifikation und terminologische kontrolle : Yahoo !, open directory und oingo im vergleich, 12 2000.

[42] J. Martinet. Un modèle vectoriel de recherche d'information pour les images. Technical report, Equipe Modélisation et Recherche d'Information Multimédia, Université Joseph Fourier, 2001.

[43] M. Mathieu-colas. Un dictionnaire électronique des mots à trait d'union. *Langue française*, pages 76–85, 1995.

[44] M. Mathieu-colas. Essai de typologie des noms composés français. Technical report, laboratoire de Linguistique Informatique, Université de Paris Nord, CNRS(UMR 0195), 1998.

[45] A. G. Maurice Grévisse. *Le bon usage*. Duculot, 1994.

[46] S. Mejri. Le figement lexical. descriptions linguistiques et structuration sémantique. Technical report, Publication de la Faculté des lettres Manouba, tunis, 1997.

[47] S. Mejri. Structuration sémantique et variation des séquences figées. *Actes de la 1ère RLM*, pages 103–112, 1998.

[48] I. Melcuk. Dependency synt ax : T heory and pract ice, 1988.

[49] A. Mikheev, C. Grover, and M. Moens. Description of the ltg system used for muc. *Seventh Message Understanding Conference (MUC 7)*, 1998.

[50] S. Miller. Algorithms that learn to extract information - bnn : Description of the sift system as used for muc-7. *Proceedings of the 7th Message Understanding*, 1999.

[51] A. Monceaux. *La formation des noms composés de structure Nom Adjectif. Élaboration d'un dictionnaire électronique*. PhD thesis, Université de Paris 7, 1993.

[52] P.-Y. F. P-A. Buvet. Classes d'objets et recherche sur le web. *Linguisticae Investigationes*, 23 :219–228(10), octobre 2001.

[53] G. Paliouras and V. Karkaletsis. Learning decision trees for named entity recognition and classification, 2000.

[54] A. Popescu-Belis. Évaluation numérique de la résolution de la référence : critiques et propositions. *T.A.L. : Traitement automatique de la langue*, 2000.

[55] L. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition, 1989.

[56] G. T. S. Hunston. A local grammar of evaluation. *Evaluation in Text : Authorial Stance and the Construction of Discourse*, pages 74 – 101, 2003.

[57] N. SAGER. Sublanguage : Linguistic phenomenon, computational tool, 1986.

[58] G. Salton and al. Extended boolean information retrieval. *Communications of the ACM*, pages 1022–1036, 1983.

[59] G. Salton and al. Introduction to modern information retrieval. *Mac Graw Hill*, 1983.

[60] A. Savary. *Recensement et descriptions des mots composés - méthodes et application*. PhD thesis, Université de Marne-la-Vallée, 14 December 2000.

[61] F. Schmidt. *Automatic Recognition of Organization Names in English Business News*. PhD thesis, Ludwig Maximilan Universität(LMU), Munich, 2004.

[62] J. Senellart. Tools for locating noun phrases with finite state transducers. *Proceedings of COLING-ACL*, pages 1212–1219, 1998.

[63] T. Shopen. Logical equivalence ist not semantic equivalence, 1972.

[64] M. Silberztein. *Dictionnaires électroniques et analyse automatique de textes : le système INTEX*. Masson, 1993.

[65] M. Silberztein. Transducteurs pour le traitement automatique des textes. *Le lexique-grammaire, Travaux de linguistique*, 1999.

[66] R. Sproat and C. Shin. *A Statistical Method for Finding Word Boundaries in Chinese Text*, volume 4, pages 336–351. Computer Processing of Chinese and Oriental Languages, 1991.

[67] M. Stricker. *Réseaux de neurones pour le traitement automatique du langage : conception et réalisatiion de filtres d'informations*. PhD thesis, Université Pierre et Marie Curie - Paris VI, 14 December 2000.

[68] S. Sun and e. al. Some issues on the statistical approach to chinese word identification. In *3rd international Conference on Chinese Information Processing*, pages 246–253, 1992.

[69] P. V. Svatek and al. Discovering company descriptions on the web by multiway analysis. In *Intelligent Information Processing and Web Mining, Springer Verlag*, 2002.

[70] A. Viterbi. The viterbi algorithm, 1973.

[71] W. Woods. Transition network grammars for natural language analysis. *Communications of the ACM*, 1970.

