Reconnaissance des entités nommées dans des documents multimodaux

Mohamed Hatmi

UNIVERSITÉ DE NANTES
FACULTÉ DES SCIENCES ET DES TECHNIQUES

**ÉCOLE DOCTORALE SCIENCES & TECHNOLOGIES
DE L'INFORMATION ET MATHÉMATIQUES - STIM**

Année 2014

Reconnaissance des entités nommées dans des documents multimodaux

THÈSE DE DOCTORAT

Discipline : Informatique

Spécialité : Traitement Automatique des Langues

*Présentée
et soutenue publiquement par*

Mohamed HATMI

Le 20 janvier 2014, devant le jury ci-dessous

Président	Pascale SÉBILLOT, Professeur des universités, INSA de Rennes
Rapporteurs	Jean-Yves ANTOINE, Professeur des universités, Université François-Rabelais de Tours
	Sophie ROSSET, Directeur de recherche, LIMSI-CNRS
Examinateurs	Emmanuel MORIN, Professeur des universités, Université de Nantes
	Christine JACQUIN, Maître de conférences, Université de Nantes
	Sylvain MEIGNIER, Maître de conférences, Université du Maine

Directeur de thèse : Prof. Emmanuel MORIN

Co-encadrante de thèse : MCF. Christine JACQUIN

Co-encadrant de thèse : MCF. Sylvain MEIGNIER

À mes parents
À mes grands-mères

Remerciements

Je voudrais tout d'abord exprimer mes plus profonds remerciements à mon directeur de thèse Emmanuel Morin pour la confiance qu'il m'a accordée durant ces années de thèse, pour sa disponibilité et pour ses qualités d'encadrement. Je lui suis également reconnaissant pour son sens pédagogique et sa bonne humeur qui m'ont permis de mener à bien cette thèse et de trouver la force dans certains moments particulièrement difficiles.

Un grand merci également à mes deux co-encadrants, Christine Jacquin et Sylvain Meignier, qui m'ont bien encadré et qui m'ont consacré beaucoup de temps. J'ai pris un grand plaisir à travailler et à apprendre à leurs côtés.

Je souhaiterais également remercier l'ensemble des membres du jury, Jean-Yves Antoine, Pascale Sébillot et Sophie Rosset, pour m'avoir fait l'honneur d'évaluer ce travail et pour leurs lectures attentives et leurs remarques constructives.

Merci à tous les membres de l'équipe TALN et de l'équipe LST pour les bons moments passés ensemble et pour les discussions que nous avons pu avoir.

Merci à mes collègues et amis doctorants du LINA pour les moments passés ensemble. J'adresse également un merci tout particulier à mes collègues du bureau avec qui j'ai passé des moments mémorables : Rima, Prajol, Ramadan, Amir et Firas.

Je tiens tout spécialement à remercier Soumaya pour son soutien et sa patience tout au long de cette thèse.

Je conclurai en remerciant de tout cœur mes parents, mes sœurs, mes frères, mes tentes, mes cousins, mes cousines et mes amis. Merci à tous ceux qui, de près ou de loin, m'ont encouragé et soutenu tout au long de ce travail.

Nantes, le 25 janvier 2014.

Table des matières

Liste des figures

Liste des tables

Introduction

Contexte

L'accès à l'information numérique est devenu une activité de notre quotidien. S'exprimer via un blog ou partager ses vidéos sur un réseau social sont devenu des activités courantes. Cela est rendu possible par le développement continu des technologies de l'information et des réseaux qui procure un accès toujours plus aisé et rapide à l'information. De ce fait, la masse d'information stockée est devenu colossale et continue de s'accroître exponentiellement. Différents modes de communication sont utilisés afin de véhiculer l'information. Cela regroupe par exemple la vidéo, la parole et l'écrit. Dans ce milieu hétérogène, l'enjeu est donc de faciliter un accès rapide et pertinent à l'information désirée, sans que l'utilisateur soit perdu face à la quantité d'information qui lui est proposée. Plusieurs applications du Traitement Automatique des Langues Naturelles (TALN) s'intéressent à développer des méthodes et des outils pour répondre à ce défi, comme par exemple l'extraction d'information, la recherche d'information, l'indexation et la traduction automatique. Dans ces différents domaines, la tâche de Reconnaissance des Entités Nommées (REN) joue un rôle transversal.

Le concept d'entité nommée est apparu au milieu des années 90 comme étant une sous-tâche de l'activité d'extraction d'information. Elle consiste à identifier certains objets textuels tels que les noms de personne, d'organisation et de lieu. Le matériel textuel exploité était de langue anglaise et concernait des dépêches de presse écrite ciblées sur un domaine spécifique. Au fil des années, les recherches concernant ces objets linguistiques se sont focalisées sur des problématiques de plus en plus complexes comme la désambiguïsation et l'annotation enrichie mais aussi sur leur reconnaissance dans des contextes différents (autre langue et autre modalité). Le déploiement d'Internet a engendré une mutation en profondeur des sources d'information. Les messages publiés sur les réseaux sociaux tels que Twitter et Facebook ainsi que les flux de vidéo et d'audio gagne de plus en plus de terrain sur les sources d'information traditionnelles telles que les articles journalistiques publiés par les agences de presse. Le traitement des entités nommées se heurte désormais à des nouvelles difficultés inhérentes aux caractéristiques de la modalité ou du type de texte à traiter.

Le travail de cette thèse s'est particulièrement concentré sur la REN dans le cadre de la modalité parole.

Motivations et objectifs

La Reconnaissance Automatique de la Parole (RAP) se heurte à plusieurs phénomènes inhérents aux caractéristiques de l'oral. La transcription produite automatiquement représente une image bruitée plus ou moins fidèle de l'oral. Les chevauchements de parole, le bruit et les mots hors-vocabulaire sont des sources potentielles d'erreurs de transcription. La qualité de transcription dépend aussi du style de parole employé (parole lue ou parole spontanée). L'ordre de mots est plus libre lorsqu'il s'agit de parole conversationnelle (dislocation de type, absence de négation, etc.). Cela pose un problème pour la modélisation du langage puisque les modèles de langues utilisés par le système de RAP sont en grande partie appris à partir d'articles journalistiques. Des erreurs qui correspondent à une mauvaise modélisation du langage peuvent donc apparaitre dans les transcriptions. Une autre propriété de la parole qui peut engendrer des erreurs supplémentaires de reconnaissance des mots est la présence des disfluences (hésitations, lapsus, faux départs, répétitions, etc.). Ces dernières correspondent à toute interruption, perturbation ou accroc dans le flux de la parole. Dans cette transcription produite automatiquement, la REN doit donc faire face à ces deux difficultés caractéristiques du traitement automatique de la parole : la présence de disfluences et d'erreurs de reconnaissance.

En outre, un degré supplémentaire de difficulté est imposé à la REN dans les transcriptions automatiques de la parole en raison de l'absence de capitalisation et le manque de ponctuation. Cela induit respectivement des ambiguïtés graphiques et de segmentation. En effet, la plupart des systèmes de RAP ne considèrent pas la majuscule pour des raisons computationnelles. D'un autre côté, la segmentation de l'oral est différente de l'écrit. Les signes de ponctuation n'ont pas d'équivalent figé dans la parole. La segmentation des unités de sens en parole se fait par des marques inhérentes à la modalité orale telles que la mélodie et l'intonation.

Les travaux de cette thèse ont été réalisés dans le cadre du projet DEPART [1]. Notre but est d'étudier de manière approfondie la REN dans le contexte de la parole. Pour ce faire, nous explorons deux axes d'étude : le premier considère la REN comme un simple traitement qui vient en aval de la tâche de reconnaissance du signal. Le deuxième consiste à interagir directement avec le système de RAP. Nous étudions alors le couplage étroit entre la tâche de transcription de la parole et la tâche de REN.

Structure de la thèse

Outre la présente introduction qui indique le contexte et les objectifs de cette thèse, ce manuscrit est organisé en deux parties principales et se termine par une conclusion.

1. Documents Ecrits et Paroles - Reconnaissance et Traduction : projet financé par la région des Pays de la Loire (`www.projet-depart.org`)

La première partie est découpée en trois chapitres d'introduction générale. Le chapitre 1 est une présentation de l'état de l'art relatif à la tâche de REN. Nous commençons par nous interroger sur le statut théorique des entités nommées en dressant un inventaire chronologique des différentes définitions qui ont été proposées. Nous présentons ensuite les indices, les phénomènes d'ambiguïté et les différentes approches utilisées pour reconnaître les entités nommées. Le chapitre 2 met l'accent sur les difficultés que posent certains phénomènes spécifiques à une ou plusieurs modalités pour la REN. Nous analysons quelques systèmes de REN caractéristiques de la modalité traitée. Le chapitre 3 s'ouvre à la problématique du multilinguisme : comment les systèmes de REN peuvent être adaptés afin de couvrir des nouvelles langues ? Pour cela, nous explorons les différentes approches utilisées pour adapter les systèmes symboliques et les systèmes à base d'apprentissage entre les langues. Nous étudions ainsi l'intérêt et le coût associé pour porter un système existant de REN vers une autre langue. Nous proposons une méthode peu coûteuse pour porter un système symbolique dédié au français vers l'anglais.

La seconde partie est composée de deux chapitres qui présentent les différentes contributions de travaux de cette thèse. Le chapitre 4 étudie les spécificités de la REN en aval du système de RAP. Nous commençons par présenter une méthode pour la REN dans les transcriptions de la parole en adoptant une taxonomie hiérarchique et compositionnelle. Nous mesurons ensuite l'impact des différents phénomènes spécifiques à la parole sur la qualité de REN. Cela permet de mener une réflexion approfondie sur les problèmes posés à la REN en parole. En réponse à ces difficultés, nous explorons, dans le chapitre 5, une autre voie de recherche qui consiste à étudier un couplage étroit entre la tâche de transcription de la parole et la tâche de REN. Dans ce but, nous proposons une nouvelle approche permettant de détourner les fonctionnalités de base d'un système de transcription de la parole pour le transformer en un système de REN. Nous menons différents types d'expérimentations afin d'optimiser et d'évaluer notre approche.

Enfin, un dernier chapitre vient conclure ce manuscrit et proposer différentes perspectives.

Première partie

État de l'art

Reconnaissance d'entités nommées : présentation générale

1

Sommaire

La reconnaissance des entités nommées (REN) apparaît comme une composante essentielle dans plusieurs domaines du Traitement Automatique des Langues Naturelles (TALN) : analyse syntaxique, résolution de coréférence, traduction automatique, recherche d'information, etc. Cette tâche relativement récente s'est complexifiée au fil du temps sur différents axes : taxonomie, modalité, langue, etc.

Dans ce chapitre, nous donnons une présentation générale de la REN. Nous commençons par dresser, en section 1.1, un inventaire chronologique des différentes définitions qui ont été proposées. La section 1.2 présente les différents indices permettant de discerner les entités nommées. La section 1.3 présente les différents phénomènes d'ambiguïté qui complexifient la reconnaissance de ces dernières. La section 1.4 décrit les mesures de performance permettant d'évaluer la qualité de reconnaissance. Finalement, la section 1.5 présente et analyse les différents travaux réalisés à ce sujet.

1.1 Définitions et approche historique

Le concept d'entité nommée est apparue dans les années 90 à l'occasion de conférence d'évaluation MUC (*Message Understanding Conference*) (Grishman et Sundheim 1996). Ces conférences avaient pour but de promouvoir la recherche en extraction d'information. Les tâches proposées consistaient à remplir de façon automatique des formulaires concernant des événements. Dans ce cadre, certains objets textuels, ayant une importance applicative particulière dans plusieurs domaines du TALN, ont été regroupés sous le nom d'entités nommées. La reconnaissance de ces dernières est donc considérée comme une sous-tâche à part entière de l'extraction d'information.

D'un point de vue définitoire, la notion d'entité nommée a évolué au fil du temps, que ce soit au niveau de ce qu'elle signifie ou au niveau des typologies qu'elle peut couvrir. La plupart des définitions proposées sont plus énumératives que linguistiques. Les typologies adoptées sont proposées au regard du besoin applicatif. Peu de travaux ont tenté une « définition linguistique ». Nous nous référons notamment ici aux travaux d'Ehrmann (2008) qui propose une définition et une caractérisation des entités nommées prenant en compte la dimension linguistique et son application en TALN. Dans la suite de cette section, nous présentons certaines définitions et typologies présentes dans divers guides des campagnes d'évaluation et travaux.

Les conférences MUC ont été organisées entre les années 1987 et 1998. La tâche de reconnaissance d'entités nommées (REN) a été créée pour la première fois lors de la campagne d'évaluation MUC-6 (1995) :

« *... la tâche d'entités nommées consiste essentiellement à identifier les noms de toutes les personnes, les organisations et les localisations géographiques dans un texte* »[1].

Les entités nommées sont donc implicitement définies en évoquant une simple énumération de ce qu'elles peuvent représenter. Le thème du corpus journalistique utilisé, portant sur les mouvements de dirigeants, a influencé indirectement l'adoption de 7 catégories réparties en 3 types :

- Enamex : pour les noms de personne, d'organisation et de lieu (*Person*, *Organization* et *Location*) ;
- Timex : pour les expressions temporelles (*Date* et *Time*) ;
- Numex : pour les expressions numériques, de monnaie et de pourcentage (*Money* et *Percent*).

La figure 1.1 présente un exemple de texte annoté de MUC-6.

1. « *... the "named entity" task, which basically involves identifying the names of all the people, organizations, and geographic locations in a text.* » (Grishman et Sundheim 1996) (page 467).

Mr. <ENAMEX TYPE="PERSON">Dooner</ENAMEX> met with <ENAMEX TYPE="PERSON"> Martin Puris</ENAMEX>, president and chief executive officer of <ENAMEX TYPE="ORGANIZATION">Ammirati Puris</ENAMEX>, about <ENAMEX TYPE="ORGANIZATION">McCann</ENAMEX>'s acquiring the agency with billings of <NUMEX TYPE="MONEY">400 million</NUMEX>, but nothing has materialized.

FIGURE 1.1 – *Exemple de texte balisé de MUC-6*

Les mêmes éléments définitoires ont été repris lors de MUC-7 (Chinchor 1998) et les campagnes associées MET-1 et MET-2 (*Multilingual Entity Task Conference*) pour le japonais et le chinois (Chinchor 1998, Grishman et Sundheim 1996) :

« *Les entités nommées sont définies comme des noms propres et des quantités d'intérêt. Les noms de personne, d'organisation et de lieu ont été annotés ainsi que les dates, les heures, les pourcentages et les montants monétaires* »[2].

En changeant le thème du corpus, les organisateurs se sont rendu compte que certains noms propres importants ne sont pas inclus dans la définition MUC, par exemple, un nouveau type « AVION » s'avérait nécessaire lors de MUC-7 pour la reconnaissance des entités nommées dans un corpus portant sur les accidents aériens. Pour pallier ces lacunes, la couverture de ce que peut être une entité nommée a été étendue à d'autres types lors de la conférence japonaise IREX (*Information Retrieval and Extraction Exercise*) (Grishman et Sundheim 1996) :

- Artéfact : catégorie utilisée afin d'annoter les noms de produits (*Pentium II*), de prix (*Nobel Prize*), etc. ;
- Optionnel : catégorie utilisée afin d'annoter une entité pour laquelle la catégorie est ambiguë.

Les campagnes CoNNL (*Conference on Natural Language Learning*) ont été organisées en 2002 pour l'espagnol et le hollandais puis en 2003 pour l'anglais et l'allemand (Tjong Kim Sang et De Meulder 2003). Ces campagnes se positionnent sensiblement de la même manière que MUC :

« *Les entités nommées sont les syntagmes contenant les noms de personne, d'organisation, de lieu, les dates et les quantités* »[3].

La notion linguistique de syntagme est employée, mais la définition reste toutefois énumérative. Les entités nommées sont reparties sur quatre catégories : personne, organisation, lieu et divers. Cette dernière catégorie vise à annoter toutes les entités n'appartenant pas aux trois autres catégories, par exemple, les nationalités (italien) et les événements (Jeux Olympiques de 2000).

Les campagnes ACE (*Automatic Content Extraction*) ont été organisées entre les années 1999 et 2008 en se focalisant sur différentes langues telles que l'anglais, l'espagnol et l'arabe (Doddington et al. 2004). Ces

2. « *Named Entities (NE) were defined as proper names and quantities of interest. Person, organization, and location names were marked as well as dates, times, percentages, and monetary amounts.* » (Chinchor 1998) (page 1).

3. « *Named entities are phrases that contain the names of persons, organizations, locations, times and quantities* » (Tjong Kim Sang et De Meulder 2003) (page 155).

conférences incluent une tâche qui s'intéresse aux relations exprimées entre les entités nommées (en intégrant les co-références) dans les transcriptions automatiques d'émissions. La typologie adoptée introduit des changements importants par rapport à la catégorisation MUC en utilisant une catégorisation plus fine. Les entités nommées sont reparties sur sept catégories principales et 37 sous-catégories :

- *Facility* : *Airport, Building-Grounds, Path, Plant, Subarea-Facility* ;
- *Geo-Political* : *Continent, County-or-District, GPE-Cluster, Nation, Population-Center, Special, State-or-Province* ;
- *Location* : *Address, Boundary, Celestial, Land-Region-Natural, Region-General, Region-International, Water-Body* ;
- *Organization* : *Commercial, Educational, Entertainment, Government, Media, Medical-Science, Non-Governmental, Religious, Sports* ;
- *Person* : *Group, Indeterminate, Individual* ;
- *Vehicle* : *Air, Land, Subarea-Vehicle, Underspecified, Water* ;
- *Weapon* : *Biological, Blunt, Chemical, Exploding, Nuclear, Projectile, Sharp, Shooting, Underspecified*.

La catégorie « *Geo-political* » vise à annoter les entités géographiques ayant une propriété métonymique : par exemple, dans la phrase « *Miami imposed a curfew* », l'entité nommée « *Miami* » réfère simultanément à un lieu et à une organisation.

Les campagnes d'évaluation ESTER (Évaluation des Systèmes de Transcription d'Émissions Radiophoniques) visaient à la mesure des performances des systèmes de transcription d'émissions radiophoniques pour le français. Deux campagnes ont été organisées dans le cadre du projet EVALDA (Évaluation des technologies de la langue en français) : ESTER 1 (2003-2005) (Le Meur et al. 2004) et ESTER 2 (2006-2008) (Galliano et al. 2009). L'évaluation s'intéresse, entre autres, à l'extraction d'entités nommées. Ces dernières ont été définies comme suit :

« *Même s'il n'existe pas de définition standard, on peut dire que les entités nommées sont des types d'unités lexicales particuliers qui font référence à une entité du monde concret dans certains domaines spécifiques notamment humains, sociaux, politiques, économiques ou géographiques et qui ont un nom (typiquement un nom propre ou un acronyme).* » (Le Meur et al. 2004) (page 4)

Cette définition met en avant la particularité référentielle des entités nommées. La référence est le lien qui existe entre une expression linguistique et l'élément du réel auquel elle renvoie. Les entités nommées sont classées en 7 catégories principales, elles-mêmes divisées en sous-catégories :

- Personnes : humain réel ou fictif, animal réel ou fictif ;
- Fonctions : politique, militaire, administrative, religieuse, aristocratique ;
- Lieux : géographique naturel, région administrative, axe de circulation, adresse (adresse postale, téléphone et fax, adresse électronique), construction humaine ;
- Organisations : politique, éducative, commerciale, non commerciale, média divertissement, géo-socio-administrative ;

- Production humaine : moyen de transport, récompense, œuvre artistique, production documentaire ;
- Date et heure : date (date absolue, date relative), heure ;
- Montant : âge, durée, température, longueur, surface et aire, volume, poids, vitesse, autre, valeur monétaire.

L'étiquette « Incertain » est utilisée pour annoter les entités n'appartenant à aucune des catégories énumérées ci-dessus. Il est à noter que seuls les 7 catégories principales ont été prises en compte dans la campagne d'évaluation.

La campagne ETAPE (Évaluations en Traitement Automatique de la Parole) (2011-2012) avait pour objectif de mesurer les performances des technologies vocales appliquées à l'analyse des flux télévisés en langue française. Cette campagne s'inscrit dans la continuité des campagnes ESTER. En ce qui concerne la définition des entités nommées, celle d' Ehrmann (2008) a été utilisée :

« *Étant donnés un modèle applicatif et un corpus, on appelle entité nommée toute expression linguistique qui réfère à une entité unique du modèle de manière autonome dans le corpus.* » (Ehrmann 2008) (page 168)

Ehrmann (2008) présente une analyse de la notion d'entité nommée d'un point de vue linguistique et TALN. Les entités nommées sont présentées comme une création du TALN, c'est-à-dire qu'elles sont créées par la communauté TALN pour répondre à des besoins spécifiques aux différentes applications TALN. Le modèle applicatif correspond à la détermination et à la structuration d'un ensemble de données pertinentes au regard d'une application. La définition proposée se base sur deux critères définitoires principaux. Le premier est qu'une entité nommée représente une expression linguistique monoréférentielle. La monoréférentialité désigne la capacité d'une unité linguistique de renvoyer à une entité extralinguistique ou à un référent unique. Le deuxième est qu'une entité nommée représente une expression linguistique autonome, c'est-à-dire qu'elle permet à elle seule l'identification du référent. Ces deux critères correspondent, de manière privilégiée, aux noms propres qui ont un rapport direct à la référence et désignent des entités de manière autonome.

En ce qui concerne la typologie, la campagne ETAPE adopte une annotation fine selon le guide d'annotation Quaero (Rosset et al. 2011b). Ce guide comporte 8 catégories principales et 32 sous-catégories. La figure 1.2 montre la hiérarchie adoptée. Les entités nommées sont, quant à elles, annotées en composants en fonction de la catégorie en question. Par exemple, un nom de rue (<loc.oro>) peut se composer d'un genre (<kind>) et d'un nom (<name>) : <loc.oro> <kind> place </kind> de l' <name> <pers.ind> <title> Abbé </title> <name.first> Georges </name.first> <name.last> Hénocque </name.last> </pers.ind> </name> </loc.oro>.

D'autres typologies fines ont été proposées pour qu'elles soient adaptées à certaines applications du TALN telles que la traduction automatique et les systèmes de questions-réponses (Tran 2006).

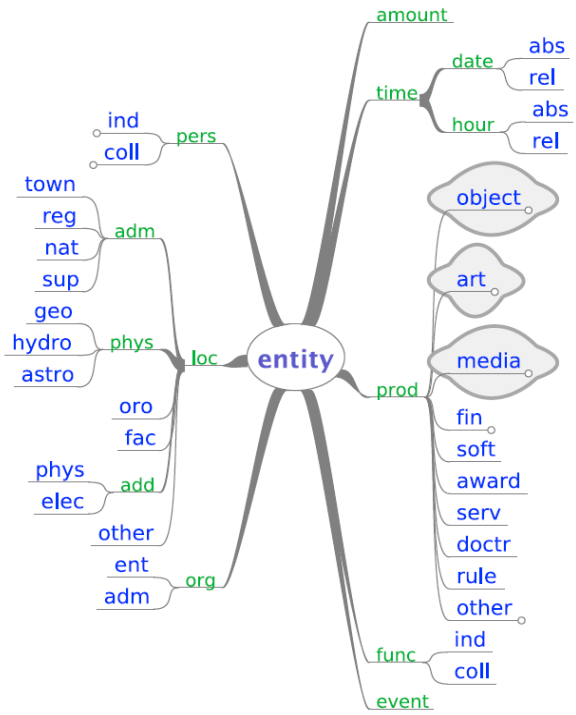

FIGURE 1.2 – *Hiérarchie des entités selon l'annotation Quaero (Rosset et al. 2011b)*

Dans une optique de recherche d'information, Paik et al. (1996) présentent une classification réalisée à partir d'une étude d'un corpus du journal *Wall Street Journal*. Cette classification comporte 9 catégories principales et 30 sous-catégories :

- *Geographic Entity* : *City, Port, Airport, Island, County, Province, Country, Continent, Region, Water, Geo. Misc* ;
- *Affiliation* : *Religion, Nationality* ;
- *Organization* : *Company, Company Type, Government, U.S. Gov., Organization* ;
- *Human* : *Person, Title* ;
- *Document* : *Document* ;
- *Equipement* : *Software, Hardware, Machines* ;
- *Scientific* : *Disease, Drugs, Chemicals* ;
- *Temporal* : *Date, Time* ;
- *Misc (divers)* : *Misc*.

Daille et al. (2000) présentent une version étendue de la typologie des noms propres proposée par le linguiste germanophone Gerhard (1985). Cette typologie est générique, c'est-à-dire qu'elle ne répond pas à un besoin applicatif particulier. Elle est constituée de 5 classes, chacune comportant des sous-catégories. Les catégories étendues ou créées apparaissent précédées d'un astérisque dans la table 1.1.

Sekine et al. (2002) proposent une hiérarchie d'entités nommées comprenant environ 150 types différents. Trois étapes ont été nécessaires pour réaliser cette hiérarchie :

1. construire trois hiérarchies initiales en menant trois études différentes : une étude manuelle en corpus (*Wall Street Journal, New York Times* et *Los Angeles Times*), une étude et une comparaison de différents typologies utilisées par les systèmes d'extraction d'information et d'autres travaux et une étude en se basant sur les thésaurus comme Wordnet et Roget Thesaurus.
2. combiner les hiérarchies résultantes.
3. affiner la hiérarchie résultante en étudiant d'autres corpus dans le but de trouver des catégories d'entités nommées ayant été oubliées.

Les différentes catégories obtenues sont illustrées dans la table 1.2.

Discussion

Dans cette partie, nous avons présenté le statut théorique des entités nommées dans le domaine du TALN. Plusieurs typologies ont été proposées pour les entités nommées : certaines ayant des catégories assez restreintes (par exemple, 4 pour CoNLL), d'autres ayant une catégorisation assez détaillée (par exemple, 150 dans Sekine et al. (2002)). Les trois catégories ENAMEX (personne, organisation et lieu) apparaissent indéniables, elles se trouvent dans les différents types de catégorisation. Toutefois, leurs sous-catégorisation se différencient entre elles, par exemple, on trouve prénom féminin et prénom masculin dans Sekine et al. (2002) et zoonymes (noms d'animaux de compagnie) dans Daille et al. (2000). Les autres catégories et leur granularité dépendent essentiellement du domaine (par exemple, les catégories *vehicle* et *weapon* pour ACE) et de l'application envisagée (par exemple, *Color* dans Sekine et al. (2002)).

L'adoption d'une topologie fine a l'avantage de pouvoir s'adapter à d'autres typologies moins fines en regroupant des catégories. En revanche, l'extension d'une topologie demande en général un travail laborieux. Il est à noter qu'en dépit de ces catégorisations fines qui ont été proposées, la plupart des systèmes de REN optent pour une catégorisation contenant un nombre limité de catégories.

Une fois une typologie adoptée, le problème revient à déterminer les indices pertinents permettant d'identifier les entités nommées. Nous présentons dans la section suivante les principaux indices.

1.2 Indices

Les indices permettent d'aider à la reconnaissance et la catégorisation des entités nommées. McDonald (1996) distingue deux types d'indices : internes et externes. Les premiers se rapportent à ce qui permet de discerner une entité nommée en se basant seulement sur les formes composant cette dernière. Les secondes en revanche s'intéressent à ce qui apparaît dans le contexte immédiat (formes situées à gauche et à droite de l'entité nommée) ou à partir d'un contexte plus large tel que le document ou le corpus.

ANTHROPONYMES	Patronymes	Dupont, Durant
	Prénoms	Alexandre, Jean-Paul
	Ethnonymes	L'Italien, les Français
	*Partis et autres organisations	Le PCF, l'ONU, l'Union européenne
	*Ensembles artistiques, groupes musicaux et troupes de théâtre	Les Toten Hosen, l'orchestre philharmonique de New York
	Pseudonymes	L'Ange vert
	Zoonymes	Médor
TOPONYMES	*Toponymes > Pays	Europe
	Pays	France, Sahara occidental
	*Pays < Toponymes > Villes	l'Ile de France
	Villes	Paris, Belo Horizonte
	Microtoponymes	Le Quartier Latin, Prenzlauerberg
	Hydronymes	La Manche, la Seine, le lac Ontario
	Oronymes	Les Andes, les Alpes
	Rues	La rue de la Paix, le Faubourg Saint Honoré
	Déserts	Le Sahara, le désert de Gobi
	Édifices	La Maison Blanche, la gare Montparnasse
ERGONYMES	Sites de production	Renault Wilword
	Marques de produits	Coca, Kleenex, Scotch
	Entreprises industrielles	Microsoft Corporation, Sud-Marine industrie
	Coopératives	Semences de Provence
	Établissement d'enseignement et de recherche	Université de Nantes
	Installations militaires	la ligne Maginot
	*œuvres intellectuelles	Matrix, l'Écume des jours
PRAXONYMES	Faits historiques	La Guerre de Cent Ans
	Maladies	La maladie d'Alzheimer
	*Événements culturels, sportifs, commerciaux, etc.	Le Festival du film de Berlin
	*Période historique	le Paléolitique
PHENONYMES	Catastrophes naturelles	Le cyclone Mitch
	Astres et comètes	La comète de Halley

TABLE 1.1 – *Typologie proposée par Daille et al. (2000) en se basant sur celle de Gerhard (1985).*

NAME	PERSON	Lastname ; Male_Firstname ; Female_Firstname
	ORGANIZATION	Company ; Company_Group ; Military ; Institute ; Market Political_Organization Government ; Political_Party ; Public_Institution Group Sports_Team ; Ethnic_Group ; Nationality
	LOCATION	GPE City ; County ; Province ; Country ; Region ; Geological_Region Water_Form ; Sea Astral_Body Planet ; Star Address Postal_Address ; Phone_Number ; Email ; URL
	FACILITY	GOE School ; Museum ; Amusement_Park ; Worship_Place Station_Top Airport ; Station ; Port ; Car_Stop Line Railroad ; Road ; Waterway ; Tunnel ; Bridge Park ; Monument
	PRODUCT	Vehicle Car ; Train ; Aircraft ; Spaceship ; Ship Drug ; Weapon ; Stock Currency ; Award ; Theory ; Rule ; Service ; Character ; Method_System Action_Movement ; Plan ; Academic ; Category ; Sprots ; Offence ; Art Picture ; Brodcast_Program ; Movie ; Show ; Music ; Printing Book ; Newspaper ; Magazine
	DISEASE	
	EVENT	Games ; Conference ; Phenomena ; War ; Natural_Disaster ; Crime
	TITLE	Position_Title
	LANGUAGE	
	RELIGION	
NATURAL_OBJECT		Animal ; Vegetable ; Mineral
COLOR		
TIME_TOP		TIMEX Time ; Date ; Era ; PERIODX Time_Period ; Date_Period ; Week_Period ; Month_Period ; Year_Period
NUMEX		Money ; Stock_Index ; Point ; Percent ; Multiplication ; Frequency ; Rank ; Age
	Measurement	Physical_Extent ; Space ; Volume ; Weight ; Speed ; Intensity ; Temperature ; Calorie ; Seismic_Intensity
	COUNTX	N_Person ; N_Organization ; N_Location N_Country ; N_Facility ; N_Product ; N_Event ; N_Animal ; N_Vegetable ; N_Mineral

TABLE 1.2 – *Typologie proposée par Sekine et al. (2002).*

1.2.1 Indices internes

Les principaux indices internes utilisés pour la reconnaissance des entités nommées sont :

- informations graphiques : la majuscule est une marque typographique qui sert à débuter chacune des formes composant une entité nommée. McDonald (1996) et Mikheev (1999) s'appuient seulement sur cet indice pour l'identification et la délimitation des entités nommées pour l'anglais : chaque mot (ou séquence de mots) ne se produisant pas dans une position ambiguë (par exemple : début d'une phrase, titre capitalisé) et dont la première lettre porte une majuscule est considéré comme entité nommée. En revanche, la majuscule ne permet pas généralement d'aider à la catégorisation des entités nommées sauf pour les acronymes (une seule forme contenant plusieurs majuscules) qui font référence, dans la plupart des cas mais pas toujours à des organisations ou des personnes.
- informations concernant la ponctuation et les caractères spéciaux et numériques : les signes graphiques peuvent être utilisés dans les acronymes et les noms des organisations et des produits. Par exemple, « I.B.M », « C&A », « O'Conner », etc. En revanche, les caractères numériques permettent d'aider à l'identification et à la catégorisation de certaines entités telles que les dates, les pourcentages et les noms des organisations. Par exemple, 2010, 3 Suisses, Élisabeth II, etc.
- informations morpho-syntaxiques : consiste à exploiter les résultat d'un étiquetage morpho-syntaxique, décrivant pour chaque mot sa catégorie grammaticale (nom propre, verbe, conjonction, etc.), afin d'élaborer des règles plus généralisatrices pour la délimitation et la catégorisation des entités nommées. Par exemple, si un mot est étiqueté en tant que nom propre alors il est fort probable qu'il fasse partie d'une entité nommée. Un étiqueteur morpho-syntaxique orienté données (Béchet et Charton 2010) ou connaissances (Vangelis et al. 1998) est utilisé afin d'étiqueter le texte.
- informations morphologiques : consiste à exploiter les informations morphologiques telles que le genre, le nombre, la forme canonique, les affixes. Par exemple, sachant que « Maria », « Marinela » et « Maricica » sont des prénoms féminins d'origines romaines, le mot « Mariana » peut être affecté à la même catégorie en se basant sur la partie commune du préfixe « Mari ». La même chose pour les suffixes, par exemple, « -escu » est un indicateur fort pour les anthroponymes en roumain, de même « -wski » en polonais, « -ovic » et « -ivic » en croato-serbe et « -son » en anglais (Silviu et David 1999).
- informations issues de lexiques : consiste à une simple interrogation des listes de noms propres et de mots clefs les plus courants. Ces listes sont préparées *a priori* manuellement ou automatiquement en utilisant des techniques d'apprentissage. Ces informations permettent dans la plupart des cas de catégoriser la forme à annoter. Par exemple, « Microsoft **Inc** », « la **Bourse** de Paris », « **Lionel** Jospin », etc.

1.2.2 Indices externes

Les indices externes se rapportent au contexte d'apparition de l'entité nommée (McDonald 1996). Ils sont nécessaires lorsque les indices internes sont ambigus pour ne pas aboutir à des erreurs de classification. Par exemple, « Washington » est à la fois une personne, une ville et un état américain. Des listes de mots déclencheurs peuvent être utilisées pour aider à catégoriser les entités nommées. Ces listes contiennent les mots qui sont susceptibles d'apparaître dans le contexte. Par exemple, un nom de personne est souvent accompagné d'un titre, d'un grade permettant d'indiquer des propriétés spécifiques. Par exemple, « **Monsieur** Washington », « **Mme** Denise », « l'**entraîneur** Aimé Jacquet », etc. D'autres indices externes peuvent aider à la délimitation et à la catégorisation tels que :

- la position du mot dans la phrase, par exemple, s'il est capitalisé et s'il ne se produit pas au début de la phrase alors il a une forte chance d'être une entité nommée.
- les informations concernant les autres occurrences de l'entité nommée potentielle dans le document ou dans le corpus. Par exemple, si le mot « Paris » se produit dans un contexte ambigu alors qu'une autre occurrence de ce mot dans le même texte est étiqueté comme un nom de lieu, alors la première occurrence peut être aussi considérée comme un nom de lieu.
- les meta-informations, par exemple, les balises XML et HTML pour la reconnaissance des entités nommées dans les documents structurés.

Malgré la présence d'indices internes et externes, la reconnaissance des entités nommées n'est pas une tâche facile. Cela est dûe notamment à la présence de plusieurs phénomènes d'ambiguïté.

1.3 Ambiguïtés

Une entité nommée est dite ambiguë quand elle est susceptible d'avoir plusieurs interprétations concernant sa délimitation ou sa typologie. Dans cette partie nous présenterons les différents phénomènes d'ambiguïté qui complexifient la tâche d'annotation des entités nommées.

1.3.1 Ambiguïtés graphiques

La majuscule comme un indicateur pour le repérage et la délimitation des entités nommées n'est pas simple à manipuler pour plusieurs raisons :

- une entité nommée peut contenir des formes commençant par une minuscule (par exemple, « le château de Versailles », « la faculté des Lettres », etc.) ou peut comporter une ou plusieurs majuscules dans d'autres positions (par exemple, « eBay », « WikiLeaks »).
- la première forme d'une phrase comporte aussi une majuscule, que ce soit une entité nommée ou non.

– l'emploi de la majuscule pour les noms propres n'est pas de règle dans toutes les langues : en allemand, tous les noms, communs ou propres, prennent une majuscule. En outre, il n'y a pas de notion de majuscules/minuscules pour les langues n'utilisant pas l'alphabet latin (par exemple, l'arabe, le chinois, etc.).
– la transcription automatique de la parole est souvent bruitée. La majuscule n'est pas toujours respectée.

1.3.2 Ambiguïtés sémantiques

À l'instar des noms communs, les entités nommées n'échappent pas à la polysémie, à l'homonymie et à la métonymie. Prenons les exemples suivants :

– **Orange** a invité M. Dupont.
– **Leclerc** a fermé ses magasins en Rhône-Alpes.
– **La France** a signé le traité de Kyoto.

Ces phénomènes complexifient la tâche de catégorisation, par exemple, est-il question de la ville d'Orange ou bien de la société de téléphonie ? De la personne Michel Edouard Leclerc ou de la chaîne de supermarché ? Faut-il préférer une annotation de France en tant qu' « organisation » ou « gouvernement » ou en tant que « lieu » ou « pays » ? (Ehrmann et Jacquet 2006).

1.3.3 Ambiguïtés liées à la délimitation

L'identification des limites d'une entité nommée se heurte à plusieurs problèmes :

– limite droite : le début d'une entité nommée est plus facile à identifier (présence d'une majuscule ou d'un mot déclencheur) alors que la limite droite est difficile à trouver car les mots qui suivent ne sont pas capitalisés, par exemple, « La Fédération nationale de la Mutualité française ».
– coordination : consiste à unir deux entités nommées en effaçant l'un des constituants communs, par exemple, « *Bill and Hillary Clinton flew to Chicago together last month (...)* ». Ici, le problème revient à séparer « Bill Clinton » de « Hillary Clinton ». Une analyse sémantique est nécessaire.
– imbrication : imbriquer une entité nommée dans une autre, par exemple, « le Comité exécutif de l' [Union des associations européennes de football] ».

Ces différents phénomènes d'ambiguïté suscitent toujours des vives discussions et plusieurs remises en cause des guides d'annotation, notamment dans le cadre des campagnes d'évaluation telles que ESTER et ETAPE. En effet, les systèmes de REN doivent respecter les règles d'annotation pour que leurs performances puissent être correctement évaluées.

Par exemple, inclure ou ne pas inclure les titres dans les noms de personne (<PERSON> M. Dorgan </PERSON> ou M. <PERSON> Dorgan </PERSON>).

Après avoir défini la notion d'entités nommées, les indices et les difficultés concernant leur reconnaissance, nous présentons dans la section suivante les différentes mesures utilisées afin d'évaluer les performances des systèmes de REN.

1.4 Mesures d'évaluation

Plusieurs mesures ont été définies pour évaluer les performances des systèmes de REN. Ces performances sont généralement mesurées en termes de rappel, de précision et de F-mesure (Rijsbergen 1979) ou en termes de SER (*Slot Error Rate*) (Makhoul et al. 1999). Plusieurs variantes pondérées existent dans la littérature pour calculer ces mesures.

Le rappel est défini par le nombre d'entités nommées correctement étiquetées au regard du nombre d'entités nommées étiquetées dans la référence. La précision est le nombre d'entités nommées correctement étiquetées au regard du nombre d'entités nommées correctement et incorrectement étiquetées. Le rappel et la précision peuvent être calculés sur la base du nombre de mots contenus dans les entités nommées au lieu du nombre d'entités nommées. La F-mesure représente la moyenne harmonique pondérée du rappel et de la précision.

$$Rappel = \frac{Nombre\ d'entités\ correctement\ étiquetées}{Nombre\ d'entités\ étiquetées\ dans\ la\ référence} \tag{1.1}$$

$$Précision = \frac{Nombre\ d'entités\ correctement\ étiquetées}{Nombre\ d'entités\ étiquetées\ fournis} \tag{1.2}$$

$$F-mesure = \frac{2 \times rappel \times précision}{rappel + précision} \tag{1.3}$$

D'un autre coté, le SER combine et pondère différents type d'erreurs. Plusieurs variantes du SER ont été définies en fonction des types d'erreurs prises en compte, des poids attribués à chaque type d'erreurs et de la base d'évaluation (nombre d'entités nommées ou nombre de mots contenus dans les entités nommées). Nous présentons ici la variante de SER utilisée afin d'évaluer les performances des systèmes de REN durant la campagne d'évaluation ESTER 2. La base de calcul est le nombre d'entités nommées. Les types d'erreurs sont définies comme suit :

- insertions (I) : le nombre d'entités détectées dans l'hypothèse n'ayant aucun mot commun avec une entité de la référence.
- suppressions (D) : le nombre d'entités de référence totalement manquées par le système.

- erreurs de type (T) : le nombre d'entités détectées dans l'hypothèse avec des frontières correctes mais une catégorie incorrecte.
- erreurs de frontière (F) : le nombre d'entités détectées dans l'hypothèse avec une catétorie correcte mais des frontières incorrectes.
- erreurs de type et de frontière (TF) : le nombre d'entités détectées dans l'hypothèse une catégorie et des frontières incorrectes.

$$SER = \frac{D + I + 0,5 \times T + 0,5 \times F + 0,8 \times TF}{Nombre\ d'entités\ étiquetées\ de\ la\ référence} \tag{1.4}$$

Les performances des systèmes de REN sont fortement liées aux approches employées. La plupart des systèmes développés sont basés principalement sur les indices internes et externes présentés dans la section 1.2. Nous présentons dans la section suivante les principales approches.

1.5 Réconnaissance des entités nommées

La plupart des systèmes de REN utilisent soit des approches orientées connaissances soit des approches orientées données. Les systèmes orientés connaissances sont fondés sur des lexiques (listes de prénom, de pays, etc.) et sur un ensemble de règles de réécriture. D'un autre côté, les systèmes orientés données sont basés sur un modèle appris à partir d'un corpus préalablement annoté. Afin de profiter des avantages de ces deux approches, d'autres systèmes combinent des techniques d'apprentissage automatique et des règles produites manuellement. Plusieurs systèmes de REN sont présentés en détails dans Poibeau (2001), Friburger (2002), Nadeau et Sekine (2007), Fourour (2004). Nous ne présenterons ici qu'un seul système pour chaque type d'approche.

1.5.1 Approches orientées connaissances

Pour les approches orientées connaissances, les règles d'extraction sont produites manuellement par des experts en se reposant essentiellement sur des descriptions linguistiques, des indices et des dictionnaires de noms propres et de mots déclencheurs. Ces règles prennent la forme de patrons d'extraction permettant de repérer et de classifier les entités nommées. Par exemple, si le mot déclencheur « Monsieur » (connaissance issue d'un lexique) précède un mot inconnu commençant par une majuscule, alors le syntagme peut être étiqueté comme un nom de personne. Les systèmes orientés connaissances permettent d'obtenir de bons résultats sur les textes bien formés (Galliano et al. 2009). Plusieurs systèmes utilisant cette approche ont été développés pour différentes langues telles que l'anglais (McDonald 1996, Wacholder et al. 1997, Kim et Woodland 2000) et le français (Brun et Ehrmann 2010, Stern et Sagot 2010, Maurel et al. 2011). Nous présentons ici le système français Nemesis.

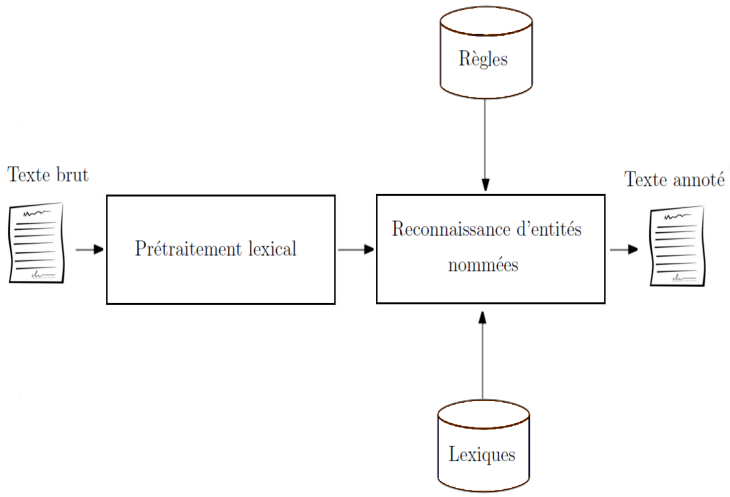

FIGURE 1.3 – *Architecture générale de Nemesis*

Nemesis : un système orienté connaissances de REN pour le français

Nemesis (Fourour 2002) est un système qui permet la délimitation et la catégorisation des entités nommées développé pour le français et pour du texte bien formé. Il se base essentiellement sur les indices internes et externes définis par McDonald (1996). L'architecture de Nemesis se compose principalement de trois modules qui s'exécutent séquentiellement : prétraitement lexical, projection des lexiques et application des règles.

1. prétraitement lexical : segmentation du texte en occurrences de formes et de phrases, puis association des sigles à leur forme étendue.
2. projection des lexiques : les lexiques ont été construits soit manuellement, soit automatiquement à partir du Web. Les éléments composant ces lexiques (79 476 éléments) sont répartis en 45 listes selon les catégories dans lesquelles ils sont utilisés : prénom connu, mot déclencheur d'un nom d'organisation (l'élément fait partie de l'entité nommée : « Fédération française de handball »), contexte d'un nom de personne (l'élément appartient au contexte gauche immédiat de l'entité nommée, mais ne fait pas partie de celle-ci : « philosophe Emmanuel Kant »), fin d'un nom d'organisation (l'élément est la dernière forme composant l'entité nommée : « Conseil régional », « Coupe du monde de football »), etc. La projection des lexiques consiste à associer les étiquettes liées aux lexiques aux différentes formes du texte. Une forme peut avoir plusieurs étiquettes (Washington | prénom-connu | lieu-connu).
3. application des règles : les règles de réécriture permettent l'annotation du texte par des balises identifiant les entités nommées (délimitation et catégorisation). Elles sont basées sur des étiquettes sémantiques référant à une forme capitalisée ou à une forme appartenant

à un lexique. En tout, Nemesis utilise 93 règles qui s'exécutent dans un ordre prédéfini. Lorsque plusieurs règles s'appliquent, Nemesis opte pour la règle ayant la priorité la plus élevée. Voici un exemple d'une règle de réécriture :

```
$Clé-oronyme $Article-min [ $Forme-capitalisée+ ]
→ ORONYME
```

et le résultat de son application :

« montagne du <ORONYME> Mont-Blanc </ORONYME> »

Pour améliorer les performances de Nemesis, de nouveaux lexiques sont appris automatiquement en se basant sur un ensemble d'heuristiques (22 pour les patronymes et 3 pour les toponymes).

L'évaluation de Nemesis a été réalisée sur un corpus composé de textes issus du journal Le Monde et du Web (31 000 mots). Les performances sur l'ensemble des entités nommées montrent un rappel de 79 % et une précision de 91 % (Fourour et Morin 2003).

1.5.2 Approches orientées données

Les approches orientées données visent à apprendre les règles d'extraction de manière autonome. L'acquisition de ces règles ainsi que de certaines ressources de connaissances s'effectue à partir d'un corpus de grande taille de manière supervisée ou semi-supervisée.

L'apprentissage supervisé consiste à apprendre les règles à partir d'un corpus préalablement annoté. La supervision concerne l'intervention humaine, par le biais d'étiquetage d'une base d'exemples, afin de guider le système lors du processus d'apprentissage. Une méthode d'apprentissage est appliquée pour entraîner le système à exploiter les différents traits singularisant les entités nommées. Ensuite le système d'apprentissage généralise le processus afin de produire un modèle permettant d'extraire les entités nommées dans de nouveaux documents. La performance des systèmes orientés données augmente proportionnellement avec la quantité et la qualité du corpus d'apprentissage. Les différents systèmes de ce type se basent notamment sur les méthodes d'apprentissage suivantes : machines à vecteurs de support (SVM) (Isozaki et Kazawa 2002), modèle de Markov à états cachés (HMM) (Bikel et al. 1997), modèle de l'entropie maximale (EM) (Andrew et al. 1998), modèle de champs conditionnels aléatoires (CRF) (Béchet et Charton 2010, Dinarelli et Rosset 2011, Raymond 2013, Hatmi et al. 2013) et arbres de décision (Isozaki 2001). Nouvel et al. (2013) proposent une méthode de fouille de données séquentielle hiérarchique permettant d'extraire des motifs d'extraction d'entités nommées à partir d'un corpus annoté.

Nous détaillons ci-dessous un exemple de système orienté données.

Le système de Raymond et Fayolle (2010) : un système orienté données de REN pour le français

Raymond et Fayolle (2010) utilisent différents algorithmes d'apprentissage automatique (CRF, SVM, et transducteurs à états finis (FST)) pour la reconnaissance d'entités nommées dans les transcriptions de la parole. Outre les mots de la transcription, trois traits ont été exploités afin de mieux caractériser les entités nommées : les étiquettes morpho-syntaxiques issues d'un processus d'étiquetage morpho-syntaxique (par exemple, nom propre, verbe, etc.), les connaissances issues des dictionnaires d'entités nommées et de mots déclencheurs, et l'importance du mot (un mot important est un mot dont l'information mutuelle partagée avec son étiquette d'entité nommée est supérieure à zéro et qui apparaît au moins trente fois dans le corpus d'apprentissage). Un corpus d'entraînement composé de 66h d'émissions radiophoniques francophones, dont 60h représentant le corpus d'apprentissage Ester 1 et 6h représentant le corpus de développement Ester 2, a été utilisé pour l'entraînement des différents classifieurs. Ce corpus a été transcrit et annoté manuellement. Le corpus de test représente 6h de transcriptions automatiques avec un taux d'erreur de mots de 26,09 % (corpus de test Ester 2). Les performances pour la reconnaissance d'entités nommées sont évaluées en termes de SER et de F-mesure. Les résultats d'évaluation montrent que les méthodes discriminantes (28,90 % de SER et 81 % de F-mesure pour le modèle SVM et 28,10 % de SER et 80 % de F-mesure pour le modèle CRF) obtiennent des performances meilleurs que les méthodes génératives (30,90 % de SER et 78 % de F-mesure pour le modèle FST).