[72] Z. Wu and G. Tseng. Chinese text segmentation for text retrieval : Achievements ans problems. *Journal of the Amarican Society for Information Science*, 1993.

[73] D. L. X. Blanco. Dictionnaire électronique français-espagnol-catalan-arabe des noms des professions et des métiers. *Penser la Francophonie : Concepts, action et outils linguistiques. Actes des premières journées scientifiques communes des réseaux de chercheurs concernant la langue*, pages 131–142, mai 2004.

[74] Y. Yang. An evaluation of statistical approaches to text categorization. *Information Retrieval*, 1(1/2) :69–90, 1999.

[75] Y. Yang and J. O. Pedersen. A comparative study on feature selection in text categorization. In D. H. Fisher, editor, *Proceedings of ICML-97, 14th International Conference on Machine Learning*, pages 412–420, Nashville, US, 1997. Morgan Kaufmann Publishers, San Francisco, US.

[76] G. Zhou and J. Su. Named entity recognition using an hmm-based chunk tagger, 2002.

[77] G. K. Zipf. Human behavior and the principle of least effort, 1949.

Deutsche Zusammenfassung

Die vorliegende Dissertation wurde im Rahmen des Projekts "Suchmaschinen am CIS" erstellt. Ziel war es, eine Jobsuchmaschine für den französischen Stellenmarkt zu realisieren. Die Effizienz und Vollständigkeit der Suche sowie die Eindeutigkeit und Aktualität ihrer Suchergebnisse sollen durch linguistisch basierte Verfahren in allen Bearbeitungsphasen auf höchstem Niveau gewährleistet werden.

Der auf lange Sicht von hoher Arbeitslosenquote und häufigen Jobwechsel geprägte Arbeitsmarkt und die inzwischen zentrale Bedeutung des Internets bei der Arbeitsvermittlung haben uns ermutigt, soziales Interesse und technologischen Fortschritt in Einklang zu bringen. Unsere Vision ist, dazu beizutragen, alle verfügbaren Stellenangebote, die bis zum heutigen Tag auf parallelen Distributionswegen über das Internet verstreut sind, auf einer zentralisierten Plattform zu sammeln.

Gemäß einer Untersuchung, die der Personalvermittler *Kelly Services* im ersten Quartal 2006 mit 19.000 beteiligten Personen aus 12 europäische Länder durchgeführt hat, ist das Internet inzwischen das an erster Stelle genutzte Kommunikationsmittel auf der Suche nach einem Arbeitplatz. 70 % der Franzosen benutzen dieses Medium als primäre Quelle um eine Arbeit zu finden. Die Studie kommt zu dem Schluss, dass das Wachstumspotential des Internets bei der Arbeitsvermittlung

keineswegs ausgeschöpft ist. Wie eine weitere, von *FocusRH-Gruppe* durchgeführte Studie über die 500 wichtigsten Internet-Jobbörsen in Frankreich zeigt, wurden im Jahr 2006 mehr als 900.000 Stellen- und Praktikumsangebote über dieses Medium veröffentlicht. Diese Zahlen spiegeln allerdings nicht die ganze Wirklichkeit wider : Eine nicht zu vernachlässigende Anzahl dieser Stellen wird unter heterogenen Formaten gleichzeitig auf mehreren Jobbörsen publiziert. Des Weiteren wird ein Großteil der auf dem Arbeitsmarkt verfügbaren freien Stellen über andere Kommunikationswege wie die Webseiten der Firmen, spezialisierte Diskussionsforen oder auch Kleinanzeigen-Verzeichnisse veröffentlicht. Diese Kommunikationswege werden teils parallel zum Stellenangebot über eine Jobbörse, teils ausschließlich genutzt. Viele Unternehmen lehnen es ab, ihre Stellenanzeigen auf den einschlägig bekannten Jobbörsen zu platzieren, einerseits aus Kostengründen und andererseits, um die Zahl von Bewerbern zu minimieren, die nicht ihren Bedürfnissen entsprechen. Neben einem Redundanzproblem besteht also das Problem, dass es eine umfangreiche Menge von auf dem Netz verstreuten Stellenangeboten gibt, die in ihrer Mehrzahl für den interessierten Bewerber zumindest schwer erreichbar bleiben.