Du fait que le corpus utilisé pour l'apprentissage est la fusion de deux corpus ayant une annotation légèrement différente (Ester 1 et Ester 2), les performances des différents classifieurs se voient affectées. Les auteurs proposent une méthode permettant de rendre les deux annotations cohérentes. Cela consiste à apprendre un modèle, en utilisant le système le plus performant (CRF), sur la concaténation des deux corpus. Ce modèle est utilisé ensuite pour ré-annoter le corpus d'entraînement. C'est ce nouveau corpus qui va servir de référence pour entraîner les différents classifieurs. Les résultats d'évaluation montrent une amélioration significative : 26,60 % de SER pour le modèle FST, 26,60 % de SER pour le modèle SVM et 22,80 % de SER pour le modèle CRF.

L'apprentissage semi-supervisé (ou légèrement supervisé) permet de réduire le travail manuel et laborieux exigé par les approches à base d'apprentissage supervisé pour annoter un corpus d'apprentissage. L'apprentissage semi-supervisé demande un corpus non annoté de grande taille et un petit degré de supervision sous la forme d'un ensemble d'amorces des règles ou de mots. La technique principale utilisée pour l'apprentissage est connue sous le nom d'amorçage qui consiste à apprendre automatiquement les patrons d'extraction en se basant sur un ensemble d'amorces pertinents par rapport à la classe sémantique des entités à extraire. Par exemple, afin d'extraire les noms de pays ou de villes, le système demande à l'utilisateur d'en introduire quelques-uns (par exemple, France, Tunisie, Paris, Nantes, etc.). Il analyse les phrases contenant les entités nom-

mées introduites et retient des indices contextuels communs (par exemple, marqueurs lexicales, typographiques, etc.). Il essaye ensuite de trouver d'autres entités nommées qui se produisent dans le même contexte. En réappliquant ce processus sur de nouveaux textes, le système apprend des nouveaux contextes pertinents et des nouveaux noms de pays ou de villes. Les systèmes basés sur cette approche affichent des résultats sensiblement identiques à ceux obtenus par les systèmes utilisant une approche à base d'apprentissage supervisé.

Michael et Yoram (1999) exploitent un corpus non étiqueté du *New York Times* et un certain nombre d'amorces des règles pour repérer et catégoriser les entités nommées. Les résultats d'évaluation de leur système affichent 91,10 % de rappel et 83,10 % de précision. Richard (2003) propose un système de reconnaissance des entités nommées à domaine ouvert (indépendant du domaine). La typologie des entités nommées n'est pas spécifiée *a priori* mais elle est identifiée automatiquement en fonction du contexte des documents. Pour chaque mot ou groupe de mots commençant par une majuscule, le système soumet une requête à un moteur de recherche (Google) afin d'identifier ses hyperonymes potentiels. Cette identification se base sur la méthode de Hearst (1992) concernant l'extraction d'hyponymes. Un exemple d'une telle requête est : *such as* X, avec X représentant une entité nommée commençant par une majuscule. Une fois collecté l'ensemble de ces hyperonymes, le système procède à leur classification afin d'identifier les catégories d'entités nommées qui apparaissent dans les documents. Le thésaurus WordNet est utilisé afin d'attribuer des noms des catégories aux classes résultantes. L'évaluation de cette méthode affiche un rappel égale à 67,39 % et une précision égale à 46,97 % en utilisant un corpus en anglais composé de 18 852 mots dont 1 051 entités nommées. Dans un travail récent, Tahmasebi et al. (2012) proposent une méthode non-supervisée pour la détection des entités nommées qui évoluent dans le temps. Leur méthode se base sur le fait que les entités nommées qui appartiennent à une même catégorie sont susceptibles de se produire dans des contextes proches lorsque l'écart entre les périodes de changement n'est pas grand. Au début, les périodes de changement concernant un terme donné sont identifiées sur toute la collection. Ensuite, un ensemble des mots contextuels est généré pour chaque période afin de détecter les mentions coréférentes qui se produisent dans le même contexte. L'évaluation de cette méthode montre un rappel important de 90 %.

1.5.3 Approches hybrides

Certains systèmes tirent profit des avantages respectifs des méthodes orientées connaissances et celles orientées données. Les règles sont, soient apprises automatiquement puis révisées manuellement (Nagesh et al. 2012), soient écrites manuellement puis corrigées et améliorées automatiquement (Andrei et al. 1998, Oudah et Shaalan 2012). Nous présentons ici un système basé sur cette approche.

LTG : un système hybride de REN pour l'anglais

Andrei et al. (1998) présentent le système LTG (*Language Technology Group*) lors de la compagne d'évaluation MUC-7. Ce système hybride a obtenu les meilleurs résultats lors de cette compétition. L'extraction des entités Enamex par LTG est faite comme suit :

1. passage des règles les plus sûres (*sure-fire rules)* : ces règles sont appliquées seulement si les indices internes et externes permettent de classer le candidat sans ambiguïté. Elles se présentent sous forme d'une liste de mots déclencheurs et un ensemble de règles contextuelles utilisant un étiquetage en parties du discours.
2. première reconnaissance partielle *(partial match 1*) : une fois les règles les plus sûres appliquées, le système génère des variantes d'entités nommées déjà reconnues en changeant l'ordre des mots ou en en supprimant. Ensuite, un algorithme probabiliste fondé sur le modèle de maximisation de l'entropie est utilisé pour l'étiquetage des noms propres.
3. passage des règles plus lâches (*rule relaxation*) : des règles plus lâches en termes de contraintes contextuelles sont appliquées. Cette étape permet aussi de résoudre le problème des conjonctions et celui des entités en début de phrases.
4. deuxième reconnaissance partielle (*partial match 2*) : cette reconnaissance partielle annote les noms propres en utilisant le modèle d'entropie maximale.
5. traitement des titres des articles (*title assignment*) : des règles et un algorithme probabiliste sont utilisés pour la désambiguïsation des entités nommées situées dans les titres car ces derniers sont entièrement écrits en majuscules.

Les résultats obtenus par ce système à la compétition MUC-7 montrent une F-mesure de 93,39 %.

Conclusion

Nous avons présenté dans ce chapitre le statut théorique des entités nommées ainsi que les indices permettant leur identification. Plusieurs systèmes ont été développés pour différentes langues. La plupart d'entre eux utilisent soit des méthodes orientées connaissances, soit des méthodes orientées données, soit des méthodes hybrides :

– les systèmes orientés connaissances sont les plus classiques lorsque la reconnaissance d'entités nommées est appliquée à des textes bien formés. Les règles sont écrites manuellement, ce qui demande une expérience linguistique et un effort humain conséquent. La reconnaissance est basée sur des indices internes et externes spécifiques à la langue en question. Des dictionnaires (noms propres, mots déclencheurs, etc.) sont éventuellement utilisés afin d'augmenter les performances.

- les systèmes orientés données sont les plus communs lorsque la reconnaissance d'entités nommées est appliquée à des textes bruités (par exemple, les transcriptions automatiques de la parole). Un modèle permettant d'étiqueter les textes est appris à partir d'un corpus annoté. Le problème revient à annoter manuellement ce corpus d'entraînement correspondant. Contrairement aux systèmes orientés connaissances, il est difficile d'intervenir manuellement pour changer les paramètres du modèle.
- les systèmes hybrides utilisent conjointement des techniques orientées connaissances et des techniques orientées données. Un ensemble de règles est écrites par un expert puis enrichi automatiquement en utilisant des techniques d'apprentissage, ce qui permet d'obtenir progressivement une meilleure couverture.

Ayant présenté les différentes approches utilisées pour la REN, nous procédons dans le chapitre suivant à l'analyse de quelques systèmes selon la modalité du texte traité.

Reconnaissance d'entités nommées et modalité

2

Sommaire

« *La modalité est la structure d'information concernant une modalité sensorielle ou une modalité de production (par exemple : vision, audition, toucher et odorat)* [1]. » (Allwood 2008).

« *Un corpus multimodal se compose de données appartenant à plus d'une modalité sensorielle ou à plus d'une modalité de production* [2] » (Allwood 2008).

L'information recherchée n'est pas obligatoirement textuelle mais elle peut être orale ou perçue. Les systèmes de REN doivent donc être adaptés afin de prendre en compte les spécificités de la modalité traitée. Les données non-textuelles telles que les données parlées et manuscrites sont généralement transformées en texte électronique avant que le processus de REN soit appliqué. La REN dans les documents textuels bien formés tels que les articles journalistiques affiche de bons résultats (F-mesure supérieure à 90 %). Cependant, cette tâche se heurte à plusieurs difficultés lorsqu'il s'agit de passer à d'autres modalités. L'absence de certains

1. « *The term "modality" can be used in many ways, but the definition we will adopt is that "multimodal information" is information pertaining to more than one "sensory modality" (i.e., sight, hearing, touch, smell or taste) or to more than one "production modality" (i.e., gesture, speech (sound), touch, smell or taste)* » (Allwood 2008) (page 207).

2. « *A multimodal digitized corpus is a computer-based collection of language and communication-related material drawing on more than one sensory modality or on more than one production modality.* » (Allwood 2008) (page 207).

indices importants et la présence des phénomènes spécifiques à la modalité traitée complexifient la tâche de REN. Nous présentons ici quelques types de systèmes de REN traitant des textes issus de différentes modalités. Nous nous intéressons dans la section 2.1 à la modalité écrite. La section 2.2 met l'accent sur la modalité parole. La section 2.3 concerne la modalité manuscrite. Finalement, la section 2.4 s'intéresse à la modalité vidéo.

2.1 La modalité écrite

La modalité écrite correspond aux données écrites sous une forme électronique (correspondant à la saisie sur un clavier d'ordinateur). Nous distinguons ici deux types de texte : le texte bien formé tels que les articles journalistiques et le texte bruité tels que les messages issus de réseaux sociaux.

2.1.1 Reconnaissance d'entités nommées à partir de textes bien formés

La REN a été réalisée de façon satisfaisante pour des textes bien formés tels que les articles journalistiques. Les systèmes de REN, qu'ils soient orientés connaissances, orientes données ou hybrides, affichent de bons résultats avec une F-mesure qui dépasse 90 % pour certaines langues telles que l'anglais et le français. La plupart de ces systèmes ont été initialement dédiés aux textes journalistiques soumis à des normes rédactionnelles strictes. Des règles de ré-écriture "rigides" sont utilisées afin de reconnaitre et de classifier les entités nommées. Ces règles se basent essentiellement sur des indices internes et contextuels.

La table 2.1 montre les résultats obtenus par certains systèmes sur le corpus de test MUC-7 [3] en termes de F-mesure.

Système	Approche	F-mesure (%)
LTG (Andrei et al. 1998)	Hybride	93,39
MENE (Andrew et al. 1998)	orienté données (supervisé)	92,20
IsoQuest (Krupka et Hausman 1998)	orienté connaissances	91,60
BALIE (Nadeau et al. 2006)	orienté données (semi-supervisé)	77,71

Table 2.1 – *Les résultats de certains systèmes de REN sur le corpus MUC-7*

Ces résultats montrent que les systèmes orientés données et les systèmes orientés connaissances obtiennent des résultats proches (0,60 % de différence dans le F-mesure entre le système MENE et le système IsoQuest). Les meilleurs résultats sont obtenus par le système hybride LTG. Les résultats obtenus pas le système à base d'apprentissage semi-supervisé sont moins bons.

2.1.2 Reconnaissance d'entités nommées à partir de textes bruités : SMS et messages issus de réseaux sociaux

Du fait que le nombre de caractères autorisés est limité, les SMS et les messages publiés sur les réseaux sociaux tels que Twitter et Facebook se caractérisent par l'omission de certaines informations en utilisant l'abréviation (écrire « L Blanc » pour « Laurent Blanc »), la phonétique (« koi »,

3. Catalogue LDC n° LDC2001T02

« jamè ») et le rébus typographique (« 2m1 » pour « demain »). Ces phénomènes rendent difficile la reconnaissance et la catégorisation des entités nommées. Une étape de normalisation est nécessaire pour optimiser les performances de REN. Cela consiste à ramener toutes les variantes des entités nommées à leurs formes canoniques non ambiguës (Khalid et al. 2008, Liu et al. 2012).

Les réseaux sociaux sont devenus une source d'information importante. Plusieurs travaux se sont intéressés à la REN à partir des commentaires et des messages publiés sur ces réseaux (Ritter et al. 2011, Liu et al. 2011, Wan et al. 2011). Nous présentons ci-dessous un système permettant la REN dans les messages publiés sur Tweeter.

Ritter et al. (2011) présentent un système qui permet la REN dans les tweets (messages publiés sur Tweeter). Ils montrent au début que l'étiqueteur en partie du discours (Stanford POS Tagger) et le chunkeur (OpenNLP) existants ne donnent pas de bons résultats lorsqu'ils sont appliqués sur des tweets. Ces outils sont donc adaptés en apprenant des nouveaux modèles à partir d'un corpus de tweets annoté manuellement (800 tweets). Cela permet une réduction de taux d'erreurs de 41 % pour l'étiquetage en partie du discours et de 22 % pour la segmentation en chunk. De plus, un classifieur a été entraîné afin de décider si les capitalisations qui apparaissent dans un tweet sont informatives ou non. Les sorties de ces différents outils sont utilisées comme des traits pour la REN. Un modèle permettant la délimitation et la segmentation des entités nommées est ensuite entraîné en utilisant un corpus composé de 2 400 tweets annotés manuellement. L'évaluation de ce système montre une F-mesure de 66 % sur un corpus de test composé de 2 400 tweets.

La difficulté de la reconnaissance des entités nommées se complexifie lorsqu'elle s'applique dans des documents issus d'autres modalités que la modalité écrite. Certains indices internes et externes ne sont plus exploitables. Plusieurs travaux traitent du problème de la REN en appliquant des méthodes développées initialement pour l'écrit sur les transcriptions automatiques. Cependant, certains phénomènes sont spécifiques aux transcriptions et n'apparaissent pas dans les textes écrits. Ces phénomènes ont un effet direct sur la qualité de REN.

2.2 La modalité orale

La reconnaissance de la parole consiste à transcrire manuellement ou automatiquement le signal audio en texte. Cette transcription peut être plus ou moins complète. Cela est dû notamment à certains phénomènes spécifiques à l'oral tels que les hésitations, les faux départs, les répétitions. À cela s'ajoute la manque de structuration (manque de casse et de ponctuation), les erreurs de la reconnaissance automatique de la parole (RAP) et les erreurs résultant des mots hors-vocabulaire pour les transcriptions automatiques. Ces différents phénomènes complexifient la tâche de REN.

Le table 2.2 montre les résultats officiels obtenus sur les transcriptions

manuelles par les différents systèmes participant lors de la campagne Ester 2 (Galliano et al. 2009) en termes de SER, précision et rappel.

Système	Approche	SER (%)	Précision (%)	Rappel (%)
Xerox	orienté connaissances	9,80	93,60	91,50
Synapse	orienté connaissances	9,90	93,00	89,30
LIA	orienté données (CRF)	23,90	86,40	71,80
LIMSI	orienté connaissances	30,90	81,10	70,90
LI Tours	orienté connaissances	33,70	79,30	65,80
LSIS	orienté données (CRF)	35,00	82,60	73,00
LINA	orienté connaissances	37,10	80,70	55,40

TABLE 2.2 – *Les résultats de la campagne Ester 2 sur les transcriptions manuelles*

Ces résultats montrent que les systèmes orientés connaissances employant une analyse syntaxique profonde (Xerox et Synapse) obtiennent les meilleurs résultats pour la transcription manuelle. Le premier système orienté données (LIA) obtient environ 14 points de SER de plus par rapport aux deux premiers systèmes. En outre, on note que les performances de ces systèmes sont dégradées par rapport aux résultats obtenus sur les articles journalistiques : par exemple, pour Nemesis (LINA), le rappel passe de 85,90 % sur les articles journalistiques à seulement 55,48 % sur les transcriptions manuelles de journaux radiophoniques.

D'un autre coté, les résultats sur les transcriptions automatiques montrent une dégradation allant de 15 à 30 % de SER pour les différents systèmes participants (transcription automatique avec un WER [4]de 12,11 %). Un corpus d'apprentissage composé de transcriptions manuelles annotées en entités nommées est utilisé afin d'effectuer l'apprentissage pour les systèmes orientés données. Le système LIA_NE (Béchet et Charton 2010) arrive en première place avec un SER de 43,40 % sur les transcriptions automatiques (WER de 12,10 %). Ce système se base sur un modèle génératif (HMM) et un modèle discriminant (CRF). Le HMM est appliqué en premier lieu en jouant le rôle d'un étiqueteur syntaxique enrichi avec des étiquettes sémantiques pour les noms propres (personne, organisation, lieu et produit). Le but est de simplifier le processus d'entraînement pour le CRF en limitant le nombre de traits. Le modèle CRF est appliqué aux sorties du HMM afin de trouver les entités nommées en se basant sur les traits contextuels des mots et de leurs étiquettes.

Les approches traditionnelles qui se positionnent en tant que post-

4. Word Error rate : utilisé pour mesurer les performances des systèmes de transcriptions de la parole. Cette mesure indique le pourcentage de mots incorrectement reconnus dans la transcription par rapport au texte de référence. Les types d'erreurs considérés sont définis comme suit :
   - Substitution (S) : le nombre de mots de la référence incorrectement reconnus dans la transcription ;
   - Suppression (D) : le nombre de mots de la référence omis dans la transcription ;
   - Insertion (I) : le nombre de mots ajoutés dans la transcription et absents de la référence.

   Le WER est donné par l'équation 2.1 :

$$WER = \frac{S + D + I}{Nombre\ de\ mots\ de\ la\ référence} \tag{2.1}$$

.

traitement au processus de transcription du signal n'ont en général pas donné une totale satisfaction concernant l'amélioration de la REN. D'autres approches sont apparues pour réaliser un couplage plus ou moins étroit entre le processus de transcription du signal et la REN.

Parada et al. (2011) ont proposé d'inclure des caractéristiques propres aux mots hors-vocabulaire dans le système de RAP. Un étiqueteur à base de CRF exploite ainsi les sorties d'un détecteur de mots hors-vocabulaire pour identifier ou ignorer des régions dans lesquelles les entités nommées sont incorrectement transcrites.

D'autres travaux exploitent des sorties intermédiaires des systèmes de reconnaissance comme les N meilleures hypothèses (Zhai et al. 2004), les graphes de mots (Favre et al. 2005) ou les réseaux de confusion (Kurata et al. 2012). Zhai et al. (2004) annotent les N meilleures hypothèses en sortie du système de RAP à l'aide d'un système qui s'appuie sur le principe d'entropie maximale. Un vote à partir des scores obtenus par le processus de transcription du signal et le système de REN est ensuite mis en place pour déterminer l'entité nommée la plus probable (même si elle n'était pas présente dans la meilleure hypothèse du système de RAP).

Kurata et al. (2012) exploitent les réseaux de confusion construits en sortie du système de RAP. Les nœuds du réseau sont d'abord regroupés selon leur similarité (*clustering*) puis, à l'aide d'un algorithme s'appuyant sur le principe d'entropie maximale, le modèle relatif au système de REN est appris en prenant en compte les clusters précédemment déterminés. Ces différentes approches ont permis d'améliorer la REN sur la modalité parole. Cependant, il n'est pas évident de comparer les résultats obtenus car les expérimentations n'ont pas toujours été menées dans des contextes similaires (transcription du signal plus ou moins difficile selon la tâche ciblée et détection des entités nommées plus ou moins difficile suivant le type de média audio : dialogue enregistré dans des centres d'appel, émissions radiophoniques...).

Les types d'approches présentées précédemment se focalisent principalement sur l'amélioration des performances de la REN qui est la tâche finale visée mais ne cherchent aucunement à améliorer le système de RAP en termes de WER par exemple. En ce sens, Wang et al. (2003) ont montré que, dans le contexte d'une autre tâche qui est celle de la compréhension de la parole, la recherche en premier lieu de la diminution du WER du système de RAP est souvent moins efficace que la prise en compte dans les modèles des spécificités de la tâche à traiter. Cependant, il est communément admis qu'une meilleure performance du système de RAP devrait aider à améliorer celle de la tâche visée. Dans cette optique et toujours pour le domaine de la compréhension de la parole, d'autres auteurs ont choisi de traiter de manière plus conjointe les différents problèmes posés. Servan et al. (2006) et Deoras et al. (2013) préconisent de commencer par apprendre de manière indépendante les différents modèles puis de construire le modèle final à travers une combinaison pondérée des différents modèles initiaux. Dans ce contexte, les premiers auteurs utilisent une approche générative (automates à états finis) qui se heurtent à certains problèmes. D'une part, l'approche ne permet pas de capturer toutes

les caractéristiques des modèles, et d'autre part, elle conduit quelquefois à une meilleure détection des étiquettes sémantiques au détriment du WER. Deoras et al. (2013) mettent aussi en œuvre une approche semblable. Ils conjuguent aussi, mais de manière différente, les scores obtenus par les différents modèles sur le graphe de mots en sortie du système de RAP. Ils utilisent de plus une approche discriminative (à base de CRF) qui leur permet de capturer plus de caractéristiques des modèles qu'une approche à base d'automates à états finis. Avec cette stratégie, ils obtiennent à la fois une amélioration du WER du système de RAP ainsi que des performances de la tâche de compréhension.

D'autres auteurs ont proposé de reculer encore les limites du couplage de la tâche de transcription du signal et de celle de REN en intégrant complètement ces deux tâches en une seule tâche (Hori et Atsushi 2006). L'idée principale est d'annoter en entités nommées les corpus d'apprentissage du système de RAP afin que celui-ci puisse directement générer une transcription annotée en entités nommées. Les entités nommées – éventuellement des mots composés – sont alors directement intégrées dans le lexique et dans le modèle de langage du système de RAP. L'expérimentation a été menée dans le cadre d'un système de question-réponse en langue japonaise et la tâche consistait à extraire les entités nommées dans 500 questions (le système de RAP utilisé ne comportait qu'une passe). Les résultats obtenus semblent prometteurs mais ils n'ont pas été comparés à ceux obtenus, dans les mêmes conditions d'expérimentation, par un système de REN état de l'art. De plus, la tâche de REN, dans des questions, semble être une tâche relativement simple du point de vue de la complexité des entités nommées.

Nous détaillons ci-dessous le système de Favre et al. (2005).

Le système de Favre et al. (2005) : reconnaissance d'entités nommées dans les graphes de mots

Une approche à deux niveaux a été utilisée dans Favre et al. (2005) afin de rechercher les entités nommées directement dans le graphe de mots issu du système de transcription de la parole :

- constitution d'une grammaire d'entités nommées : cette grammaire, implémentée par un transducteur à états finis, est composée d'un ensemble de règles de réécriture obtenue sur des corpus étiquetés. Elle prend en entrée n'importe quelle séquence de mots et génère en sortie tous les étiquetages possibles. Les mots du lexique utilisés par le système de transcription automatique de la parole sont exploités dans la grammaire afin de contrôler la généralisation des règles.
- composition entre le graphe de mots et la grammaire représentée sous forme de graphe : cette étape consiste à fusionner les hypothèses de transcription et d'étiquetage. Ensuite, les chemins sont réévalués en utilisant un étiqueteur HMM afin de construire la liste des n-meilleures séquences de mots correspondant à chaque type d'entité.

Le corpus d'entraînement Ester, composé d'émissions radiophoniques entre 2002 et 2004, a été utilisé pour entraîner le système. En revanche, deux corpus ont été utilisés pour le test : un corpus contenant des émissions enregistrées à la même période que le corpus d'entraînement et un corpus enregistré six mois plus tard. Les performances sont évaluées en termes de SER et de F-mesure. Les résultats obtenus pour les deux corpus sont différentes : la F-mesure perd 10 points, de 73 % (41 % de SER) pour le corpus ayant une date proche de celle du corpus d'apprentissage à 63 % (54 % de SER) pour le corpus avec un décalage de six mois. Cela s'explique par le fait que plusieurs entités nommées qui apparaissent dans le deuxième corpus ne sont pas reconnues par le système de transcription de la parole (mots hors-vocabulaire).

Un corpus de texte journalistique correspondant aux dates de diffusion des émissions dans le corpus de test a été collecté afin d'adapter les modèles d'entités nommées. Les auteurs montrent que, pour chaque jour, 25 % des entités nommées sont communes entre le corpus de test et le corpus journalistique collecté. Ce corpus journalistique est rajouté au modèle de langage de système de RAP. Les entités nommées se produisant deux fois ou plus dans le corpus journalistique sont rajoutées au lexique de système de RAP, aux dictionnaires d'entités nommées et à la grammaire de REN. L'amélioration globale de la détection de toutes les entités nommées est marginale (un gain de 1 % en F-mesure et de 3 % en rappel Oracle) pour chaque jour.

Certaines tâches du traitement automatique de la parole s'intéressent à l'identification de certaine entités nommées particulières. Par exemple, la tâche d'identification nommée du locuteur (INL) qui consiste à déterminer qui parle et quand en attribuant au locuteur un prénom et un patronyme. Une étape fondamentale de cette tâche consiste à transcrire le signal audio et à reconnaitre certaines entités nommées (Bigot et al. 2013). La phase de détection d'entités nommées peut être réalisée à l'aide de systèmes de REN. Les entités nommées de type personne sont ensuite exploitées pour extraire les noms complets de la transcription, les autres entités nommées vont servir de classes sémantiques pour généraliser la transcription. Nous présentons ici le système Milesin.

Milesin : un système d'identification nommée du locuteur par analyse conjointe du signal et de sa transcription

Milesin (Jousse 2011) est un système d'INL complètement automatique. Le système commence par identifier les entités nommées dans les transcriptions en utilisant le système de REN LIA_NE. Les noms de personne sont ensuite utilisés comme point de départ pour la phase d'attribution locale. Dans cette phase, un arbre de décision binaire reposant sur le principe des arbres de classification sémantique (SCT) est utilisé afin d'associer à chaque nom détecté dans la transcription quatre scores correspondant aux probabilités de chacune des étiquettes : précédent, courant, suivant. Ces étiquettes représentent la position du nom détecté dans le tour de parole. À partir des scores obtenus, une décision locale est prise en attri-

buant l'étiquette la plus probable à l'exemple. Pour les décisions correspondant aux étiquettes « courant », « suivant » ou « précédent », un nom de locuteur potentiel et son score sont alors associés aux tours de parole correspondants puis aux locuteurs dans une seconde phase de décision.

L'évaluation de Milesin est effectuée sur des transcriptions d'émissions radiophoniques en français provenant de la campagne ESTER 1 et du projet ANR EPAC. Le corpus de test contient 10 heures de données (1 194 tours de parole) dans lesquels 1 102 noms complets sont détectés. Les résultats montrent un rappel de 86,22 % à 95 % de précision sur les transcriptions manuelles. En revanche, les performances sur les transcriptions automatiques sont moins bonnes et n'affichent que 38,77 % de rappel et 70,87 % de précision.

Outre les transcriptions de la parole, les transcriptions issues d'un processus d'OCR se caractérisent aussi par l'existence de phénomènes caractéristiques à la modalité qui complexifient la REN.

2.3 La modalité manuscrite

La reconnaissance optique de caractères (OCR pour *Optic Character Recognition*) consiste à transformer un fichier image contenant un texte écrit en fichier texte. Il s'agit donc de reconnaitre, dans une image, les lettres composant un texte. Tout comme dans les transcriptions de la parole, les transcriptions issues d'un processus d'OCR ne sont pas parfaites et contiennent des erreurs commises par le système de reconnaissance automatique de l'écriture. Par contre, la majuscule et la ponctuation sont conservées dans la transcription si l'image initial en contient. Il a été démontré dans Miller et al. (2000) que pour chaque augmentation de 1 % de WER, le taux de F-mesure pour les entités nommées se dégrade de 0,7 point. Dinarelli et Rosset (2012) proposent une étape de correction des erreurs de transcription avant de commencer le processus de REN. Cela consiste notamment à corriger certaines erreurs de segmentation en phrases et en mots et de transcription. Nous détaillons ici quelques travaux concernant l'extraction d'entités nommées à partir des transcriptions issues d'OCR.

Le système de REN présenté dans (Grover et al. 2008) permet la reconnaissance des noms de personne et de lieu dans des documents provenant des archives du parlement britannique datant de XVIIe et XIXe siècles. Les documents sont scannés et transformés en fichier texte en utilisant un système de reconnaissance de l'écriture. Les transcriptions obtenues sont erratiques à cause des erreurs de transcription (par exemple, la confusion entre les notes marginales et les corps de texte) et de la nature des documents traités (présence de plusieurs caractères non-alphabétiques, utilisation fréquente des guillemets, absence de capitalisation pour certains noms propres, etc.). Un système orienté connaissances de REN est ensuite appliqué sur ces transcriptions. Les lexiques et les règles ont été adaptés afin de prendre en compte les phénomènes spécifiques aux transcriptions. Les résultats d'évaluation montre une F-mesure d'environ 71 %.

Dans (Packer et al. 2010), le but est de reconnaitre les noms de personne à partir de documents historiques. Quatre systèmes de REN utilisant des approches différentes ont été appliqués sur les transcriptions de ces documents : à base de lexique, à base de règles, à base de modèle de Markov à entropie maximale (MEMM) et à base de CRF. Les meilleurs résultats sont obtenus en combinant les résultats de ces différents systèmes.

Subramanian et al. (2011) proposent de rechercher les entités nommées directement dans le graphe de mots issu du système d'OCR plutôt que dans la sortie finale. L'évaluation est réalisée en utilisant deux systèmes d'OCR : le premier pour la reconnaissance de texte dans les documents vidéos en anglais et le deuxième pour la reconnaissance de texte manuscrit en arabe. Les résultats montrent un gain en rappel par rapport à la reconnaissance dans la sortie finale de 18,30 % pour l'anglais et de 4,60 % pour l'arabe.

2.4 La modalité vidéo

Plusieurs travaux proposent d'exploiter conjointement différentes modalités afin d'améliorer la qualité de REN. Cela concerne notamment la modalité vidéo. Nous détaillons ici un système visant à identifier les noms de personne dans des documents vidéo.

Percolo : un système multimodal de détection de personnes dans des documents vidéo

Percolo (Bechet et al. 2012) est un système permettant de détecter la présence de personnes dans des documents vidéo. Cela consiste à identifier et nommer à chaque instant d'une vidéo l'ensemble des personnes présentes à l'image. Percolo se base sur une analyse conjointe des indices identifiés à partir de données issues de deux modalités différentes :

- parole : segmentation et regroupement en locuteurs, transcription automatique et extraction d'entités nommées (en utilisant le système de REN LIA_NE).
- image : détection et suivi de visages, reconnaissance automatique des textes incrustés dans les vidéos.

Une fois les noms des locuteurs et des visages potentiels déterminés, un processus de fusion multimodale se charge de propager ces noms détectés vers les segmentations en visages et locuteurs afin de prendre les décisions finales sur la présence d'une personne dans le document. Les premiers résultats du système Percolo montre un rappel de 48,80 % et une précision de 64,20 % sur un corpus de test composé de 3 heures de signaux vidéos.

Conclusion

La REN donne de bonnes performances sur l'écrit (texte bien formé). Ces performances se voient dégradées lorsqu'il s'agit de textes bruités. Cette dégradation est due notamment à des problèmes liés à l'ambiguïté graphique (manque de capitalisation), à des problèmes de segmentation (manque de ponctuation) ainsi qu'à des problèmes intrinsèques à la modalité en question telles que les disfluences pour la modalité orale. Pour pallier ces problèmes, certains restaurent les ponctuations et la capitalisation dans la transcription (Gravano et al. 2009). D'autres, comme Sudoh et al. (2006), détectent les entités nommées qui ont été mal retranscrites afin qu'elles ne soient plus prises en compte par le processus de REN.

Les approches traditionnelles de REN se positionnent en tant que post-traitement au processus de transcription que se soit pour la modalité orale ou manuscrite ; elles n'ont en général pas donné une totale satisfaction concernant l'amélioration de la REN. D'autres approches sont apparues pour réaliser un couplage plus ou moins étroit entre le processus de transcription et la REN en exploitant des sorties intermédiaires comme les N meilleures hypothèses ou les graphes de mots. Les résultats obtenus montrent que la REN à partir des textes bruités pose encore de nombreuses difficultés.

Outre les problèmes liés à la modalité du texte, la dimension multilingue n'est pas aussi facile à prendre en compte. Nous exposons dans le chapitre suivant les différents travaux réalisés à ce sujet.

Reconnaissance d'entités nommées et multilinguisme

3

Sommaire

La REN se base essentiellement sur des indices propres à une ou quelques langues, par exemple, la majuscule pour l'anglais et le français. La génération des règles, que ce soit manuellement ou automatiquement à travers un corpus annoté, requière des connaissances linguistiques considérables. Certaines langues se caractérisent par des traits plus discriminants pour la distinction des entités nommées que dans d'autres langues. Les ressources multilingues telles que les dictionnaires bilingues et les corpus parallèles peuvent être donc utilisées afin de tirer profit de certains indices dans une langue pour améliorer la qualité de REN dans une autre langue (Burkett et al. 2010, Green et al. 2012, Darwish 2013).

Il est presque impossible de construire des modèles ayant un bon taux de reconnaissance et traitant plusieurs langues. Plusieurs travaux tentent donc à rendre adaptables les systèmes de REN à plusieurs langues. L'approche utilisée pour les différentes langues reste en fait la même, mais les ressources changent en fonction de la langue traitée. Un détecteur de langue est généralement appliqué afin de sélectionner les ressources adéquates. Certaines ressources peuvent être partagées par plusieurs langues proches. Dans ce chapitre, nous présentons comment les systèmes orientés connaissances et les systèmes orientés données sont adaptés afin de couvrir plusieurs langues. La section 3.1 s'intéresse à l'adaptation des systèmes orientés connaissances. Nous étudions l'intérêt et le coût associé pour porter un système existant de REN vers une autre langue. Nous menons différents types d'expérimentations afin d'évaluer ce coût. La section 3.2 met l'accent sur l'adaptation des systèmes orientés données.

3.1 Adaptation des systèmes orientés connaissances

L'adaptation des systèmes orientés connaissances nécessite un traitement lourd incluant, notamment, la modification des règles contextuelles et le développement de nouvelles listes de mots déclencheurs et de noms propres. La rareté et le coût de construction de ces ressources représentent un problème majeur pour certaines langues (Poibeau 2003, Gamon et al. 1997).
Poibeau (2003) présente un système de REN qui vise à traiter douze langues différentes. Le traitement séquentiel est le même pour n'importe quelle langue. Ce traitement consiste à :

- analyser le texte en utilisant un système à base de règle classique basé sur le lexique et les règles contextuelles. Pour chaque langue, ces sources de connaissance sont développées séparément, ce qui nécessite un travail laborieux pour certaines langues dont les ressources existantes sont rares telles que la langue malgache.
- enrichir le lexique et les règles contextuelles en utilisant les techniques d'apprentissage automatique. Le but est d'identifier de nouvelles entités qui ne sont pas reconnues lors de la première phase.
- réviser l'annotation dans certains contextes. Par exemple, dans un texte anglais, les occurrences isolées du mot « Washington » sont considérées comme un nom de lieu. Quand le système trouve ensuite ce mot dans un autre contexte tel que « Mr. Washington », il révise les premiers occurrences en changeant leur annotation par la catégorie personne.

Ce système hybride est expérimenté sur l'anglais en utilisant le corpus MUC-6. Le résultat de l'évaluation montre 90 % de F-mesure. Afin de faciliter la création des ressources pour les nouvelles langues, Poibeau propose de développer des ressources partagées par plusieurs langues proches. Par exemple, une modification de dictionnaire de noms de personne pour l'anglais, en supprimant certaines entités ambiguës, permet d'obtenir un dictionnaire pour le français ou l'espagnol.

Dans ce cadre, nous étudions ici la possibilité de porter un système de REN pour le français et pour le texte bien formé vers l'anglais d'une façon simple et peu coûteuse (Hatmi 2012). Il s'agit ici du système Nemesis (Fourour 2002), présenté en détail dans la section 1.5.1. La méthode proposée consiste à adapter séparément et séquentiellement les deux principaux éléments constitutifs de Nemesis : les lexiques et les règles de réécriture :

- adaptation des lexiques : l'ensemble des lexiques pour l'anglais a été construit en traduisant tout simplement les lexiques existants pour le français. La traduction est faite automatiquement en utilisant un dictionnaire bilingue [1] sans aucune information contextuelle. Cela concerne principalement l'ensemble des lexiques de mots déclencheurs. Les lexiques des noms de personne et d'entreprise sont conservés. Cette tâche n'est pas coûteuse en termes de temps et ne demande pas une expertise linguistique (des outils en ligne comme

1. Catalogue ELRA-M0033

Google Translate peuvent être utilisés pour les langues pour lesquelles les dictionnaires électroniques ne sont pas disponibles). Lorsque le dictionnaire comporte plusieurs traductions pour un mot, l'ensemble des traductions sont conservées (les problèmes de polysémie ne sont pas traités à ce niveau). Une fois la phase de traduction terminée, une liste des mots outils en anglais a été utilisée pour écarter certaines entrées pouvant produire de bruit, par exemple, le mot « *even* » qui se présente comme un nom de personne dans le lexique français. En définitive, l'ensemble des lexiques pour l'anglais compte 83 305 entrées.

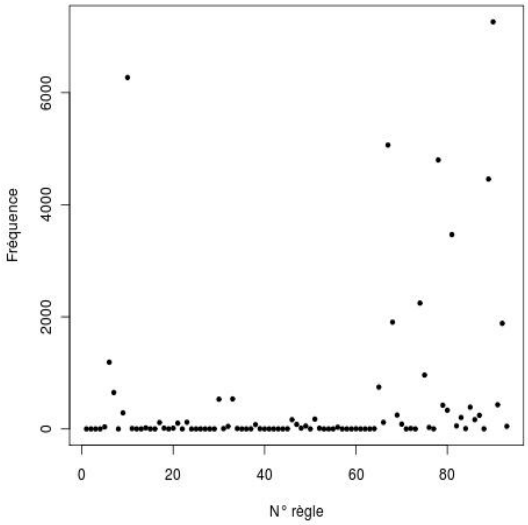

FIGURE 3.1 – *Nombre de déclenchements des règles pour le français*

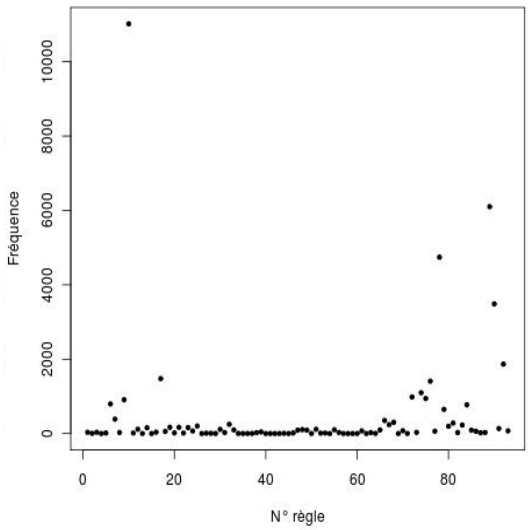

FIGURE 3.2 – *Nombre de déclenchements des règles pour l'anglais*

– adaptation des règles : le but ici est de sélectionner les règles à modifier pour la langue anglaise. Au début, la fréquence et la précision relatives de chacune des règles utilisées par Nemesis (93 règles) sont mesurées en utilisant le corpus Le Monde (2007) pour le français

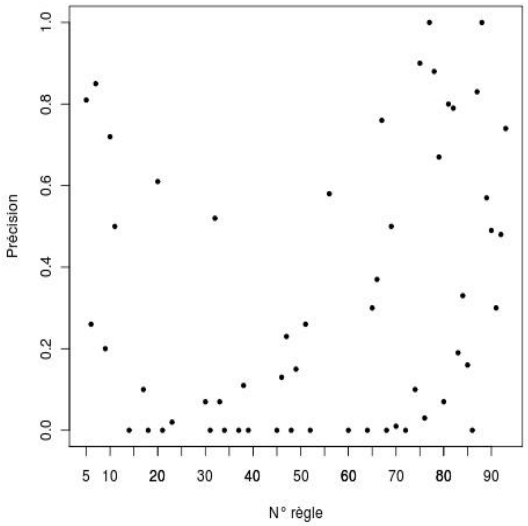

FIGURE 3.3 – *Précision des règles déclenchées pour l'anglais (corpus BBN)*

(1 010 000 mots) et le corpus de développement BBN [2] pour l'anglais (938 330 mots). La fréquence représente le nombre de fois où chaque règle est déclenchée pour reconnaître une entité nommée. La figure 3.1 présente les résultats obtenus sur le corpus Le Monde pour le français et la figure 3.2 présente les résultats obtenus sur le corpus de développement BBN pour l'anglais (après adaptation des lexiques). Les résultats montrent que la loi de Zipf est respectée pour les deux langues : un nombre limité de règles est à l'origine de la plupart des entités extraites, les autres règles sont rarement déclenchées. Par exemple, pour l'anglais, 11 règles couvrent 86 % des entités extraites. Aussi, de nombreuses règles ne sont pas déclenchées (38 pour l'anglais contre 17 pour le français). La précision mesure la quantité d'entités correctement reconnues parmi les réponses retournées. En effet, ce n'est pas parce qu'une règle est fréquente qu'elle est pour autant précise. La précision de chacune des règles déclenchées est calculée en utilisant le corpus AFP (Stern et Sagot 2010) [3] pour le français (38 831 mots) et le corpus de développement BBN pour l'anglais. La figure 3.3 présente les résultats obtenus pour l'anglais. Les résultats montrent que plusieurs règles déclenchées obtiennent une précision relativement faible (32 règles ont une précision inférieure à 50 %, ce qui représente environ 60 % des règles déclenchées).

En se basant sur les critères de fréquence et de précision, les règles sont ensuite classées en différentes catégories, par exemple : règles fréquentes [4] ayant une bonne précision [5] dans les deux langues (7 règles), règles fréquentes pour l'anglais avec une faible précision [6] (3 règles), règles déclenchées seulement pour le français (19 règles), règles non déclenchées dans les deux langues (15 règles), etc. Ces dif-

2. BBN Pronoun Coreference and Entity Type Corpus : catalogue LDC n° LDC2005T33
3. Ce corpus est disponible librement dans le cadre de la distribution de SXPipe
4. Fréquence > 1 000
5. Précision > 70 %
6. Précision < 50 %

férentes catégories permettent de déterminer quelles sont les règles à modifier pour la langue anglaise. Celles n'appartenant pas aux catégories de bonne précision pour l'anglais sont sélectionnées. Certaines règles déclenchées seulement pour le français sont éliminées car elles sont très spécifiques à cette langue, par exemple :

```
[ $Clé-organisation $Article-min $Item $Article-min
$Forme-capitalisée+ ] → ORGANISATION
```

et un exemple d'application :

« <ORGANISATION> Centre de recherche de Solaize </ORGANISATION> »

Ensuite, les règles sélectionnées sont modifiées de manière à respecter la syntaxe de la langue anglaise (29 règles), comme la règle suivante pour le français :

```
$Fonction $Adjectif-de-nationalité [ $Forme-capita-
lisée+ ] → PERSON
```

et le résultat de son application :

« Le président français <PERSON> Jacques Chirac </PERSON> »

et la règle pour l'anglais après adaptation :

```
$Adjectif-de-nationalité $Fonction [ $Forme-capita-
lisée+ ] → PERSON
```

et le résultat de son application :

« The French President <PERSON> Jacques Chirac </PERSON> »

Afin d'évaluer les performances de cette approche, Nemesis a été appliqué, dans un premier temps, au corpus de test BBN (235 274 mots) et au corpus AFP (38 831 mots). Aucune adaptation n'a été réalisée. Les résultats de l'évaluation montrent que la reconnaissance des noms de personne demeure satisfaisante (perte d'environ 5 points de F-mesure : 77,33 % pour le français contre 71,82 % pour l'anglais). Cela s'explique par le fait que les lexiques de Nemesis contiennent des prénoms anglais et qu'il y a des règles de réécriture communes. La reconnaissance se voit fortement dégradée pour les autres catégories (noms de lieu : 88,82 % pour le français contre 37,46 % pour l'anglais, noms d'organisation : 54,24 % pour le français contre 1,50 % pour l'anglais).

L'apport de l'adaptation des lexiques et des règles est ensuite calculé. La table 3.1 affiche les gains obtenus pour chaque adaptation. Ces résultats sont comparés à ceux obtenus avec un système natif (Stanford NER)[7] (Finkel et al. 2005).

L'adaptation des lexiques a permis un apport significatif concernant la reconnaissance des noms de lieu (environ 37 points de F-mesure). La reconnaissance des noms d'organisation et d'entreprise se voit améliorée mais elle reste toutefois faible. En effet, ce problème semble lié à une couverture insuffisante des lexiques de Nemesis et à une ambiguité liée à la

7. Ce système est disponible gratuitement à : http://nlp.stanford.edu/software/CRF-NER.shtml

	Nemesis Anglais (adaptation des lexiques)	**Nemesis Anglais (adaptation des lexiques et des règles)**	**Stanford NER**
	Corpus BBN (anglais)	**Corpus BBN (anglais)**	**Corpus BBN (anglais)**
	F-mesure (P/R)	**F-mesure (P/R)**	**F-mesure (P/R)**
Personne	77,30 (74,48/80,34)	79,51 (81,08/78,01)	90,90 (88,07/93,90)
Lieu	74,38 (72,50/76,35)	79,17 (78,59/79,76)	90,80 (87,6/94,20)
Organisation	21,02 (24,20/18,58)	38,15 (41,90/35,02)	84,60 (89,20/80,04)
Entreprise	27,52 (50,72/18,88)	30,15 (68,34/19,34)	

TABLE 3.1 – *Performances de Nemesis pour le français et l'anglais en termes F-mesure, Précision (P) et Rappel (R)*

catégorisation des organisations et des entreprises (entreprise catégorisée en tant qu'organisation et vice-versa).

L'adaptation des règles montre un gain relativement bon pour la catégorie organisation (environ 17 points de F-mesure) et une légère amélioration concernant les autres catégories. Les résultats globaux sont bien en dessous de ceux obtenus avec Stanford NER (Finkel et al. 2005), surtout pour la catégorie organisation. Stanford NER est un système orienté données développé pour l'anglais. Pour ce dernier, les noms d'entreprises sont catégorisés comme étant organisation.

Cette approche peu coûteuse pour porter Nemesis vers l'anglais montre des résultats satisfaisants pour la reconnaissance des noms de personne et de lieu. Elle montre ses limites pour la reconnaissance des noms d'organisation et d'entreprise. Nous nous intéressons dans la section suivante à l'adaptation des systèmes orientés données.

3.2 Adaptation des systèmes orientés données

Les techniques d'apprentissage sont indépendantes de la langue. Seuls les corpus d'apprentissage annotés changent afin de créer un modèle pour une langue donnée. Huang (2006) présente un système multilingue de REN à base de HMM. Ce système est expérimenté sur l'anglais, l'arabe et le chinois. Un corpus annoté est utilisé afin de créer un modèle pour chaque langue. Du fait de l'indisponibilité d'un corpus d'apprentissage suffisant pour la langue chinoise, un système de REN existant a été utilisé afin d'annoter automatiquement un corpus de grand taille. Ce corpus imparfaitement annoté est utilisé ensuite afin de créer un modèle pour le chinois.

Le problème majeur des systèmes orientés données est donc la disponibilité des corpus d'apprentissage. Ces derniers ne sont pas faciles à constituer et ne sont pas disponibles pour toutes les langues. Plusieurs travaux visent à automatiser le processus d'annotation. Nous distinguons ici deux

approches principales. La première se base sur l'utilisation des ressources encyclopédiques multilingues (Kim et al. 2012, Richman et Schone 2008) et la deuxième repose sur l'exploitation des corpus comparables multilingues (Klementiev et Roth 2006) et des corpus parallèles (Fu et al. 2011, Che et al. 2013, Wang et al. 2013). Nous détaillons ici quelques échantillons de ces systèmes.

3.2.1 Utilisation de ressources encyclopédiques

Richman et Schone (2008) utilisent l'encyclopédie multilingue Wikipédia pour créer automatiquement un corpus d'apprentissage annoté dans n'importe quelle langue utilisée sur Wikipédia. La méthode proposée ne demande pas de connaissances sur la langue sauf pour l'anglais qui est considéré comme une langue référentielle. L'identification des entités nommées potentielles se base sur trois types de wikiliens :

- les liens internes qui mènent d'une page à une autre dans la même langue.
- les liens interlangues qui mènent d'une page dans une langue à une page dans une autre langue.
- les liens thématiques qui mènent à une page de catégorie. En effet, chaque article Wikipédia est classé, selon son sujet, dans une ou plusieurs catégories. Les catégories sont elles-mêmes classées dans d'autres catégories, thématiquement plus larges.

Pour l'anglais, les catégories Wikipédia auxquelles la page d'une entité potentielle appartient sont extraites en utilisant les liens thématiques. Ces catégories sont ensuite comparées à de simples listes de mots-clés créées manuellement pour chaque catégorie d'entités nommées. Par exemple, « *People by* », « *People in* » et « *Human names* » pour la classe « *Person* ». Si un mot-clé apparaît dans un nom de catégorie Wikipédia, alors l'entité potentielle est assignée à la catégorie d'entités nommées à laquelle appartient le mot clé. Sinon, le système passe aux catégories Wikipédia parentes. Par exemple, pour classifier l'entité « *Jacqueline Bhabha* », le système commence par extraire les catégories Wikipédia suivantes : « *British lawyers* », « *Jewish American writers* » et « *Indian Jews* ». Aucun mot-clé n'est trouvé, alors il extrait les catégories parentes de second ordre : « *Lawyers by nationality* », « *British legal professionals* », « *American writers by ethnicity* », « *Indian people by religion* », etc. Il trouve que « *people by* » apparaît dans l'un de ces catégories. Donc, l'entité « *Jacqueline Bhabha* » est assignée à la catégorie « *Person* ».

Pour les autres langues, le système commence par chercher l'équivalent anglais de la page de l'entité potentielle en utilisant les liens interlangues. Si l'équivalent anglais existe, il le catégorise comme expliquer ci-dessus et considère que l'entité potentielle dans l'autre langue a la même catégorie. Sinon, le système se réfère aux catégories Wikipédia. En associant ces catégories à leurs équivalents en anglais, il détermine la catégorie de l'entité nommée potentielle. Par exemple, « *Breton town of Erquy* » est une page en français qui n'a pas un équivalent en anglais. Le système commence par

déterminer ses catégories Wikipédia en français : Commune des Côtes-d'Armor, Ville portuaire de France, Port de plaisance et Station balnéaire. Il les associe ensuite à leurs équivalent en anglais : « *Communes of Côtes-d'Armor* », « *Unknown* », « *Marinas* », « *Seaside resorts* ». Il trouve que la première catégorie représente une sous-catégorie de « *Cities* », « *Towns* » and « *Villages in France* ». Donc, l'entité potentielle est associée à la catégorie « GPE ».

Cette méthode pour la création d'un corpus annoté a été évaluée sur plusieurs langues. Les résultats obtenus montrent une F-mesure d'environ 85 % pour le français, l'espagnole et le polonais et une F-mesure d'environ 80 % pour ukrainien, la russe et le portugais.