Viele Internet-Arbeitsvermittlungen mit unterschiedlichen Konzepten wurden und werden im anhaltenden Boom des Internets und seiner Technologien gegründet und versuchen sich auf dem Markt zu behaupten. Das Spektrum der Vermittlungsfirmen hat sicherlich dazu beigetragen, den Arbeitsmarkt transparenter und sowohl für den Arbeitgeber als auch den Arbeitssuchenden leichter zugänglich zu machen. Die Konkurrenzsituation hat allerdings zu einer Redundanz der Daten und einer Erhöhung des Noise geführt, anstatt eine Konvergenz des elektronischen Materials auf einer einzelnen gemeinsamen Plattform zu schaffen. So hat ein Großteil der Online-Anbieter in erster Linie die Quantität der publizierten Stellenanzeigen, der in der Datenbank angelegten Lebensläufe und der Besucher der Webseite im Blick. Es hat sich ein regelrechter Wettlauf entwickelt die größte Jobsuchmaschine zu sein. Die Qualität der Suchergebnisse sowie die soziale Verantwortung, die ein Arbeitsvermittler innehat, wurden dabei zumindest teilweise aus dem Blick verloren. Jobsuchmaschinen dieses Typs beschränken sich in der Transformationsphase der Daten einer Suchanfrage auf die traditionellen, rein statistischen Indexierungs- und Suchmethoden, anstatt den Vorteil der schon semi-struktuierten Daten zu nutzen und so die Qualität der gelieferten Suchergebnissen zu verbessern.

Eine optimierte Jobsuchmaschine sollte es dem Arbeitssuchenden erlauben sich rasch ein ganzheitliches Abbild des Arbeitsmarktes zu machen. Die Ergebnisse einer spezifischen Suche sollen praktisch alle relevanten Arbeitsangebote enthalten, aber auch nur diese. Die Darstellung der Suchergebnisse sollte über eine einzige Oberfläche und in Echtzeit für alle freien Stellen erfolgen, die aktuell im Internet veröffentlicht sind. Vor dem dargelegten Hintergrund mit der hohen Nutzungsquote des Internets als Medium zwischen Arbeitgeber und Arbeitnehmer, haben wir es uns zum Ziel gesetzt, eine solche optimierte Jobsuchmaschine für den französisch-sprachigen Raum zu entwickeln. Wir bieten ein System an, das alle online verfügbaren Stellenangebote findet und zusammenstellt. Eine linguistische Analyse jedes einzelnen Jobangebots gewährleistet höchste Selektivität und somit auch höchste Benutzerfreundlichkeit. Wir haben Lösungen erstellt, die es erlauben, die auf dem Netz verfügbaren Job-Web-Seiten zu erreichen und sie in semantisch strukturierte oder semi-strukturierte Einträge zu verwandeln. Dies ermöglicht allgemein eine Verbesserung der Leistung zukünftiger Informationretrieval Systeme.

Die vorliegende Dissertation ist aus 9 Kapiteln aufgebaut. Vor dem Hintergrund der aktuellen Situation des Online-Stellenmarktes wird eine einführende Beschreibung des Projektes, seiner Zielsetzung und Struktur gegeben (Kapitel 1). Es folgt ein Überblick über die verschiedenen existierenden Arten von Online-Arbeitsvermittlungen (Kapitel 2). Dabei wird im besonderen Maße auf die verschiedenen Typen der Jobsuchmaschinen eingegangen und ein Vergleich zwischen den 10 bekanntesten und am meisten benutzten Jobsuchmaschinen in Frankreich gezogen. In den folgenden Kapiteln werden der Stand der Technik der hier relevanten Bereiche in der gängigen Nomenklatur dargelegt und die jeweiligen Beschränkungen diskutiert. Die Teilaspekte des Projekts betreffen funktionelle Suchmaschinen und ihre internen Systeme (Kapitel 3), Techniken der Informationsextraktion (Kapitel 4) sowie elektronische Lexika des Formats DELA für Lexeme und insbesondere Mehrwortlexeme (Kapitel 5). In Kapitel 6 wird das neu erstellte *RecPAE System* zur automatischen Erkennung von Firmen-Homepages vorgestellt. Die Vorteile dieses Systems bei der Realisierung des Projekts sowie seine Grenzen werden erörtert. Das im Rahmen dieser Arbeit erstellte Lexikon der französischen Berufsbezeichnungen, das mittlerweile mehr als 80.000 flektierte Einträge enthält, wird in Kapitel 7 inklusive Klassifizierungsordnung, Bearbeitung und interner typologische Analyse präsentiert. Im letzten Kapitel wird der auf diesen beiden Vorarbeiten basierende zentrale Aspekt des neu implementierten Systems

dargelegt : die Informationsextraktion aus französischen Stellenanzeigen und ihre Umwandlung aus einer reinen Textform in ein semantisch strukturiertes Dokument. Eine Zusammenfassung und Diskussion zu Erweiterungsmöglichkeiten des Systems schließen die Arbeit ab.