3.2.2 Utilisation de corpus comparables multilingues et de corpus parallèles

Klementiev et Roth (2006) proposent l'exploitation des corpus multilingues comparables qui sont faciles à construire (par exemple, deux ensembles d'articles journalistiques sur un même thème mais dans deux langues différentes). L'approche utilisée ici repose simultanément sur deux hypothèses :

- la première concerne la synchronicité des entités nommés : étant donnés deux corpus multilingues comparables ayant des dates de publication proches, les entités nommées dans un corpus sont susceptibles d'apparaitre avec leurs équivalents dans l'autre corpus. Un score de similitude des distributions au cours du temps est calculé pour chaque paire constituant l'entité et son équivalent potentiel. Ici, l'écueil est que les entités appartenant au même corpus et ayant des distributions similaires peuvent être confondues (par exemple, Taliban et Afghanistan).
- la seconde est que les entités nommées sont généralement composées des mots qui sont phonétiquement translittérées à travers les langues. Ici, l'écueil est la difficulté d'avoir un bon modèle de translittération.

Une étape préliminaire consiste à entraîner un modèle linéaire de translittération en utilisant des entités composées d'une seule forme : pour une entité donnée, le modèle génère plusieurs translittérations potentielles dans une autre langue. La meilleur translittération est celle ayant le score de similitude des distributions au cours du temps le plus élevé. La paire résultante (entité, translittération) est utilisée afin d'entraîner itérativement le modèle. Cette méthode montre une précision de 66 % sur un corpus comparable anglais-russe composé de 2 022 documents (727 entités nommées prises en compte).

Che et al. (2013) proposent l'utilisation des corpus parallèles afin d'améliorer les performances de la REN pour deux langues différentes. Les corpus parallèles sont composés de paires de phrases chinois-anglais tel que, pour chaque paire, une phrase représente une traduction de l'autre. L'idée sous-jacente derrière l'exploitation des corpus parallèles est de bénéficier de certains indices saillants pour la REN, qui sont présents

dans une langue, pour aider à l'annotation de certaines entités ambiguës dans l'autre langue. Par exemple, le mot chinois « 本 » signifie « *this* » en anglais. Ce mot peut représenter aussi une translittération du prénom « Ben ». Pour résoudre cette ambiguïté, il est possible d'utiliser des indices spécifiques à l'anglais tels que la majuscule pour annoter correctement ce mot en anglais et projeter ensuite l'annotation obtenue aux segments de la langue chinoise. Deux types de contraintes bilingues sont définies : dures et souples. Les premières consistent à ce que les mots composant une paire d'alignement mot à mot doivent avoir les mêmes étiquettes d'entité nommée. Les secondent concernent certains cas particuliers tels que dans le cas d'erreurs d'alignement automatique de mots. Les résultats expérimentaux, établis sur un corpus parallèle chinois-anglais d'environ 8 249 paires de phrases alignées, montrent une amélioration d'environ 5 % en termes de F-mesure pour la langue chinoise.

Conclusion

La difficulté de la REN se complexifie dès lorsqu'elle s'applique dans un contexte multilingue. Certains indices spécifiques à une langue donnée ne sont pas valides pour repérer et classifier les entités nommées dans un autre texte ayant une langue différente. L'utilisation des ressources multilingues comme les dictionnaires bilingues, les corpus comparables et parallèles et les ressources encyclopédiques permet l'exploitation de ces indices afin d'améliorer la qualité de REN pour certaines langues dont elles en sont privées. Cela permet aussi la construction des certaines ressources importantes pour la REN telles que les corpus d'apprentissage et les dictionnaires de noms propres. Le problème qui se pose ici concerne l'acquisition de certaines ressources multilingues. Par exemple, les corpus parallèles alignés qui sont difficiles à trouver ou à constituer.

La portabilité entre les langues des systèmes de REN est coûteuse en termes de temps et de connaissances linguistiques requises. L'adaptation des systèmes orientés connaissances souffre du coût de développement de nouveaux lexiques et de la mise à jour des règles contextuelles. D'un autre côté, l'adaptation des systèmes statistiques se heurte au problème du coût de préparation d'un nouveau corpus d'apprentissage. Dans ce chapitre, nous avons montré comment les systèmes de REN peuvent être adaptés entre les langues en exploitant les ressources multilingues. Nous avons aussi étudié l'intérêt et le coût associé pour porter un système existant de REN pour du texte bien formé vers une autre langue. Nous avons proposé une méthode peu coûteuse pour porter un système orienté connaissances dédié au français vers l'anglais. Les expérimentations réalisées montrent des résultats satisfaisants pour la reconnaissance des noms de personne et de lieu et des résultats qui restent insuffisants pour les noms d'organisation et d'entreprise. Un traitement plus approfondi concernant la langue et la modalité traitées est donc nécessaire. Dans le chapitre suivant, nous nous intéressons à la langue française et à la modalité parole. Nous étudions les enjeux et les limites concernant la REN dans les transcriptions automatiques de la parole en suivant une taxonomie fine.

Deuxième partie

Contributions

4 Reconnaissance d'entités nommées dans les transcriptions de la parole

Sommaire

Avec l'essor des documents sonores et audiovisuels (enregistrements radiophoniques, conversations téléphoniques, émissions télévisées, etc.), la reconnaissance automatique de la parole (RAP) devient une technique importante pour indexer comme pour rechercher de l'information dans ce flux. Dans ce contexte, le traitement des entités nommées soulève des difficultés particulières, que se soit du point de vue de la transcription ou de la reconnaissance. En effet, la transformation de la parole sous la forme d'un texte exploitable se heurte aux problèmes des erreurs de transcription et des erreurs résultant des mots hors-vocabulaire. Les transcriptions produites sont souvent plus au moins bruitées. De plus, comme décrit dans la section 2.2, certains indices importants pour la REN ne sont plus exploitables lorsqu'elle s'applique sur des transcriptions de la parole. Cela est dû notamment à certains phénomènes spécifiques à l'oral telles que les hésitations, les faux départs, les répétitions. À cela s'ajoute

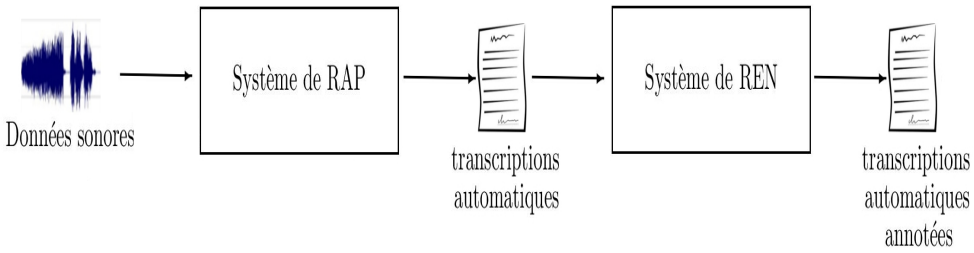

FIGURE 4.1 – *REN dans les transcriptions de la parole qui vient en aval de la tâche de reconnaissance du signal*

le manque de structuration (manque de ponctuation) et de majuscule du résultat de la transcription automatique.

Dans ce chapitre, nous étudions les spécificités de la REN en sortie du système de RAP (Figure 4.1). Notre objectif est de mesurer l'impact des différents phénomènes spécifiques à la parole sur la qualité de REN :

- impact de l'absence de la majuscule ;
- impact de la présence de disfluences ;
- impact de la présence d'erreurs de transcription.

Nous adoptons une taxonomie d'entités nommées hiérarchique et compositionnelle. Nous évaluons aussi l'impact de l'adoption de cette taxonomie fine.

Ce chapitre est organisé comme suit : la section 4.1 présente les spécificités des corpus, ainsi que les mesures de performance utilisés dans le cadre de ce travail. La section 4.2 introduit quelques notions concernant les modèles probabilistes, en mettant l'accent sur les champs aléatoires conditionnels. La section 4.3 décrit l'approche que nous mettons en œuvre. La section 4.4 présente une évaluation des performances de notre approche à travers différentes expérimentations. Finalement, la section 4.5 vient discuter les résultats obtenus.

4.1 Description des corpus et mesures de performance

Ce travail se situe dans le cadre de la campagne d'évaluation ETAPE (Évaluations en Traitement Automatique de la Parole) (Gravier et al. 2012). Cette campagne avait pour objectif de mesurer les performances des technologies vocales appliquées à l'analyse des flux télévisés en langue française [1]. L'évaluation des performances des systèmes concernait trois tâches principales : la segmentation, la transcription et l'extraction d'informations. Nous nous sommes intéressés à la tâche d'extraction d'informations qui porte sur la reconnaissance d'entités nommées. L'annotation manuelle des corpus mis à disposition a été conduite suivant le guide d'annotation Quæro (Rosset et al. 2011b). Nous commençons donc par présenter les spécificités de ce schéma d'annotation. Nous décrivons ensuite les corpus et les mesures de performance utilisés dans ce travail.

4.1.1 Schéma d'annotation Quæro

Le guide d'annotation Quæro propose une nouvelle définition pour des entités nommées étendues. Les entités sont organisées d'une manière hiérarchique. Elles sont structurées autour de huit catégories principales et de 32 sous-catégories. La table 4.1 montre la hiérarchie des entités. Par rapport au guide d'annotation ESTER 2 (Galliano et al. 2009) [2], nous trouvons les mêmes sept catégories principales mais nous relevons un changement concernant l'organisation des sous-catégories. Nous constatons un nombre plus grand de sous-catégories pour les noms de lieu et de production humaine. Par contre, nous relevons l'absence de sous-catégories pour le type « *amount* ». Les noms de personne, d'organisation et de fonction ne possèdent que deux sous-catégories chacun. Nous relevons aussi l'utilisation dans le guide d'annotation Quæro de la catégorie « *event* » pour l'annotation des événements, comme pour la hiérarchie d'entités nommées proposée dans Sekine et al. (2002). Cette catégorie n'est pas considérée dans la campagne ETAPE ni dans notre travail vu qu'elle a été peu annotée dans le corpus d'entraînement.

La couverture de certaines catégories d'entités est étendue afin d'inclure certains syntagmes construits autour de noms communs. Par exemple, « la civilisation précolombienne », « les Mayas », « le socialiste », etc.

En plus du critère hiérarchique, les entités nommées sont organisées d'une manière compositionnelle. La compositionnalité concerne la manière dont les entités nommées sont formées. Les éléments constitutifs de l'entité sont étiquetés avec un composant ou une autre entité imbriquée, sauf éventuellement les prépositions. Cela permet d'obtenir une annotation plus fine à l'intérieur même des entités étiquetées avec les catégories.

1. www.afcp-parole.org/etape.html
2. www.afcp-parole.org/camp_eval_systemes_transcription/docs/Conventions_EN_ESTER2_v01.pdf

Catégories	Sous-catégories
Personne (pers)	- **Individu** (pers.ind) - **Groupe de personnes** (pers.coll) : présence d'un nom propre (les Beatles) ou présence d'un gentilé (victimes françaises)
Fonction (fonc)	- **Individu** (fonc.ind) : métier (pompier), fonction (Miss Italie), - **Groupe de personnes** (fonc.coll) : métier (journalistes), fonction (chefs d'État), rôle social (intellectuels)
Organistion (org)	- **Service** (org.ent) : commercialise des produits ou rend des services non uniquement administratifs (société Peugeot) - **Administration** (org.adm) : joue un rôle principalement administratif (Mairie de Paris)
Lieu (loc)	- **Lieu administratif ville** (loc.adm.town) : ville, village, etc. - **Lieu administratif région** (loc.adm.reg) : région, comté, etc. - **Lieu administratif national** (loc.adm.nat) : pays - **Lieu administratif supranational** (loc.adm.sup) : continent, etc. - **Lieu physique terrestre** (loc.phys.geo) : espaces géographiques naturels tels que les montagnes, les glaciers, etc. - **Lieu physique aquatique** (loc.phys.hydro) : noms d'étendue d'eau tels que les rivières, les océans, etc. - **Lieu physique astronomique** (loc.phys.astro) : les planètes, etc. - **Oronyme** (loc.oro) : les rues, les places, les routes, etc. - **Bâtiment** (loc.fac) : les gares, les musées, les stades, etc. - **Adresse physique** (loc.add.phys) : un point dans une voie. - **Adresse électronique** (loc.add.elec) : coordonnées électroniques tels que les numéros de téléphone, les e-mails, les URLs, etc.
Production humaine (prod)	- **Marque d'objet** (prod.object) : objets manufacturés - **Œuvre artistique** (prod.art) : les peintures, les sculptures, etc. - **Production médiatique** (prod.media) : tout ce qui est diffusé dans la presse, à la radio ou à la télévision, etc. - **Produit financier** (prod.fin) : les produits bancaires (PEP), les indices boursiers (CAC 40), les monnaies (l'Euro), etc. - **Logiciel** (prod.soft) - **Prix divers** (prod.award) - **Ligne de transport** (prod.serv) - **Doctrine** (prod.doctr) : le socialisme, le communisme, etc. - **Rule** (prod.rule) : loi, décret, accord, etc. - **Productions autres** (prod.other)
Quantité (amount)	
Point dans le temps (time)	- **Date absolue** (time.date.abs) - **Date relative** (time.date.rel) : demain, jour de Noël, etc. - **Heure absolue** (time.hour.abs) - **Heure relative** (time.hour.rel) : ne peuvent être déterminées qu'avec un raisonnement s'ancrant sur la connaissance de l'heure de l'énonciation (dans deux heures).
Événement (event)	

TABLE 4.1 – *Catégories et sous-catégories d'entités nommées de la campagne ETAPE.*

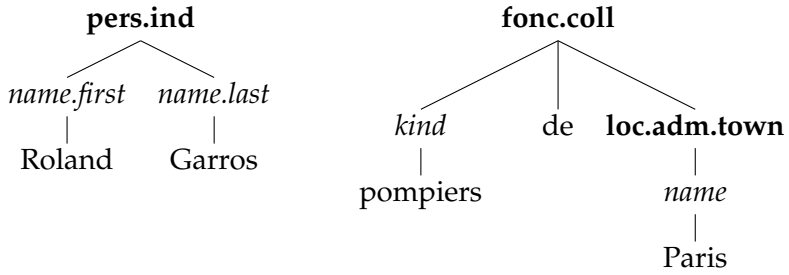

FIGURE 4.2 – *Exemples de découpage d'entités en composants et en entités imbriquées*

La figure 4.2 montre quelques exemples de découpage d'entités en composants et en entités imbriquées. Dans le premier exemple, nous distinguons les deux composants prénom (*name.first*) et nom (*name.last*) à l'intérieur de l'entité personne (*pers.ind*). Dans le deuxième exemple, nous distinguons le composant hyperonyme (*kind*), la préposition « de » et l'entité de type lieu (*loc.adm.town*) à l'intérieur de l'entité fonction (*fonc.coll*). Nous distinguons aussi le composant nom (*name*) à l'intérieur de l'entité lieu (*loc.adm.town*). Certains composants sont transverses, c'est-à-dire qu'ils peuvent être communs à différentes entités, d'autres étant spécifiques à une catégorie. La table 4.2 présente ces différents composants.

Types	**Composants**
Transverses	name, name.nickname, kind, extractor, demonym.nickname, qualifier, val, unit, range-mark
Personne (pers)	name.last, name.first, name.middle
Adresse (loc.phys.add)	address-number, po-box, zip-code
Quantité (amount)	object
Point dans le temps (time)	day, week, month, year, century, millenium, reference-era, time-modifier

TABLE 4.2 – *Composants des entités nommées de la campagne ETAPE.*

Nous recensons ici quelques directives d'annotation comme elles sont expliquées dans le guide d'annotation :

- métonymie : dans une métonymie, une catégorie d'entité est employée pour désigner une autre catégorie d'entité. On annote la catégorie à laquelle appartient intrinsèquement l'entité mentionnée. On surannote celle-ci avec la catégorie désignant l'expression dans le contexte (« l' <org.adm> <loc.fac> <name> Élysée </name> </loc.fac> </org.adm> »).
- imbrication : les seules imbrications obligatoires d'entités sont : la métonymie (« la <org.adm> <loc.adm.nat> <name> France </name> </loc.adm.nat> </org.adm> demande l'arrêt des hostilités ») et les cas comme « <func.ind> <kind> maire </kind> de <loc.adm.town> <name> Paris </name> </loc.adm.town> </func.ind> » ou « <func.ind> <kind> ministre </kind> des <org.adm> <name> affaires étrangères </name> </org.adm> </func.ind> », etc.
- disfluences :
    1. les phénomènes typiques de l'oral (hésitations, mots cou-

pés etc.) doivent être intégrés dans l'entité s'ils coupent strictement l'entité et n'apportent pas de correction à l'élément (« <prod.object> <org.ent> <name> Peugeot </name> </org.ent> euh <name> 107 </name> </prod.object> »).

2. s'il s'agit d'une reformulation, d'une correction, il y a autant d'entités que de reformulations (« <prod.object> <org.ent> <name> Renault </name> </org.ent> </prod.object> euh <prod.object> <org.ent> <name> Peugeot </name> </org.ent> <name> 107 </name> </prod.object> »).

– coordination : lorsque plusieurs entités se suivent avec mise en facteur d'une partie commune, on annote séparément chaque entité, bien qu'elle soit incomplète (« <loc.phys.geo> <kind> vallées </kind> de la <name> Lorraine </name> </loc.phys.geo> », « <loc.phys.geo> de l' <name> Alsace </name> </loc.phys.geo> »).
– élaboration : lorsqu'on a à la fois un acronyme ou un sigle et une définition, explication, expansion, élaboration de ce nom, on a affaire à deux entités distinctes (« <org.ent> <name> Agipi </name> </org.ent> <org.ent> <name> association d'assurés pour la prévoyance », « la dépendance et l'épargne-retraite </name> </org.ent> »).

4.1.2 Présentation des corpus

Le corpus ETAPE que nous utilisons dans ce travail est composé de transcriptions manuelles d'émissions télévisées et radiophoniques. La taille de ce corpus représente environ 42,5 heures d'enregistrements audio provenant de différentes stations de radio et de télévision françaises : BFM TV, La Chaîne Parlementaire (LCP), TV8 et France Inter. Les émissions enregistrées consistent en des émissions d'information, des débats d'actualité, des discussions, des reportages et des questions au gouvernement (assemblée nationale). Les transcriptions obtenues sont enrichies par une annotation des entités nommées selon le guide d'annotation Quæro.

Le corpus ETAPE est divisé en trois parties distinctes : une première partie d'entraînement qui sert à l'apprentissage du système (60 %), une autre partie de développement pour l'ajustement des règles et des paramètres du système (20 %) et une dernière partie d'évaluation qui est utilisée pour évaluer les performances du système (20 %). Les caractéristiques de ces différentes parties sont présentées dans la table 4.3.

La table 4.4 présente les caractéristiques du corpus ETAPE en termes de nombre de mots, de nombre de mots composants les entités nommées et de nombre d'entités nommées. La figure 4.3 montre le pourcentage de chaque catégorie d'entité nommée par rapport au nombre total d'entités dans les différentes parties du corpus. Nous remarquons que la répartition des entités nommées est homogène dans les trois corpus. La table 4.5 présente le nombre des entités nommées par catégorie, ainsi que le nombre des composants.

Genre	Train	Dev	Test	Source
Émissions d'information télévisées	7h30	1h35	1h35	BFM Story, Top Questions (LCP)
Débats d'actualité télévisés	10h30	2h40	2h40	Pile et Face, Ça vous regarde, Entre les lignes (LCP)
Émissions d'amusement télévisés	-	1h05	1h05	La place du village (TV8)
Émissions de radio	7h50	3h00	3h00	Un temps de Pauchon, Service Public, Le masque et la plume, Comme on nous parle, Le fou du roi
Total	25h30	8h20	8h20	42h10

TABLE 4.3 – *Caractéristiques du corpus ETAPE (Gravier et al. 2012) : Train (corpus d'entraînement), Dev (corpus de développement), Test (corpus de test).*

Corpus	Nombre de mots	Nombre de mots composants les EN	Nombre d'EN
ETAPE-train	312 368	35 621 (11,40 %)	19 105
ETAPE-dev	100 217	11 058 (11 %)	5 609
ETAPE-test	100 585	9 587 (9,50 %)	5 367

TABLE 4.4 – *Caractéristiques du corpus ETAPE : ETAPE-train (corpus d'entraînement), ETAPE-dev (corpus de développement), ETAPE-test (corpus de test).*

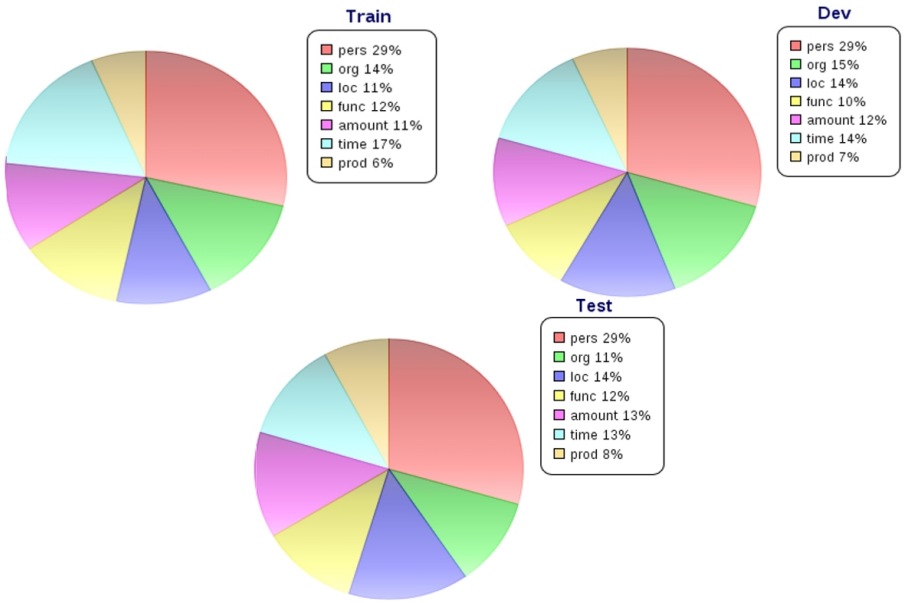

FIGURE 4.3 – *Pourcentage de chaque catégorie d'entité nommée par rapport au nombre total d'entités dans les différentes parties du corpus ETAPE.*

Catégorie	Train	Dev	Test	Composant	Train	Dev	Test
amount	2 272	683	705	kind	4 981	1 428	1 163
pers.ind	4 541	1 430	1 398	extractor	1 410	15	4
pers.coll	1 181	309	177	qualifier	1 410	380	250
time.date.abs	698	214	192	val	2 572	792	808
time.date.rel	1 317	407	348	unit	1 302	369	463
time.hour.abs	127	51	46	object	924	296	225
time.hour.rel	341	146	84	range-mark	156	40	71
loc.oro	24	5	2	day	111	28	36
loc.fac	357	133	81	week	141	44	39
loc.add.phys	5	1	4	month	187	59	74
loc.add.elec	42	21	18	year	454	128	95
loc.adm.town	655	245	279	century	13	5	3
loc.adm.reg	149	95	47	reference-era	23	10	2
loc.adm.nat	766	240	276	time-modifier	1 389	467	337
loc.adm.sup	136	26	33	award-cat	7	2	2
loc.phys.geo	25	53	30	demonym	0	0	206
loc.phys.hydro	16	3	5	name	6 387	2 110	1 593
loc.phys.astro	1	3	0	name.last	3 346	992	928
prod.object	47	41	58	name.first	3 213	1 024	1 032
prod.art	288	81	87	name.nickname	58	23	71
prod.media	622	217	164	title	375	89	75
prod.fin	106	19	84				
prod.soft	18	0	0				
prod.award	33	10	13				
prod.serv	18	3	0				
prod.doctr	12	6	3				
prod.rule	113	17	5				
prod.other	8	4	7				
prod.unk	0	0	2				
func.ind	1 411	358	383				
func.coll	1 019	219	243				
org.ent	1 578	345	307				
org.adm	1 179	224	286				
Total	19 105	5 609	5367	Total	28 459	8 301	7 477

TABLE 4.5 – *Nombre des entités nommées par catégorie et le nombre des composants dans le corpus ETAPE*

4.1.3 Mesures de performance

La qualité de la REN est évaluée en termes de F-mesure et de coût d'appariement (SER) (Makhoul et al. 1999, Galibert et al. 2011). La base de calcul pour ces deux mesures est le nombre d'entité nommées. L'équation correspondante à la F-mesure est donnée dans la section 1.4. Pour le SER, nous utilisons la variante définie pour la campagne d'évaluation ETAPE (Galibert et al. 2011). Les types d'erreurs considérés sont : les suppressions (D), les insertions (I), les erreurs de type (T), les erreurs de frontière (F) et les erreurs de type et de frontière (TF). Le SER est alors donné par l'équation 4.1 :

$$SER = \frac{D + I + 0,5 \times T + 0,5 \times F + TF}{Nombre\ d'entités\ étiquetées\ de\ la\ référence} \tag{4.1}$$

L'annotation de référence est effectuée sur les transcriptions manuelles. Les annotations obtenues sont ensuite projetées automatiquement sur les différentes sorties des systèmes de transcription afin d'annoter les références automatiques. Cette projection fournit une annotation de référence des entités nommées correspondant à ce qui a été dit, et non à ce qui est reconnu (Rosset et al. 2011a).

4.2 REN comme étant un problème d'étiquetage de séquences

Différents types de modèles probabilistes sont employés pour la REN. Cette tâche est formalisée comme un problème d'étiquetage de séquences. L'objectif revient à trouver la meilleur séquence d'étiquettes au sens d'une séquence d'observations. Cela peut être modélisé comme suit : étant donnée une séquence d'observations (séquence de mots dans notre cas) $X = \{x_1, x_2..., x_n\}$, il faut trouver la séquence d'étiquettes $E = \{e_1, e_2..., e_n\}$ qui maximise la probabilité conditionnelle :

$$\widehat{e} = arg\ \max_{e}\ \mathbb{P}(E|X) \tag{4.2}$$

avec $\forall j, e_j \in E = \{e_1, e_j..., e_n\}$, E est un ensemble fini des étiquettes.

Il est coutume de distinguer deux familles de modèles probabilistes pour résoudre ce problème, à savoir les modèles génératifs et les modèles discriminants.

Les modèles génératifs, tels que les Modèles de Markov Cachés (HMM), s'intéressent à maximiser la probabilité jointe de la paire (X, E). En appliquant la règle de Bayes, cette probabilité se décompose en deux contributions à estimer, l'une correspond à la probabilité conditionnelle de génération des observations $\mathbb{P}(X|E)$, et l'autre à la probabilité de la séquence d'étiquettes $\mathbb{P}(E)$. Des hypothèses sur l'indépendance des observations sont généralement imposées en raison des problèmes de complexité (explosion combinatoire).

Les modèles discriminants considèrent la probabilité conditionnelle $\mathbb{P}(E|X)$ plutôt que la probabilité jointe $\mathbb{P}(X, E)$. Cela permet de relaxer les hypothèses faites sur l'indépendance des observations et de prendre une décision globale sur l'ensemble des observations afin de prédire une étiquette. En outre, les modèles discriminants ne nécessitent pas l'estimation de la probabilité de génération des observations.

Une description détaillée des modèles génératifs et discriminants peut être trouvée dans Zidouni (2010).

Plusieurs études ont montré que les modèles discriminants ont un certain avantage sur les modèles génératifs (Lafferty et al. 2001, Sha et Pereira 2003, Raymond et Riccardi 2007). Les champs aléatoires conditionnels (CRF pour *Conditional Random Fields*) se sont imposés comme l'un des modèles discriminants les plus performants pour le problème d'étiquetage de séquences. Les CRF sont des modèles graphiques non dirigés. Ils permettent de manipuler un grand nombre de descripteurs et de remédier au problème de biais de label rencontré avec les modèles de markov à entropie maximale (MEMM pour *Maximum Entropy Markov Model*) (Lafferty et al. 2001). Un modèle CRF permet de calculer la probabilité conditionnelle d'une séquence d'étiquettes $E = \{e_1, e_2..., e_n\}$ étant donné une séquence de mots $X = \{x_1, x_2..., x_n\}$. Cette probabilité est définie comme suit :

$$\mathbb{P}(E|X) = \frac{1}{Z(X)}\ exp\ (\sum_{i=1}^{n}\sum_{k=1}^{K}\lambda_k f_k(e_i,\ x_i,\ v(x_i),\ X)) \tag{4.3}$$

- $f_k(e_i,\ x_i,\ v(x_i),\ X)$ est un ensemble de fonctions caractéristiques sur lesquelles repose le modèle pour la prédiction des étiquettes. Cela représente généralement des fonctions binaires qui retournent 1 si un phénomène à la position courante est observé, 0 sinon. Ces fonctions caractéristiques prennent en paramètre l'étiquette candidate e_i, le mot courant x_i, son contexte $v(x_i)$, ainsi que l'ensemble de l'observation X. Elles sont définies par l'utilisateur, ce qui nécessite une certaine expertise du domaine. Par exemple, la présence de la majuscule pour la REN.
- λ_k sont des paramètres de pondération associés à chaque fonction caractéristique f_k. Ces poids sont fixés à partir de données d'apprentissage afin d'accorder plus d'importance aux fonctions caractéristiques les plus discriminantes pour la prédiction des étiquettes. Un poids négatif peut être associé à une fonction caractéristique afin d'indiquer qu'un phénomène ne doit pas se produire.
- $Z(x)$ est un coefficient de normalisation calculé sur toutes les séquences candidates. Il est défini comme suit :

$$Z(X) = \sum_{l \in L(X)}\ exp\ (\sum_{i=1}^{l}\sum_{k=1}^{K}\lambda_k f_k(e_i,\ x_i,\ v(x_i),\ X)) \tag{4.4}$$

  $L(X)$ est l'ensemble de toutes les séquences candidates X.

La phase d'apprentissage consiste à estimer les paramètres de pondération $\Lambda = \{\lambda_k\}$ à partir des données annotées. Ces paramètres doivent

maximiser la vraisemblance du modèle par rapport aux données d'apprentissage. Une fois un modèle d'annotation appris, l'application des CRF à de nouvelles données revient alors à trouver la séquence d'étiquettes la plus probable étant donnée une séquence de mots en entrée. Celle-ci est généralement obtenue en appliquant l'algorithme de *Viterbi*.

Vus les avantages que présentent les CRF par rapport aux autres modèles probabilistes, nous avons opté pour ce type de modèle pour la mise en œuvre de notre système de REN. Dans nos expériences, l'apprentissage des différents modèles est réalisé à l'aide de l'outil libre CRF++ toolkit [3]. La section suivante présente la méthode proposée.

4.3 REN suivant une taxonomie hiérarchique et compositionnelle

Nous commençons par présenter la problématique de l'adaptation d'une taxonomie hiérarchique et compositionnelle. Nous présentons ensuite la méthode utilisée pour résoudre ce problème.

4.3.1 Problème de classification multi-étiquettes

Comme indiqué dans la section 4.1.1, le schéma d'annotation Quæro est structuré d'une manière hiérarchique et compositionnelle. Par conséquent, un mot appartenant à une entité nommée peut se voir attribuer deux ou plus étiquettes en même temps. Par exemple, dans la phrase « Vous êtes <func.ind> <kind> directeur </kind> de l' <org.ent> école nationale d' assurance </org.ent> </func.ind> », les étiquettes « *func.ind* » et « *kind* » sont attribuées simultanément au même mot « directeur ». On se trouve donc en face d'un problème de classification multi-étiquettes. La plupart du temps, les modèles probabilistes standards d'étiquetage de séquences sont mono-étiquettes, c'est-à-dire que chaque mot doit appartenir à une seule catégorie. En effet, les modèles graphiques utilisent des données d'apprentissage annotées (en général par des experts humains) afin de créer un modèle probabiliste permettant de prédire la séquence d'étiquettes pour une séquence de mots donnée. Dans ce but, le corpus d'apprentissage utilisé est présenté de manière à associer à chaque mot une étiquette. Les mots appartenant aux entités nommées sont annotés avec leurs étiquettes correspondantes. Les autres mots sont annotés avec l'étiquette *O* (pour *outside*). Afin de résoudre le problème de segmentation qui se produit lorsque plusieurs entités nommées ayant la même catégorie se succèdent, un indicateur de position est associé aux étiquettes. Cela permet d'indiquer pour chaque mot appartenant à une entité nommée s'il correspond au début (*B* pour *begin*) ou à l'intérieur de l'entité (*I* pour *inside*). Cette présentation des données est connue sous le nom d'encodage *BIO* (Ramshaw et Marcus 1995). Un exemple est donné dans la table 4.6.

3. http://crfpp.googlecode.com/svn/trunk/doc/index.html?source=navbar

Séquence de mots :	je	suis	avec	le	ministre	Bruno	Le	Maire
Séquence d'étiquettes :	O	O	O	O	fonc-B	pers-B	pers-I	pers-I

TABLE 4.6 – *Exemple d'encodage BIO.*

4.3.2 Méthode proposée

La méthode proposée afin de résoudre le problème de classification multi-étiquettes consiste à diviser ce dernier en plusieurs problèmes de classification mono-étiquettes. Nous définissons donc trois niveaux d'annotation : i) un premier niveau s'intéresse à l'annotation des 32 catégories, ii) un second niveau s'occupe de l'annotation des composants et iii) un dernier niveau permet l'imbrication des annotations. De manière générale, chaque niveau d'annotation comporte deux phases distinctes : (a) apprentissage : apprendre un modèle d'annotation à partir de données d'apprentissage , (b) annotation : utiliser le modèle obtenu afin d'annoter une nouvelle séquence de mots. Ce processus est résumé en figure 4.4 et nous le précisons dans les paragraphes qui suivent.

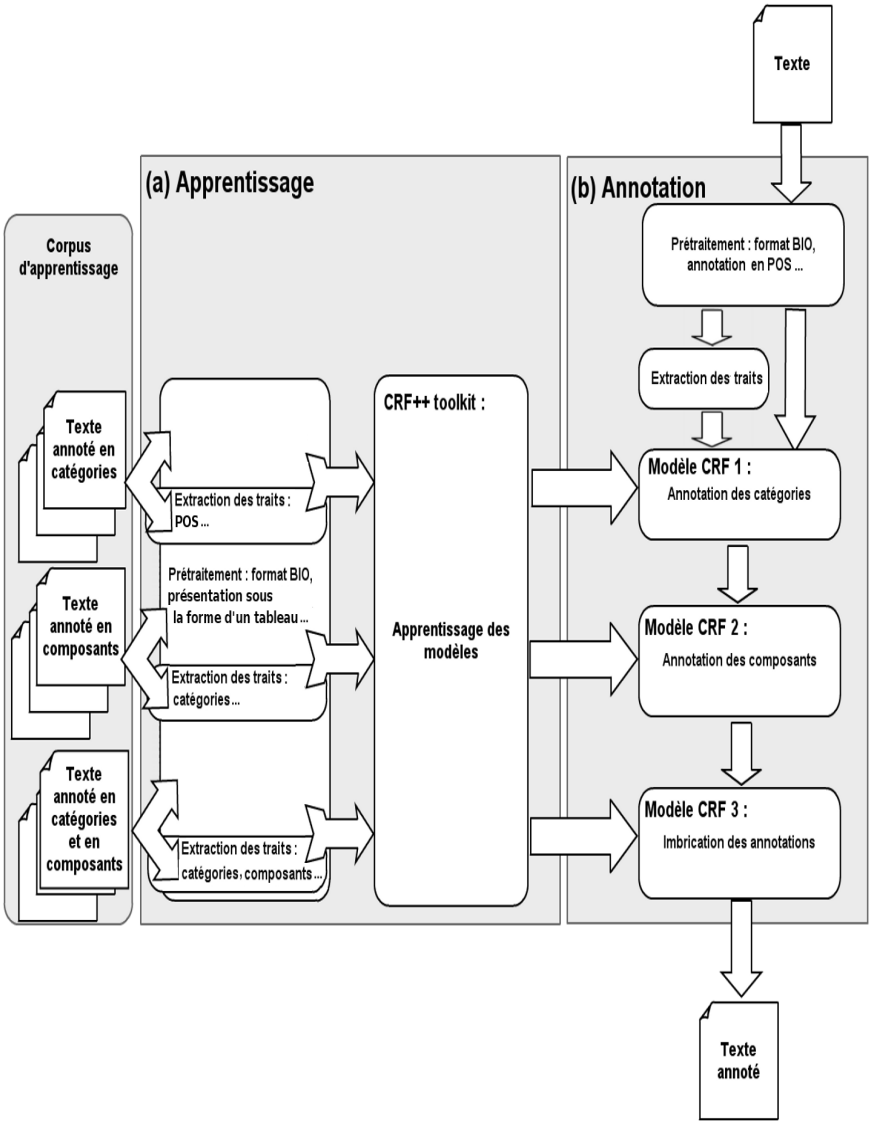

FIGURE 4.4 – *Processus général d'annotation selon la taxonomie Quæro*

(a) Apprentissage des modèles d'annotation

Cette étape consiste à apprendre un modèle CRF séparé pour chaque niveau d'annotation. Le corpus d'apprentissage utilisé est présenté sous la forme d'un tableau de n lignes et de m colonnes. Chaque ligne présente un exemple d'apprentissage (x, y). x correspond aux $m-1$ premières colonnes représentant le mot de la transcription, ainsi que ses différents traits. y correspond à la dernière colonne qui représente l'annotation du mot au format *BIO*. Nous présentons ci-dessous les différents traits utilisés pour chaque niveau d'annotation.

Premier niveau d'annotation : catégories

Mot	**POS**	**Majuscule ?**	**Catégorie au format** *BIO*
vous	PPERP	0	O
êtes	VEP	0	O
directeur	NMS	0	func.ind-b
de	DETFS	0	func.ind-i
l'	DETFS	0	func.ind-i
école	NFS	0	org.ent-b
nationale	AFS	0	org.ent-i
d'	PREPADE	0	org.ent-i
assurance	NFS	0	org.ent-i
à	PREP	0	O
Paris	XLOC	1	loc.adm.town-b

TABLE 4.7 – *Format du corpus d'apprentissage utilisé pour apprendre le modèle d'annotation des catégories.*

Le premier modèle CRF vise à annoter un texte avec les catégories d'une manière plate, c'est-à-dire sans imbrication des annotations. Pour ce faire, nous avons tout d'abord séparé les annotations imbriquées dans le corpus d'apprentissage (par exemple, transformer « <func.ind> maire de <loc.adm.town> Paris </loc.adm.town> </func.ind> » en « <func.ind> maire de </func.ind> <loc.adm.town> Paris </loc.adm.town> »). Nous avons ensuite présenté ce corpus sous la forme d'un tableau de couples (x, y) afin de permettre l'apprentissage du modèle. x correspond au mot de la transcription et ses différents traits. y correspond à la catégorie du mot au format *BIO*. Pour la prédiction des étiquettes, nous avons inclus comme traits les types d'informations suivants :

- les informations contextuelles : une fenêtre contextuelle de $[-2; +2]$ par rapport à la position courante est utilisée afin de prédire, à partir des mots de la transcription et de leurs traits, l'étiquette du mot courant.
- la présence de la majuscule : ce trait est employé seulement pour les transcriptions manuelles dont les noms propres sont capitalisés. Il prend 1 si la majuscule est présente et 0 sinon.
- les étiquettes morpho-syntaxiques des mots : pour associer aux mots leur étiquettes morpho-syntaxiques, nous avons utilisé le logiciel libre LIA_TAGG [4]. Cet étiqueteur est basé sur une approche générative à base de HMM. Le jeu d'étiquettes des parties du discours (POS, pour *part of speech*) utilisé comporte une centaine de classes dont certaines à caractère sémantique, par exemple, *XPERS* pour les noms de personnes et *XLOC* pour les noms de lieux. Les mots inconnus de LIA_TAGG sont annotés avec l'étiquette *MOTINC*. Voici un exemple d'une phrase annotée en POS :

4. Téléchargeable sur : `http://lia.univ-avignon.fr/fileadmin/documents/Users/Intranet/chercheurs/bechet/download_fred.html`

je	suis	avec	le	ministre	Bruno	Le	Maire
PPER1S	VE1S	PREP	DETMS	NMS	XPERS	MOTINC	MOTINC

TABLE 4.8 – *Exemple d'une phrase annotée en POS en utilisant LIA_TAGG.*

Nous avons enrichi le lexique de base de LIA_TAGG avec environ 111 600 noms propres (dont 30 300 noms de personne, 18 700 noms d'organisation et d'entreprise et 62 600 noms de lieu) [5]. Cela permet un premier niveau d'annotation basé sur le lexique.

La table 4.7 présente le format du corpus d'apprentissage utilisé pour apprendre le modèle CRF. À l'issu de cette étape, nous avons créé deux version du modèle CRF permettant l'annotation en catégories : une première version, dédiée au texte capitalisé, exploite la présence de la majuscule comme trait, une deuxième version, dédiée au texte non capitalisé, n'utilise pas ce trait.

Deuxième niveau d'annotation : composants

Mot	**Catégorie au format** *BIO*	**Composant au format** *BIO*
vous	O	O
êtes	O	O
directeur	func.ind-b	kind-b
de	func.ind-i	O
l'	func.ind-i	O
école	org.ent-b	name-b
nationale	org.ent-i	name-i
d'	org.ent-i	name-i
assurance	org.ent-i	name-i
à	O	O
Paris	loc.adm.town-b	name-b

TABLE 4.9 – *Format du corpus d'apprentissage utilisé pour apprendre le modèle d'annotation des composants.*

Le deuxième niveau d'annotation s'intéresse à annoter un texte avec les composants. Pour l'apprentissage du modèle correspondant, nous avons présenté le corpus d'apprentissage sous la forme d'un tableau de couples (x, y). x correspond au mot de la transcription et ses différents traits. y correspond à l'étiquette du mot au format *BIO*. Nous avons inclus comme traits les deux types d'informations suivants :

– les informations contextuelles : une fenêtre contextuelle de $[-2; +2]$ par rapport à la position courante est utilisée afin de prédire l'étiquette du mot courant.

5. Ces lexiques sont majoritairement issus du système Nemesis de (Fourour 2002) (section 1.5.1)

– les étiquettes concernant les catégories : nous avons utilisé les annotations en catégories afin d'associer à chaque mot une étiquette au format *BIO*.

La table 4.9 montre le format du corpus d'apprentissage utilisé pour apprendre le modèle CRF concernant le deuxième niveau d'annotation.

Troisième niveau d'annotation : imbrication des entités

Mot	**Catégorie au format** *BIO*	**Composant au format** *BIO*	**Imbrication au format** *BIO*
vous	O	O	O
êtes	O	O	O
directeur	func.ind-b	kind-b	func.ind-b
de	func.ind-i	O	func.ind-i
l'	func.ind-i	O	func.ind-i
école	org.ent-b	name-b	func.ind-i
nationale	org.ent-i	name-i	func.ind-i
d'	org.ent-i	name-i	func.ind-i
assurance	org.ent-i	name-i	func.ind-i
à	O	O	O
Paris	loc.adm.town-b	name-b	loc.adm.town-b

TABLE 4.10 – *Format du corpus d'apprentissage utilisé pour apprendre le modèle permettant l'imbrication des annotation.*

Dans le premier et le deuxième niveau d'annotation, les entités nommées et les composants sont annotés d'une manière plate, c'est-à-dire sans imbrication des entités. Nous avons donc besoin de modifier la limite droite de certaines entités afin d'éviter de chevaucher d'autres entités. Par exemple, dans la phrase « Vous êtes <**func.ind**> <kind> directeur </kind> de l' <**/func.ind**> <org.ent> école nationale d' assurance </org.ent> », nous avons besoin de changer la limite droite de l'entité « *func.ind* » afin d'inclure l'entité « *org.ent* » (« Vous êtes <**func.ind**> <kind> directeur </kind> de l' <org.ent> école nationale d' assurance </org.ent> <**/func.ind**> »). Pour ce faire, nous avons entraîné un troisième modèle CRF incluant deux types d'informations :

– les informations contextuelles : une fenêtre contextuelle de $[-2; +2]$ par rapport à la position courante est utilisée afin de prédire l'étiquette du mot courant.
– les étiquettes concernant les catégories et les composants : nous avons utilisé les annotations manuelles concernant les catégories et les composants afin d'associer aux mots leur étiquettes au format *BIO*.

La table 4.10 montre le format du corpus d'apprentissage utilisé pour apprendre le modèle CRF concernant le troisième niveau d'annotation.

(b) Annotation d'un nouveau texte

Une fois les différents modèles appris, l'annotation d'un nouveau texte consiste en quatre étapes successives :

1. prétraitement : présenter le texte sous la forme d'un tableau dont la première colonne correspond au mot de la transcription et les autres colonnes correspondent aux différents traits utilisés dans le premier niveau d'annotation (les étiquettes POS avec LIA_TAGG et la majuscule si le texte est capitalisé).
2. annotation en catégories : appliquer le modèle CRF du premier niveau d'annotation afin de prédire une étiquette d'entité nommée pour chaque mot de la transcription.
3. annotation en composants : appliquer le modèle CRF du deuxième niveau d'annotation sur la sortie du premier modèle CRF. Le but ici est de décomposer les entités nommées en composants. Cela ne concerne que les entités reconnues dans le premier niveau d'annotation, c'est-à-dire que les composants ne peuvent apparaître qu'à l'intérieur des entités nommées.
4. imbrication des annotations : appliquer le modèle CRF concernant le troisième niveau d'annotation sur la sortie du deuxième modèle CRF. Le but ici est de changer la limite droite de certaines entités afin d'inclure d'autres entités.

4.4 Expérimentations

Nous utilisons les données d'évaluation fournies dans le cadre de la campagne d'évaluation ETAPE (corpus **Etape-test**) pour mesurer les performances de notre système. Les annotations sont évaluées en termes de F-mesure et de SER sur les transcriptions manuelles et automatiques.

4.4.1 Évaluation sur les transcriptions manuelles

Les transcriptions manuelles sont réalisées par des experts humains. Elles ne contiennent pas d'erreur de transcription (WER = 0) et les majuscules ainsi que les signes de ponctuation sont présents. Les majuscules sont employées seulement pour les noms propres et non pas au début des phrases. Nous menons trois évaluations différentes sur ces données : nous mesurons tout d'abord l'apport des différents descripteurs, puis nous évaluons l'impact de la présence de certains phénomènes spécifiques à la parole et, enfin, nous étudions l'impact de l'adoption d'une taxonomie fine.

Impact de la prise en compte des différents traits

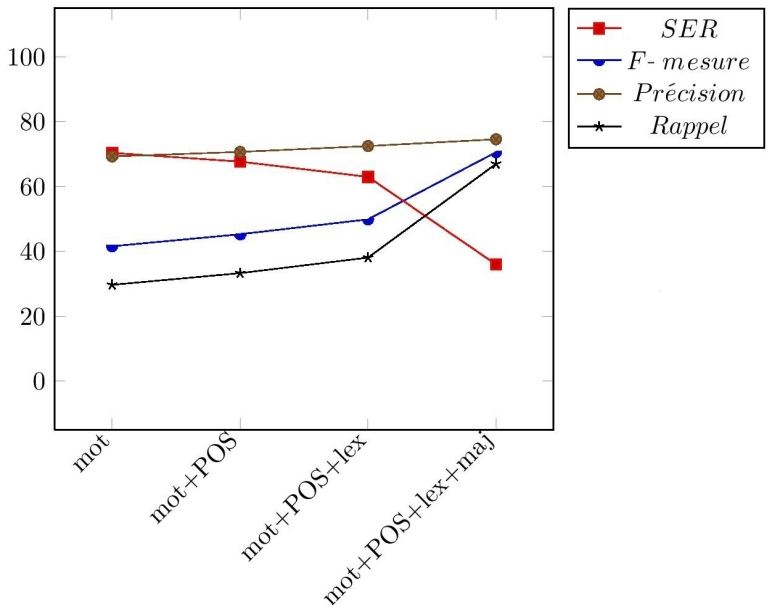

FIGURE 4.5 – *Performances du système de REN en fonction des différentes caractéristiques sur les transcriptions manuelles*

Nous évaluons l'impact des différents traits discriminants employés pour reconnaître les entités nommées. Pour ce faire, nous testons différentes versions de système en fonction des traits utilisés :

- *mot* : mots composant la transcription.
- *mot+POS* : mots composant la transcription et leurs étiquettes morpho-syntaxiques obtenues en utilisant l'étiqueteur LIA_TAGG.
- *mot+POS+lex* : mots composant la transcription et leurs étiquettes morpho-syntaxiques obtenues en utilisant l'étiqueteur LIA_TAGG après l'enrichissement de lexique avec des connaissances connues *a priori*.
- *mot+POS+lex+maj* : mots composant la transcription, leurs étiquettes morpho-syntaxiques, les connaissances connues *a priori* et les informations concernant la présence de majuscule.

Une fenêtre contextuelle de $[-2;+2]$ par rapport à la position courante est utilisée pour les différentes expérimentations. La figure 4.5 montre les résultats obtenus pour les sept catégories principales après l'ajout progressif des différents traits. En utilisant seulement les mots composant la transcription, le système se base seulement sur les formes des mots, que ce soit dans la position courante ou dans le contexte immédiat, pour la prédiction des étiquettes. Cela pose un problème pour l'annotation des formes qui n'apparaissent pas dans le corpus d'entraînement. On le voit notamment à travers le rappel relativement faible (environ 29,70 %). L'ajout des informations morpho-syntaxiques permet d'élargir la portée des règles d'annotation. Nous notons une augmentation de F-mesure d'environ 4 % et une diminution du SER d'environ 3 %. L'utilisation des informations lexicales qui correspondent à des connaissances connues *a priori* permet de

limiter les cas d'ambiguïtés (contexte ambigu). Cela permet une amélioration des performances qui concerne principalement les noms de personne (un gain d'environ 17 % en F-mesure) et de lieu (un gain d'environ 6 % en F-mesure). Le gain est significatif lorsqu'on ajoute les informations concernant la présence de la majuscule (environ 20 % en termes de F-mesure et 27 % en termes de SER). En effet, la majuscule est employée seulement pour les noms propres dans le corpus d'apprentissage. Cela représente donc un indicateur saillant de la présence d'une entité nommée pour le modèle d'annotation.

Impact de la présence des disfluences

Nous analysons dans cette section l'impact de la présence des disfluences sur la qualité de REN. Les disfluences sont des phénomènes inhérents à la parole spontanée. Elles correspondent à toute interruption, perturbation ou accroc dans le flux de la parole. Ces phénomènes peuvent se produire lorsqu'il s'agit de prononcer une entité nommée. Par exemple, il arrive que le locuteur utilise des pauses pleines pour compenser un problème lexical. Nous recensons quatre types principaux de disfluences :

- les hésitations : correspondent aux pauses pleines telles que euh, hum, etc. Par exemple, « AF **euh euh euh** 347 ».
- les répétitions : correspondent aux mots (ou groupes de mots) qui se répètent successivement plus d'une fois. Par exemple, « AF **AF** 347 ».
- les auto-corrections : se produisent lorsque le locuteur fait une ou plusieurs erreurs, et se corrige dans le même énoncé. Le mot (ou les mots) est (sont) prononcé(s) complètement. Par exemple, « AF **346** 347 ».
- les amorces : se produisent lorsque le locuteur arrête de prononcer un mot avant la fin normale de celui-ci. Par exemple, « AF 3**(47)** ».