Im Folgenden wird eine Zusammenfassung der in den Kapiteln 1 und 6-8 ausführlich erläuterten Arbeiten gegeben, die ich auf dem Weg zu der neu implementierten Jobsuchmaschine erbracht habe.

Unser System

Zur Realisierung der optimierten Jobsuchmaschine wurde im Rahmen dieser Arbeit ein Multi-Level System aufgebaut, das in all seinen Analyse- und Bearbeitungsphasen von linguistischen Theorien und im besonderen vom Konzept « lokaler Grammatiken » getragen wird. Das gesamte System basiert auf zwei hier ausgearbeiteten, interagierenden Modulen und erinnert so auf den ersten Blick mit den entsprechenden drei Hauptprozessen an den Aufbau einer gewöhnlichen Suchmaschine.

Entdeckung und Sammlung der Dokumente (Stellenanzeige)

Um die zentralisierte Plattform der Stellenangebote zu verwirklichen, sollen die im Internet verfügbaren Stellen automatisch gefunden und gesammelt werden. Angesichts der oben dargelegten Beobachtungen wurden dazu zwei Entdeckungsstrategien entwickelt : Über die erste Variante werden ausschließlich die Web-Seiten von Unternehmen auf freie Stellen durchsucht ; zu diesem Zweck wurde ein System zur automatischen Erkennung von Firmen-Homepages entwickelt. Die zweite Variante setzt einen fokalisierten Crawler ein, der durch die erlernte Terminologie der spezialisierten sub-Sprache von Stellenanzeigen jede gecrawlte Webseite in die Klassen »Stellenanzeige » oder « nicht Stellenanzeige »einordnet.

Erkennung von Firmen-Homepages

Eine Datenbank von Unternehmens-URLs ist Voraussetzung, die veröffentlichten Stellen auf den Firmen-Homepages zu finden. Eine solche Datenbank wird über das neu entwickelte *RecPAE-System* zur automatischen Erkennung von Firmen-Homepages erstellt und auf dem aktuellen Stand des Internets gehalten. Es handelt sich um ein automatisches binäres Klassifikationssystem, das für jede URL entscheidet, ob sie der Klasse « Organisation » angehört oder nicht.

Dieses System sucht nach in 10 Klassen organisierten Deskriptoren, die während der Lernphase als für die Entscheidungsfindung relevant definiert wurden. Die verschiedenen Analysephasen sind :

- Analyse der HTML-Struktur
- Analyse der URLs, der Meta-Informationen, des Titels
- Analyse und Klassifikation der Ankertexte in vordefinierte semantische Klassen
- Extraktion des Firmennamens
- Extraktion der Adresse, der Telefonnummer, der siret-Nummer, ...
- Extraktion typischer Formulierung und komplexer terminologischer Begriffe

Nach Identifizierung der Firmen-Homepages beginnt die Suche nach den Stellenangeboten, die auf den einschlägigen Webseiten veröffentlicht sind. Diese Suche erfolgt über eine Analyse der HTML-Struktur der Seiten. Zu diesem Zweck wurde während der Lernphase die Klasse der « Jobs », die den Ankertexten auf den Firmenwebseiten entsprechen, auf über 80 Sequenzen vervollständigt. Diese Ankertexte führen zu der jeweiligen Seite für offene Stellen des untersuchten Unternehmens.

Fokalisierte Crawler

Die zweite Strategie zur Sammlung von Stellenangeboten konzentriert sich auf die Terminologie der Sub-Sprache der Stellenanzeige.