Nous recensons 2 441 hésitations dans le corpus d'évaluation. Environ 8 % de ces hésitations touchent les entités nommées, que se soit à l'intérieur même de l'entité (36 fois) ou dans le contexte immédiat (342 fois). Nous pouvons remarquer que ces phénomènes touchent les différentes catégories. Les entités contenant des hésitations affichent une précision de 58 %, ce qui est faible par rapport à la précision globale du système (74,60 %). En regardant plus précisément les différents types d'erreurs, nous constatons un nombre important d'entités qui ne sont pas annotées (12 parmi les 36) qui concerne notamment les entités construites autour de noms communs. Par exemple, « groupe de euh local taliban » et « ex euh gaulliste ».

Les répétitions qui se produisent à l'intérieur même des entités nommées sont moins fréquentes (seulement 12 fois). Ces répétitions concernent en grande partie des mots vides. Par exemple, « patron du du FMI » et « vers le le 12 ». Dans ce cas, les entités mal reconnues le sont, à cause d'erreurs de délimitation (6 parmi les 12).

Impact de l'adoption d'une taxonomie fine

Nous menons une évaluation séparée pour chaque niveau de granularité de la taxonomie Quæro. La table 4.11 présente les performances obtenues en termes de SER.

	Insertion (%)	**Suppression (%)**	**Substitution (%)**	**SER (%)**
7 catégories	9,40	19,50	13,50	36,00
32 catégories	9,70	20,40	15,50	38,20
composants	10,70	19,30	11,70	36,70
32 catégories +composants	9,70	20,30	13,70	37,50

TABLE 4.11 – *Performances en termes de SER pour les différents niveaux de granularité.*

Nous remarquons que l'utilisation d'une typologie plus fine (32 catégories) engendre une dégradation des performances d'environ 2,20 % par rapport à l'utilisation seule des sept catégories principales (*pers*, *org*, *loc*, *prod*, *func*, *time*, *amount*). Cette dégradation est due notamment à une augmentation de nombre de substitutions (environ 2 %). Cela montre une difficulté de distinction entre certaines sous-catégories qui concerne notamment les noms de lieu et de produit.

La table 4.12 présente les performances réalisées par catégorie et par composants. Ces résultats s'avèrent satisfaisants pour certaines catégories classiques telles que « *pers.ind* » et « *loc.nat* » et moins bons pour d'autres catégories telles que « *loc.fac* » et « *prod.object* ». On constate notamment un rappel faible pour ces catégories. Cela est dû essentiellement à une faible apparition de ces catégories dans le corpus d'apprentissage et à une ambiguïté d'annotation.

La table 4.13 montre les différents types d'erreurs comptabilisés dans le SER par catégorie. Les erreurs de suppression concernent particulièrement des entités nommées ne contenant pas de noms propres. Par exemple : « exportateurs de fruits et légumes espagnols », « spécialiste en climatologie », « chiraquiens ». En ce qui concerne les erreurs de substitutions, nous relevons une confusion entre les sous-catégories. Ceci concerne notamment les noms d'organisation (confusion entre « *org.ent* » et « *prod.media* », « *org.adm* » et « *org.ent* »), les sous-catégories des produits et de « *time*. Les erreurs de frontières concernent les « *amount* » et les « *time* » et les « *personne* » (64 fois).

La table 4.14 présente les performances réalisées par composant. Nous observons que les composants concernant les noms de personne (« *pers.first* » et « *pers.last* ») et de « *time* (« *day* », « *week* », « *month* » et « *year* ») sont les mieux reconnus. Les composants transverses affichent de moins bons résultats. Nous voyons clairement apparaître des difficultés concernant les composants de type « *kind* », « *qualifier* » et « *title* ». D'une manière générale, nous remarquons que les erreurs de substitution concernant les composants sont moins fréquentes que celles concernant leurs catégories.

7 catégories	P (%)	R (%)	F (%)	32 catégories	P (%)	R (%)	F (%)
amount	67,6	63,5	65,5	amount	67,6	63,5	65,5
pers	83	83,6	83,3	pers.ind	85,2	84,4	84,8
				pers.coll	50,3	49,2	49,7
				pers.other	0	0	0
time	69	76,3	72,4	time.date.abs	51,9	36,5	42,8
				time.date.rel	72,8	76,7	74,7
				time.hour.abs	62,5	54,3	58,1
				time.hour.rel	67,6	82,1	74,2
loc	86,8	73,7	79,7	loc.oro	0	0	0
				loc.fac	33,3	8,6	13,7
				loc.add.phys	0	0	0
				loc.add.elec	83,3	83,3	83,3
				loc.adm.town	78,8	65,2	71,4
				loc.adm.reg	57,8	55,3	56,5
				loc.adm.nat	87,5	86,2	86,9
				loc.adm.sup	95,5	63,6	76,4
				loc.phys.geo	100	13.3	23.5
				loc.phys.hydro	0	0	0
				loc.phys.astro	0	0	0
prod	64,3	36,2	46,3	prod.object	33,3	3,4	6,2
				prod.art	28,6	11,5	16,4
				prod.media	68,5	67,7	68,1
				prod.fin	55,6	11,9	19,6
				prod.soft	0	0	0
				prod.award	80	30,8	44,4
				prod.serv	0	0	0
				prod.doctr	0	0	0
				prod.rule	50	60	54.5
				prod.other	0	0	0
				prod.unk	0	0	0
func	72,7	45,5	56	func.ind	74,6	49,9	59,8
				func.coll	66,7	37	47,6
org	57,7	52,4	54,9	org.ent	41,5	50,8	45,7
				org.adm	67,7	46,2	54,9
Total	74,6	67	70,6	Total	73,9	66	69,7

TABLE 4.12 – *Performances en termes de F-mesure par catégories principales et par sous-catégories. (F : F-mesure, P : précision, R : rappel).*

	Insertion	Suppression	Erreur type	Erreur frontière
amount	63	101	16	139
pers	77	143	75	79
time	171	196	55	94
loc	26	139	125	14
prod	28	154	96	11
func	55	282	15	45
org	73	177	101	19

TABLE 4.13 – *Types d'erreurs par catégorie d'entité nommée.*

Composant	P (%)	R (%)	F (%)
kind	73,2	38,4	50,4
extractor	100	25	40
qualifier	45,2	24,4	31,7
title	47,1	32	38,1
val	86,2	83,8	85
unit	90,9	84,4	87,6
object	57,8	66,2	61,7
range-mark	92,2	66,2	77
day	88,2	83,3	85,7
week	91,2	79,5	84,9
month	72,1	66,2	69
year	90,2	87,4	88,8
century	100	100	100
reference-era	16,7	50	25
time-modifier	62,2	66,5	64,3
award-cat	0	0	0
demonym	70.6	49	57.9
name	65,8	70,6	68,1
name.last	83,4	82,4	82,9
name.first	82.5	90.7	86.4
name.nickname	0	0	0

TABLE 4.14 – *Performance du système par composant sur le corpus d'évaluation. (F : F-mesure, P : précision, R : rappel).*

4.4.2 Évaluation sur les transcriptions automatiques

Les transcriptions automatiques sont générées automatiquement par les systèmes de RAP. Par rapport aux transcriptions manuelles, elles se caractérisent par l'absence de majuscule, le manque de ponctuation et la présence d'erreurs de transcription. Ces dernières varient d'un système de RAP à un autre. La table 4.15 affiche les résultats obtenus par notre système pour les différents catégories et composants sur les transcriptions automatiques dans le cas d'un taux d'erreur de mot (WER) égale à 23 %. Dans l'ensemble, il est évident que la REN présente des difficultés supplémentaires lorsqu'il s'agit des transcriptions automatiques par rapport aux transcriptions manuelles. Les résultats globaux montrent une dégradation d'environ 25 % en termes de SER et d'environ 17 % en termes de F-mesure. Manifestement, cela est du aux erreurs de transcriptions et à l'absence de la majuscule. On remarque que les performances des catégories dont les entités nommées représentent des noms propres sont les plus touchées. Par exemple, « *pers.ind* » et « *loc.adm.town* ». En ce qui concerne les composants, nous observons un taux de reconnaissance satisfaisant pour les composants relatifs aux « *amount* » et aux « *time* ».

Catégorie	P (%)	R (%)	F (%)	Composant	P (%)	R (%)	F (%)
amount	51,1	48,1	49,6	kind	64,9	43,3	42,3
pers.ind	62	48,9	54,71	extractor	50	25	33,3
pers.coll	44,4	36,9	40,3	qualifier	37,1	22,8	28,2
pers.other	0	0	0	title	38,2	26,8	31,5
time.date.abs	41	27,6	33	val	71,4	69	70,2
time.date.rel	63,7	67,4	65,5	unit	75,1	75,8	75,4
time.hour.abs	37,5	29,2	32,9	object	46,4	51,9	49
time.hour.rel	64,4	59, 7	62	range-mark	81,8	46,5	59,3
loc.oro	50	50	50	day	65,7	79,3	71,8
loc.fac	50	6,4	11,3	week	80	75,6	77,7
loc.add.phys	0	0	0	month	62,1	57,7	59,8
loc.add.elec	60	37,5	46,1	year	84,2	63,4	72,4
loc.adm.town	53,6	38	44,4	century	100	66,6	80
loc.adm.reg	60	45,6	51,8	reference-era	25	50	33,3
loc.adm.nat	73,6	82,8	78	time-modifier	56,4	59,5	57,9
loc.adm.sup	59,2	48,5	53,3	award-cat	0	0	0
loc.phys.geo	0	0	0	demonym	65,5	39,9	49
loc.phys.hydro	0	0	0	name	63,8	52	57,4
loc.phys.astro	0	0	0	name.last	47,1	36,6	41,2
prod.object	0	0	0	name.first	68	44,2	53,5
prod.art	0	0	0	name.nickname	78,5	17,2	28,2
prod.media	73,2	46,4	56,8				
prod.fin	62,5	6,4	11,6				
prod.soft	0	0	0				
prod.award	0	0	0				
prod.serv	0	0	0				
prod.doctr	33,3	33,3	33,3				
prod.rule	0	0	0				
prod.other	0	0	0				
prod.unk	0	0	0				
func.ind	64,7	43	51,6				
func.coll	54,5	30,1	38,8				
org.ent	40,1	36,7	38,3				
org.adm	60,4	37	46				
Total	56,3	44,7	50	Total	62,7	48,1	54,5

TABLE 4.15 – *Performance du système par catégorie et par composant sur le corpus d'évaluation (WER=23). (F : F-mesure, P : précision, R : rappel).*

Impact de la présence des erreurs de transcription

Nous évaluons ici comment les performances de la REN évoluent en fonction du WER. La figure 4.6 montre les résultats obtenus par notre système pour différents WER : 0 % (s0 : transcriptions manuelles), 23 % (s23), 25 % (s25), 30 % (s30) et 35 % (s35). Les résultats détaillés sont présentés dans la section A.2.

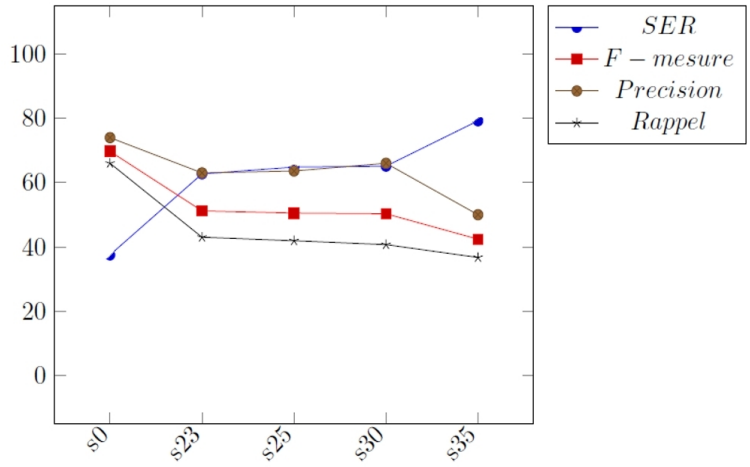

FIGURE 4.6 – *Performances du système de REN en fonction des différents taux d'erreur de mots (WER)*

Trivialement, nous pouvons remarquer que l'augmentation du WER dégrade les performances du système de REN (Miller et al. 2000). La dégradation est plus importante lorsqu'on passe d'un taux de WER de 30 % à un taux de 35 %.

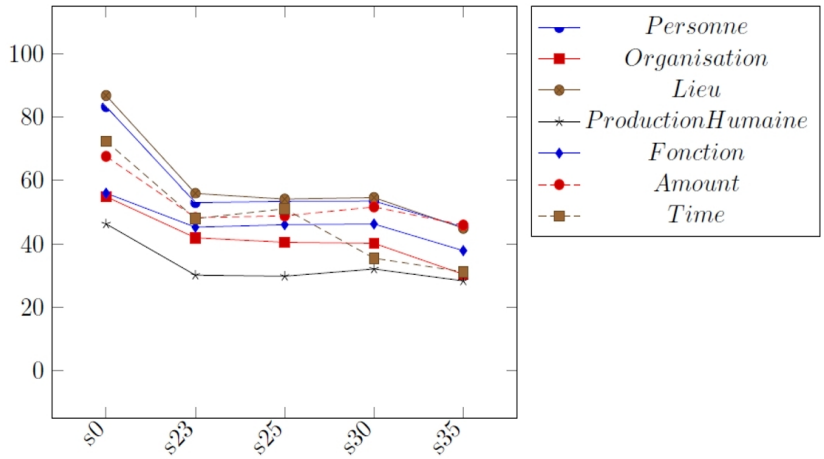

FIGURE 4.7 – *SER par catégorie en fonction des différents taux d'erreur de mots (WER)*

La figure 4.7 montre le SER obtenu pour les sept catégories principales pour les différents WER. Elle montre que les résultats pour les différentes catégories sont affectés par l'augmentation du WER. Nous remarquons que la dégradation est plus marquée pour les noms de personne et de lieu. Cela s'explique par le faite que ces entités se composent généralement par des noms propres dont la transcription est plus difficile.

Impact de la prise en compte des majuscules

Nous évaluons ici l'impact de la prise en compte de la majuscule sur la qualité de REN. Nous nous plaçons dans le cas d'un WER de 24 %. Les majuscules ont été ajoutées manuellement aux noms propres. La table 4.16 montre les performances de notre système sur les transcriptions sans majuscule (WER = 24 sans majuscules) et sur les transcriptions capitalisées (WER = 24 avec majuscules).

	P (%)	**R (%)**	**F (%)**	**SER (%)**
WER = 24 sans majuscules	63,67	43,28	51,56	65,67
WER = 24 avec majuscules	62,39	39,63	48,47	60,11

TABLE 4.16 – *Performance du système avec et sans la prise en compte de la majuscule sur le corpus d'évaluation (WER=24). (F : F-mesure, P : précision, R : rappel).*

Comme pour les transcriptions manuelles, les résultats obtenus montrent l'influence positive de la prise en compte de la majuscule. Nous notons une augmentation de F-mesure d'environ 3 % et une diminution du SER d'environ 5 %. L'augmentation d'environ 4 % du rappel montre que l'intégration de la majuscule comme trait permet de reconnaitre des entités nommées qui n'ont pas été reconnues lorsqu'on n'exploite pas cette information.

4.5 DISCUSSION

Nous discutons ici les performances de notre système en le comparant aux systèmes de la campagne d'évaluation ETAPE. La table 4.17 reporte les résultats de notre système, nommé **LL_NE**[6], ainsi que les résultats des autres systèmes participant à la campagne ETAPE (après l'étape d'adjudication).

	Manuel	**Rover**	**WER=23**	**WER=24**	**WER=25**	**WER=30**	**WER=35**
S1	**35,38**	-	-	67,14	-	-	-
S2	36,38	**57,17**	**59,04**	64,44	**61,88**	**61,59**	**76,09**
S3	37,31	67,72	67,87	67,44	70,69	69,88	84,92
S4	37,99	63,22	66,91	63,52	68,57	68,1	79,9
S5	39,16	64,28	68,92	65,38	69,66	69,02	86,09
S6	44,64	69,51	73,58	71,93	73,61	74,61	85,62
S7	49,96	87,3	97,73	76,29	91,90	93,85	98,65
S8	85,56	97,88	100,48	94,2	98,79	98,27	100,71
LL_NE	37,50	57,89	62,76	**60,11**	64,79	65,17	79,21

TABLE 4.17 – *Performances en termes de SER des systèmes participants à la campagne d'évaluation ETAPE*

6. La version du système développé dans le cadre de notre participation à la campagne ETAPE ne permet pas l'annotation des composants. Nous nous sommes intéressés seulement aux catégories (premier et troisième niveaux d'annotation). Nous avons étendu ensuite notre système afin d'annoter aussi les composants.

Le système S1 (Friburger 2002) obtient les meilleurs résultats sur les transcriptions manuelles. Ce système est basé sur une approche orientée connaissances qui consiste à délimiter et à catégoriser les entités nommées par cascades de transducteur. Les systèmes S2, S3 et S6 sont basés sur une approche à base de CRF. Le système S2 (Raymond 2013) utilise 68 modèles CRF binarisés. Un modèle est créé par type ou par composant. Les résultats des différents modèles CRF sont ensuite alignés ensemble. Le système S3 (Dinarelli et Rosset 2011) utilise un modèle CRF pour l'annotation des composants et un PCFG (*Probabilistic Context Free Grammar*) pour reconstituer les entités. Le système S4 (Nouvel 2012) utilise des méthodes de fouille de données séquentielles hiérarchiques. Des motifs séquentiels pour la REN sont extraits automatiquement à partir de données d'apprentissage (enrichies par des informations morpho-syntaxiques et sémantiques). Ces motifs permettent de détecter séparément le début et la fin des entités nommées, d'estimer les étiquettes probables et d'en déterminer l'annotation la plus vraisemblable.

Nos résultats sont proches de ceux obtenus par ces différents systèmes. Cela montre qu'avec la disponibilité d'un corpus d'apprentissage de grande taille, les méthodes statistiques permettent d'offrir une solution rapide et facile pour la REN. L'évaluation sur un corpus de test ayant les mêmes caractéristiques que le corpus d'entraînement (même domaine, mêmes sources, etc.) représente un point important pour les approches orientées données. De manière générale, nous voyons que les différents systèmes présentent des dégradations assez importantes lorsqu'il s'agit des transcriptions automatiques. Le WER a un effet directe sur la qualité de REN.

Le rover consiste à combiner les sorties de plusieurs systèmes afin de sélectionner à chaque fois la meilleure proposition. Les résultats obtenus sur ces transcriptions sont meilleurs que ceux obtenus sur les autres transcriptions automatiques. Cela est dû principalement à l'amélioration de la qualité de la transcription. Pour les transcriptions ayant un WER égale à 24 %, les majuscules ont été manuellement rétablies pour les entités nommées. Les résultats pour ces transcriptions s'avèrent meilleurs que ceux obtenus sur les transcriptions ayant un WER égale à 23 %, et cela malgré une dégradation de 1 % de la qualité de la transcription. Cela s'explique par le fait que l'intégration de la majuscule comme trait permet une amélioration des résultats de REN. Nous notons que le système S1 affiche des résultats moins bons sur ces transcriptions par rapport aux systèmes à base de CRF.

4.6 Conclusion

Nous avons présenté dans ce chapitre une étude approfondie concernant la REN dans les transcriptions de la parole. Cette tâche, qui est considérée ici comme étant un post-traitement qui vient en aval de la phase de transcription, montre des résultats satisfaisants pour les transcriptions manuelles. Les résultats expérimentaux montrent que l'adoption d'une

taxonomie fine entraîne une dégradation relativement faible sur les performances globales du système (environ 1,50 % de SER). Ces résultats prouvent aussi l'importance relative de la présence de la majuscule et de la ponctuation pour la REN. D'un autre coté, les résultats restent insuffisants pour les transcriptions automatiques. L'absence de majuscule et de ponctuation ainsi que les erreurs de transcription ont un effet direct sur la qualité de REN. Il serait donc intéressant d'envisager une interaction plus forte avec le niveau RAP afin de contourner ces problèmes. Pour cela, nous proposons dans le chapitre suivant une méthode permettant d'incorporer directement la tâche de REN dans le processus de RAP.

5 Intégration de la reconnaissance des entités nommées au processus de reconnaissance de la parole

Sommaire

Comme il est présenté dans le chapitre précédent, la REN en parole consiste généralement à appliquer des méthodes développées pour l'écrit aux sorties issues d'un processus de RAP. Les systèmes de REN sont adaptés afin de prendre en compte les phénomènes spécifiques aux trans-

criptions par rapport aux textes bien formés. L'absence de capitalisation [1] et le manque de ponctuation induisent respectivement des ambiguïtés graphiques et de segmentation. La présence de disfluences et d'erreurs de reconnaissance constitue deux autres difficultés caractéristiques du traitement automatique de la parole auxquelles se confrontent les systèmes de REN.

Dans ce contexte, la REN ne peut être considérée comme un simple traitement qui vient en aval de la tâche de reconnaissance du signal. Il est donc intéressant d'interagir avec le système de RAP. Un premier niveau d'interaction consiste par exemple à s'appuyer sur les scores de confiance associés aux mots reconnus pour détecter les entités nommées mal reconnues (Sudoh et al. 2006). L'interaction avec le système de RAP peut être encore plus forte lorsque l'on exploite des sorties intermédiaires comme les N meilleures hypothèses (Zhai et al. 2004) ou les graphes de mots (Favre et al. 2005) afin d'améliorer la précision de la REN. Un pas supplémentaire peut être franchi lorsque l'on cherche à tirer parti des connaissances exploitées en RAP pour prédire des régions où les entités nommées seront incorrectement retranscrites, voire même pour corriger les transcriptions (Parada et al. 2011). Dans ce travail, nous souhaitons aller encore plus avant dans cette interaction entre RAP et REN en produisant directement des transcriptions étiquetées en entités nommées. En ce sens, nous proposons d'inclure au sein du système de RAP un processus de REN qui pourra directement exploiter les connaissances mobilisées pour la tâche de RAP telles que les mots du vocabulaire et le modèle de langage. L'idée générale étant que le système de RAP devient ainsi un système dédié à la tâche de REN en produisant en sortie des transcriptions étiquetées en entités nommées. Le système sera peut-être moins performant au niveau du taux d'erreur global sur les mots. En revanche, cette approche devrait améliorer la détection et la catégorisation des entités nommées.

Ce chapitre est organisé comme suit : nous commençons par décrire les ressources radiophoniques et les mesures de performance utilisées dans le cadre de ce travail en section 5.1. La section 5.2 commence par présenter les différents systèmes de RAP et de REN que nous avons mobilisés avant de décrire l'approche que nous mettons en œuvre. La section 5.3 évalue les performances de notre proposition à travers différentes expérimentations et propose une analyse des erreurs rencontrées. Finalement, la section 5.4 vient conclure ce travail.

1. La majorité des systèmes de RAP ne génèrent pas de majuscule. Cela s'explique d'une part par les limites du matériel à la fin des années 2000. À cette époque, comme le nombre de mots des systèmes était limité à 65 K, il était préférable d'ajouter des nouveaux mots à la place d'une variante en majuscule d'un mot déjà présent dans le lexique. D'autre part, cela est dû aussi à des considérations pratiques. Il est plus facile de retirer les majuscules du corpus d'apprentissage pour le modèle de langage que de ne garder que les majuscules pertinentes. Il est difficile de gérer les majuscules sur le premier mot des phrases, par exemple.

5.1 Présentation des corpus et des mesures de performance

Dans cette section, nous présentons le schéma d'annotation adopté dans ce travail, le corpus utilisé et les mesures de performance permettant d'évaluer notre approche.

5.1.1 Schéma d'annotation ESTER 2

Le guide d'annotation adopté est celui de la campagne ESTER 2 [2]. Les entités nommées sont ainsi réparties en 7 catégories principales, elles-mêmes divisées en 38 sous-catégories. La table 5.1 présente les différentes catégories et sous-catégories.

Seules les catégories principales sont considérées dans ce travail.

Nous recensons ici quelques directives d'annotation comme elles sont expliquées dans le guide d'annotation :

- annotation en contexte : le type d'une entité nommée dépend uniquement du contexte dans laquelle elle est employée. Il faut donc tenir compte de l'intention du locuteur et de ce qu'il a voulu désigner. Par exemple, l'entité « Ariane » est annotée en tant que personne ou production humaine, en fonction de son contexte, dans les deux phrases suivantes : « <pers.hum> Ariane </pers.hum> n'est pas venue hier. » et « <prod.vehicule> Ariane </prod.vehicule> a réussi son décollage. ».
- étendue de l'entité nommée : ne prendre en compte dans l'étiquette que l'entité nommée. C'est-à-dire que toutes les particules sont exclues de l'annotation. Par exemple, « Monsieur <pers.hum> Durand </pers.hum> », « La <loc.admi> Grande-Bretagne </loc.admi> », etc.
- imbrication : annoter les entités nommées de lieu, de fonction, d'organisation et de personne lorsque celles-ci sont imbriquées. Par exemple, « <loc.fac> université d' <loc.admi> Orléans </loc.admi> </loc.fac> ».
- disfluences : les hésitations ne sont à intégrer dans l'étiquette que si elles sont placées à l'intérieur de l'entité nommée. Par exemple, « <pers.hum> Jacques euh Chirac </pers.hum> ».