Dabei wurden während der Lernphase eine Reihe von Floskeln und komplexen Formulierungen gesammelt, die sich in zwei Typen einteilen lassen : Einerseits nominale Phrasen, die die Stellenanzeigen semantisch strukturieren und sozusagen die Überschriften der Sektionen in der Anzeige darstellen. Diese wurden in 13 semantische Kategorien organisiert, die an die Spaltenstruktur unserer Datenbank erinnern. Andererseits Redewendungen, die spezifische Verben oder Nomen unserer Sub-Sprache beinhalten. Bei diesen handelt es sich um eine Reihe von Formulierungen, die sich eindeutig auf die Stellenangebote beziehen, da sie in einem anderen Zusammenhang keinen Sinn ergeben. Ihre Anwesenheit ermöglicht es unserem fokalisierten Crawler zu entscheiden, ob eine Web-Seite in die Kategorie „Stellenanzeige“ einzuordnen ist. Diese Formulierungen sind besonderes wichtig für den Fall unstrukturierter Angebote, die Floskeln des ersten Typs nicht enthalten.

Informationsextraktion und Umwandlung der Darstellung der Dokumente

Die zweite Phase des vorgestellten Systems – der Hauptteil der vorliegenden Arbeit – behandelt die Informationsextraktion und automatischen Umwandlung der reinen Textform eines Stellenangebots in ein semantisch strukturiertes Dokument. Wir haben uns zu diesem Zweck auf die Erstellung einer bedeutenden Anzahl lokaler Grammatiken und elektronischer Lexika konzentriert, die es uns erlauben, die Datenbank der Stellenangebote automatisch zu füllen. Die Struktur der Datenbank kann als ein Formular betrachtet werden, über das die in jeder Stellenanzeige vorliegende Information strukturiert wird. Die gängigen Konzepte von Jobsuchmaschinen – auch neuerer Generation- erfordern hingegen ein manuelles Ausfüllen der entsprechenden Formulare. Über ein im ersten Arbeitsschritt des Systems erstelltes Dokument versuchen wir folgende Informationen automatisch zu extrahieren :

- das Datum der Veröffentlichung der Anzeige
- das Datum des letzten Bewerbungsmöglichkeit
- das Einstellungsdatum
- der Name des Unternehmens, das einstellt
- die persönlichen Daten des Unternehmens : Postadresse, E-mail, Web-Seite, Telephon, Fax .
- der Kontakt, an den die Bewerbungsunterlagen geschickt werden

sollen
- der Tätigkeitsbereich des Unternehmens
- der Name der offene Stelle
- die Referenz der offene Stelle
- die Art des Vertrags
- die Dauer der Stelle falls befristet
- der Ort der offenen Stelle
- das vorgeschlagene Gehalt
- die gewünschte Arbeitserfahrung des Kandidaten
- die gewünschte Ausbildung des Kandidaten

All diese Informationen sind nicht notwendigerweise in jeder Stellenanzeige erwähnt. Einige Unternehmen beschränken sich auf ein Minimum an Informationen, besonders diejenigen, die die Anzeige direkt über ihren eigenen Webauftritt publizieren. Diese Anzeigen sind oft sehr kurz gehalten und verzichten darauf entsprechend der Ebene, auf der sich das Stellenangebot befindet, Informationen zur Firma wie Firmenname, (Kontakt-)Adresse oder Tätigkeitsbereich zu wiederholen, da diese dem menschlichen Besucher der Seite ja schon bekannt sind. Es wurden daher bestimmte Vorfahrtsregeln und Ausnahmen zur Routine hinzugefügt, die es erlauben, ein Dokument in der Klasse „Stellenangebot“ zu klassifizieren, selbst wenn nicht alle gesuchte Informationen gefunden werden.

Ausblick

Teile des hier konzipierten Systems einer optimierten Jobsuchmaschine können auch für andere Ziele benutzt werden. So hält das RecPAE System die Datenbank der Firmen mit einem Webauftritt immer auf dem aktuellen Stand. Eine solche Datenbank kann für die verschiedensten Anwendungen von großem Nutzen sein. So konsultieren immer mehr Menschen das Internet um beispielsweise einen Dienstleister oder Anbieter einer bestimmten Branche oder Region zu finden. Unser aktuelles Ziel ist es ein Klassifikations-System zu entwickeln, das automatisch jedes Stellenangebot in die entsprechende Berufsbranche einordnen kann. Zu diesem Zweck wird eine Ontologie der Berufsbezeichnungen erstellt, die es erlaubt, eine zukünftige Suche auf die semantischen Beziehungen zwischen den Suchbegriffen zu erweitern.