5.1.2 Corpus ESTER 2

Pour nos différentes expérimentations, nous avons utilisé le corpus distribué dans le cadre de la campagne d'évaluation ESTER 2 (Galliano et al. 2009). Ce corpus est constitué d'émissions radiophoniques accompagnées de leurs transcriptions manuelles, ainsi que de transcriptions manuelles

2. www.afcp-parole.org/camp_eval_systemes_transcription/docs/Conventions_EN_ESTER2_v01.pdf

Catégories	Sous-catégories
Personne (pers)	- **Humain réel ou fictif** (pers.hum) - **Animal réel ou fictif** (pers.anim)
Fonction (fonc)	- **Politique** (fonc.pol) - **Militaire** (fonc.mil) - **Administrative** (fonc.admi) - **Religieurse** (fonc.rel) - **Aritocratique** (fonc.ari)
Organistion (org)	- **Politique** (org.pol) - **Éducative** (org.edu) - **Commerciale** (org.com) - **Non commerciale** (org.non-profit) - **Média & divertisement** (org.div) - **Géo-socio-dministrative** (org.gsp)
Lieu (loc)	- **Géographique naturel** (loc.geo) - **Région administrative** (loc.admi) - **Éducative** (loc.edu) - **Axe de circulation** (loc.line) - **Adresse postale** (loc.addr.post) - **Téléphone et fax** (loc.addr.tel) - **Adresse électronique** (loc.addr.elec) - **Construction humaine** (loc.fac)
Production humaine (prod)	- **Moyen de transport** (prod.vehicule) - **Recomponse** (prod.award) - **Œuvre artistique** (prod.art) - **Production documentaire** (prod.doc)
Date et heure (time)	- **Date absolue** (time.date.abs) - **Date relative** (time.date.rel) - **Heure** (time.hour)
Montant (amount)	- **Âge** (amount.phys.age) - **Durée** (amount.phys.dur) - **Température** (amount.phys.temp) - **Longueur** (amount.phys.len) - **Surface et aire** (amount.phys.area) - **Volume** (amount.phys.vol) - **Poids** (amount.phys.wei) - **Vitesse** (amount.phys.spd) - **Autre** (amount.phy.other) - **Valeur monétaire** (amount.cur)

TABLE 5.1 – *Catégories et sous-catégories d'entités nommées de la campagne ESTER 2.*

rapides de radios africaines. La plupart des émissions enregistrées correspondent à des journaux d'information et à des émissions conversationnelles provenant de quatre sources : France Inter, Radio France International (RFI), Africa number one (Africa 1) et Radio Télévision du Maroc (TVME).

Une première partie de ce corpus a été utilisée comme corpus d'apprentissage pour construire les modèles de langage du système de transcription. Une seconde partie a été utilisée comme corpus de développement pour adapter le vocabulaire (corpus de développement *Dev1*) et pour ajuster certains paramètres du système proposé (corpus de développement *Dev2*). La dernière partie est utilisée pour évaluer les performances de notre système [3].

Les transcriptions manuelles des ressources radiophoniques composant les corpus de développement et de test sont enrichies par une annotation manuelle des entités nommées selon le shéma d'annotation ESTER 2.

La table 5.2 présente les principales caractéristiques des différents corpus considérés issus d'ESTER 2. La table 5.3 présente le nombre des entités nommées par catégorie. La figure 5.1 montre le pourcentage de chaque type d'entité nommée par rapport au nombre total d'entités dans les différentes parties du corpus.

Corpus	**Période**	**Nombre d'heures**
Apprentissage	1999-2004	200,0
Développement 1 (*Dev1*)	Juillet 2007	6,0
Développement 2 (*Dev2*)	Déc. 2007-Fév. 2008	2,5
Test	Déc. 2007-Fév. 2008	4,5

TABLE 5.2 – *Description des corpus utilisés (issus d'ESTER 2).*

Catégorie	**Apprentissage**	***Dev1***	***Dev2***	**Test**
Personne	17 336	1 225	436	701
Organisation	13 470	1 292	490	807
Lieu	19 076	1 303	446	864
Fonction	3 036	389	173	300
Production humaine	427	38	28	24
Montant	3 179	266	72	156
Date et heure	9 828	801	286	437

TABLE 5.3 – *Nombre des entités nommées par catégorie pour les différentes parties du corpus.*

3. Le corpus d'apprentissage et le corpus de développement *Dev1* sont les mêmes que ceux utilisés pour la campagne ESTER 2. En revanche, le corpus de test utilisé pour la campagne ESTER 2 a été découpé en deux corpus – notre corpus de développement *Dev2* et notre corpus de test – afin de disposer d'un corpus de développement indépendant du corpus *Dev1* pour l'ajustement des paramètres de notre système (cf. section 5.3.1).

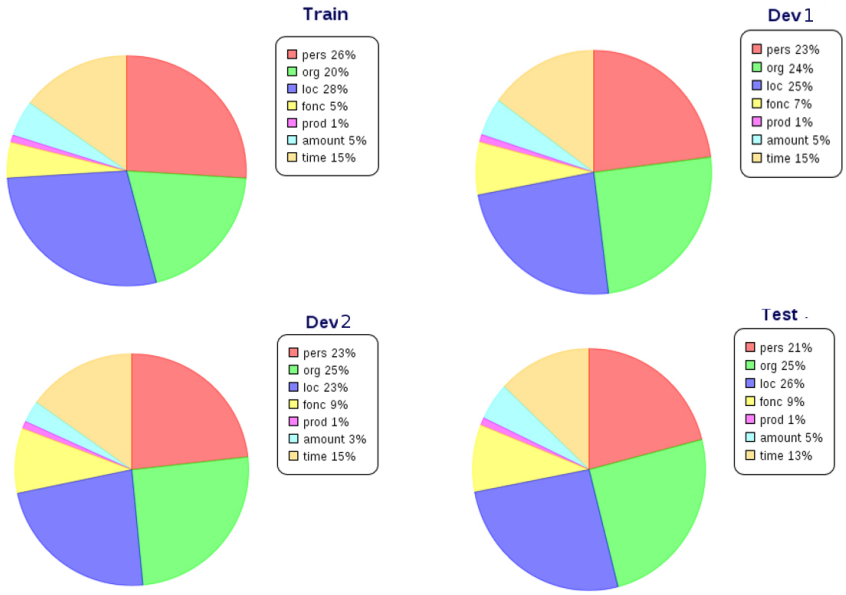

FIGURE 5.1 – *Pourcentage de chaque type d'entité nommée par rapport au nombre total d'entités dans les différentes parties du corpus ESTER 2.*

5.1.3 Mesures de performance

La qualité de la RAP est évaluée en termes de taux d'erreur sur les mots (WER) (section 2.2). La qualité de la REN est évaluée en termes de F-mesure et de SER. La F-mesure est calculée sur la base du nombre de mots contenus dans les entités nommées. Le SER est calculé sur la base du nombre d'entités nommées. Nous utilisons la variante de SER définie pour la campagne d'évaluation ESTER 2. Les types d'erreurs considérés sont : suppression (D), insertion (I), erreur de type (T), erreur de frontière (F) et erreur de type et de frontière (TF). Le SER est alors donné par l'équation 5.1 :

$$SER = \frac{D + I + 0,5 \times T + 0,5 \times F + 0,8 \times TF}{Nombre\ d'entités\ étiquetées\ de\ la\ référence} \tag{5.1}$$

Il est à noter que si l'un des mots d'une entité nommée est mal transcrit alors l'entité nommée sera considérée comme étant une erreur, soit de frontière (si sa catégorie est bien reconnue), soit de type et de frontière (si sa catégorie n'est pas bien reconnue). Par exemple, pour la référence annotée en entités nommées « **<pers>** denis astagneau **</pers>** », si la transcription correspondante annotée en entités nommées est « **<pers>** de ni astagneau **</pers>** », cette entité nommée sera considérée comme étant une erreur de frontière.

5.2 Intégration de la REN au système de transcription

Notre approche consiste à intégrer la tâche de REN au niveau du système de transcription. Dans cette optique, nous nous sommes appuyés sur

le système de RAP du LIUM et sur le système de REN LIA _NE (Béchet et Charton 2010). Nous commençons par présenter ces deux systèmes avant d'introduire notre nouveau système.

5.2.1 Système de transcription automatique de la parole du LIUM

Le système de transcription du LIUM a été développé à partir du logiciel open source Sphinx (Huang et al. 1993, Lee et al. 1990) auquel ont été ajoutées de nombreuses contributions depuis 2004. Pour nos expérimentations, nous utilisons le système développé durant la campagne d'évaluation ESTER 2. Ce système a obtenu la seconde place lors de cette campagne avec un taux d'erreur sur les mots (WER) de 19,20 %.

Le but du système de transcription est de trouver la séquence de mots la plus probable ($\hat{W} = w_1, w_2, w_3, ..., w_k$) étant donné la séquence d'observations acoustiques ($X = x_1, x_2, x_3, ..., x_p$) :

$$\widehat{W} = arg \max_{w} \ \mathbb{P}(W|X) \tag{5.2}$$

L'application de la règle de Bayes met en évidence la nécessite de deux types de modèles probabilistes, à savoir un modèle de langage qui se charge de calculer la valeur de $\mathbb{P}(W)$ et un modèle acoustique qui se charge de calculer la valeur de $\mathbb{P}(X|W)$:

$$\widehat{W} = arg \max_{w} \ \frac{\mathbb{P}(W)\mathbb{P}(X|W)}{\mathbb{P}(X)} = arg \max_{w} \ \mathbb{P}(W)\ \mathbb{P}(X|W) \tag{5.3}$$

Ces deux modèles doivent être le plus pertinent possible par rapport au domaine ciblé, notamment au niveau du vocabulaire. Le système de RAP de LIUM est dédié à la transcription des émissions radiophoniques en français. Ainsi, une partie des données d'apprentissage est constituée d'émissions radiophoniques transcrites manuellement. Ces données ont été fournies dans le cadre de la campagne ESTER 2.

Modèles de langage

Le modèle de langage se charge de contraindre le système de RAP à générer des séquences d'hypothèses de mots syntaxiquement correctes. Il permet de calculer la probabilité *a priori* $\mathbb{P}(W)$ d'une séquence de mots $W = w_1, w_2, w_3, ..., w_k$:

$$\mathbb{P}(W_k) = \mathbb{P}(w_1) \times \prod_{2}^{k} \mathbb{P}(w_i|w_1...w_{i-1}) \tag{5.4}$$

Le système de RAP de LIUM utilise des modèles de langage de type *n*-grammes. pour ce type de modèle, la probabilité d'un mot est prédite par la suite de *n-1* mots précédents (des historiques de longueur 3 ou 4). Les

termes de l'équation précédente pour un modèle 3-grammes se résument alors comme suit :

$$\mathbb{P}(W_k) = \mathbb{P}(w_1) \times \mathbb{P}(w_2|w_1) \times \prod_{3}^{k} \mathbb{P}(w_i|w_{i-2}\ w_{i-1}) \tag{5.5}$$

L'apprentissage d'un modèle de langage n-gramme consiste à estimer, à partir d'un ou plusieurs corpus d'apprentissage normalisés, les probabilités n-gramme du modèle de langage. Les corpus, fournis dans le cadre de la campagne ESTER 2, sont utilisés comme corpus d'apprentissage pour le système du LIUM. Ces corpus sont augmentés notamment par le corpus *French Gigaword Corpus* [4], qui est composé d'un un très grand nombre de dépêches AFP (Agence France-Presse) et et APW (*Associated Press Worldstream*), et le corpus Web qui est composé des articles collectés à partir du Web tels que les articles du site Afrik.com et le site l'Humanité.

Pour chaque corpus, le nombre de séquence de 1 à n mots (n = 3 et n = 4) est compté et stocké. La probabilité d'apparition d'un mot est estimée par le critère de maximum de vraisemblance. Par exemple, pour le modèle 3-grammes et pour un mot w précédé des mots, cela donne :

$$\mathbb{P}(w|w_i w_j) = \frac{c(w_i w_j w)}{c(w_i w_j)} \tag{5.6}$$

La technique de lissage *modified Kneser-Ney* (Chen et Goodman 1996) est utilisée afin d'attribuer une probabilité non nulle aux événements non vu dans les corpus d'apprentissage. Aucun *cut-off* n'a été appliqué, c'est-à-dire que tous les n-grammes observés dans les corpus d'apprentissage, même une seule fois, sont pris en compte. les deux modèles de langage sont des modèles d'ordre 3 (3-grammes) et 4 (4-grammes), estimés avec la librairie SRILM *toolkit* (Stolcke 2002). Ainsi, un modèle 3-grammes et un modèle 4-grammes sont créés pour chaque corpus d'apprentissage. Ces modèles sont ensuite fusionnés par interpolation linéaire. Les coefficients d'interpolation ont été optimisés sur le corpus développement ESTER 2 (67 568 mots). Les modèles obtenus sont composés de 122 984 1-grammes, 29 042 901 2-grammes, 162 038 928 3-grammes et 376 037 558 4-grammes.

Modèles acoustiques

Les modèles acoustiques employés par le système de RAP du LIUM sont, comme la plupart des systèmes de RAP, basés sur les Modèles de Markov Cachés (HMM) (Jelinek 1976, Rabiner 1989). Ces modèles sont des automates stochastiques dont le but est de calculer la probabilité d'émission d'une séquence de vecteurs de paramètres du signal de parole. Cela consiste à estimer les probabilités de transition d'un état q_i vers un autre état (ou lui-même), soit $\mathbb{P}(q_j \mid q_i) = a_{ij}$, et les probabilités d'émission des observations dans chaque état, soit $b_j(x_n) = \mathbb{P}(x_n \mid q_j)$. Les

4. Catalogue LDC n° LDC2009T28

modèles HMM sont d'ordre 1, c'est-à-dire que la probabilité de passer dans l'état suivant dépend uniquement de l'état courant. Les unités de base modélisées appartiennent à un jeu de 35 phonèmes du français et de 5 types de *filler* (silence, musique, bruit, inspiration, "euh" prolongé). La modélisation des phonèmes est de type tri-phones, c'est-à-dire que la réalisation d'un phonème prend en compte les phonèmes qui l'entourent à gauche et à droite. La modélisation d'un mot s'effectue à partir de la concaténation des modèles de phonèmes qui composent ce mot.

Le système de RAP du LIUM utilise différents modèles acoustiques spécialisés en fonction du sexe du locuteur (homme/femme) et de la bande passante (téléphone/studio). Ces modèles ont été appris à partir d'un corpus d'apprentissage composé de 240 heures d'enregistrements audio transcrits manuellement (3,3 millions de mots), dont le corpus d'apprentissage ESTER 2. Les modèles obtenus sont composés soit de 6 500 états partagés, soit de 7 500 états partagés.

Dictionnaire de phonétisation

Le dictionnaire phonétisé consiste en quelques dizaines de milliers de mots sélectionnés à partir des corpus d'apprentissage. Ce vocabulaire doit garantir une couverture satisfaisante de l'ensemble des mots relatifs au domaine en question. Cela joue un rôle fondamental pour la qualité des transcriptions car si un mot n'est pas présent dans le vocabulaire, il ne pourra pas apparaître dans les transcriptions. Une ou plusieurs phonétisations sont associées à chaque entrée du vocabulaire sous la forme de phonèmes. Par exemple, « nn an tt » pour « nantes ». Cette phonétisation est réalisée soit manuellement, soit automatiquement en faisant appel à des outils de phonétisation automatique tel que LIA_PHON (Béchet 2001). La taille du vocabulaire dans le système de RAP du LIUM est fixée à 122 984 mots (227 000 entrées dans le dictionnaire phonétisé).

Processus de transcription

Une fois les ressources constituées (modèles acoustiques, modèle de langage et dictionnaire phonétisé), elles sont ensuite combinées pour générer la séquence de mots la plus probable. Le système de RAP du LIUM utilise comme base le décodeur CMU *Sphinx*[5] (Huang et al. 1993, Lee et al. 1990), diffusé sous licence libre. Différentes adaptations ont été effectuées à ce décodeur afin d'améliorer ces performances. La figure 5.2 présente l'architecture générale du système de transcription automatique LIUM. Un descriptif plus détaillé de ce système est présenté en détail dans Estève (2009), Dufour (2010).

La transcription est obtenue en enchaînant 5 passes de décodage.

5. http ://cmusphinx.sourceforge.net/

La première passe permet d'obtenir une hypothèse qui est utilisée pour adapter les modèles acoustiques aux locuteurs.

La passe 2 génère un graphe qui servira à guider le décodage de la passe 3. La réduction de l'espace de recherche au graphe obtenu à l'issue de la passe 2 permet d'utiliser les triphones inter-mots des modèles acoustiques. L'utilisation des contextes inter-mots permet un décodage plus précis que celui des 2 premières passes, là où seuls les triphones intra-mots sont exploités. Mais cette méthode est plus gourmande en ressources (en temps processeur comme en occupation mémoire), d'où la nécessité de réduire l'espace de recherche à un graphe prédéfini.

La passe 3 crée ainsi un graphe qui est réévalué à l'aide d'un modèle de langage quadrigramme dans la passe 4 (dans les 3 passes précédentes, le modèle de langage est un modèle trigramme).

Finalement, la cinquième et dernière passe transforme le graphe généré en fin de passe 4 en un réseau de confusion duquel la transcription finale est extraite.

De cette description technique, il faut retenir le fait que deux modèles de langage sont utilisés : un modèle trigramme (dans les 3 premières passes) et un modèle quadrigramme (dans les deux dernières passes). Ces modèles de langage partagent le même vocabulaire, qui est identique dans toutes les passes de la transcription.

Dans la suite de ce chapitre, nous réfèrencerons ce système par SRAP_SANS_REN.

5.2.2 Système de reconnaissance des entités nommées LIA _NE

Le système LIA _NE (Béchet et Charton 2010) utilisé dans ce travail a été développé par le Laboratoire Informatique d'Avignon (LIA). Ce système repose sur une méthode orientée données permettant de déterminer conjointement les frontières et les catégories des entités nommées.

Le système LIA _NE utilise, dans un premier temps, un modèle génératif à base de Modèles de Markov Cachés (HMM) afin d'annoter le texte en parties du discours. Il applique ensuite un modèle discriminatif à base de CRF pour annoter les entités nommées. Les résultats du modèle HMM permettent au modèle CRF d'éliminer un certain nombre d'ambiguïtés et de construire des règles d'une portée plus large en se basant sur les traits contextuels des mots et sur leurs parties du discours.

Le CRF a été appris sur le corpus d'apprentissage de la campagne d'évaluation ESTER 1 (Galliano et al. 2005). Ce corpus – qui représente une sous-partie du corpus d'apprentissage ESTER 2 (cf. table 5.2) – est constitué d'une centaine d'heures d'émissions radiophoniques francophones manuellement transcrites et annotées en entités nommées. La figure 5.3 montre les résultats de la REN pour quelques phrases extraites du corpus ESTER 2.

Le choix du système LIA _NE est motivé par les résultats qu'il a obtenu

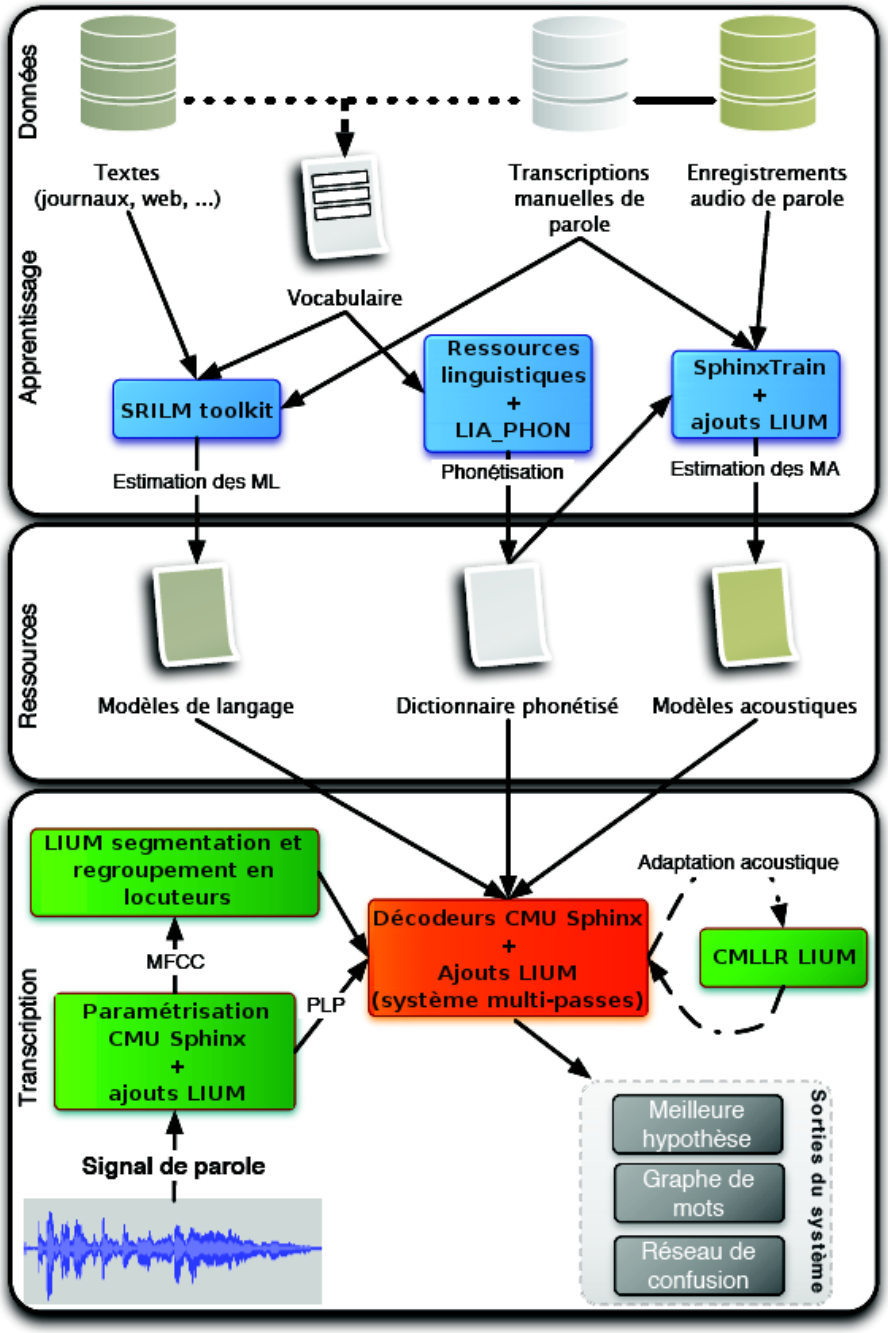

FIGURE 5.2 – *Architecture générale du système de RAP du LIUM (Estève 2009)*

(a) le journal **[pers** Denis Astagneau **]** bonsoir **[pers** Denis **]**
(b) son **[fonc** directeur général **]** **[pers** Philibeaux Dillon **]** démissionne
(c) sur place pour **[org** France Inter **]** **[pers** Frédéric Barrère **]**
(d) depuis **[time** vingt cinq ans **]** dans le **[loc** cinquième arrondissement **]** **[pers** Benoît Collombat **]**
(e) autour de **[amount** moins trois **]** à **[amount** moins cinq degrés **]** en général
(f) il est **[time** vingt heures **]** à **[loc** Johannesburg **]**
(g) **[amount** deux milliards d' euros **]** ont été mobilisés

FIGURE 5.3 – *Exemple de texte annoté par* LIA _NE *au format d'ESTER 2.*

lors de la campagne ESTER 2 (premier système si on omet les 2 systèmes non libres développés par des entreprises privées). Il a obtenu un SER de 23,90 % sur des transcriptions manuelles et un SER de 51,60 % sur les transcriptions automatiques (17,83 % de WER). De plus, la majuscule peut être intégrée en tant que descripteur dans le modèle CRF afin d'améliorer les performances du système lorsqu'il s'agit d'un texte bien formé.

5.2.3 Couplage de la REN au processus de transcription

Contrairement aux approches traditionnelles séparant la tâche de transcription de celle de REN, notre objectif est d'intégrer la tâche de REN au sein du processus de transcription. De cette manière, le système de transcription devient un système dédié à la tâche de REN et produit en sortie des transcriptions étiquetées en entités nommées.

Le problème revient donc à trouver la séquence de mots étiquetés en entités nommées la plus probable $(\widehat{W}, \widehat{E}) = (w_1, e_1), (w_2, e_2), (w_3, e_3), ..., (w_k, e_k)$, étant donnée la séquence d'observations acoustiques $(X = x_1, x_2, x_3, ..., x_p)$:

$$(\widehat{W}, \widehat{E}) = \arg\max_{w} \ \mathbb{P}(W, E|X) \tag{5.7}$$

Pour parvenir à cet objectif, notre démarche consiste à nous appuyer sur les ressources mobilisées par le système de transcription pour les enrichir et ainsi construire notre système de REN.

Les deux ressources exploitées sont, d'une part, les mots du vocabulaire, et d'autre part, les modèles de langage construits à partir des corpus d'apprentissage du système de transcription. Ces corpus doivent avoir une couverture importante par rapport au domaine ciblé car si un mot n'apparaît pas dans le vocabulaire, il ne pourra pas apparaître dans les transcriptions.

Nous supposons également que les entités nommées qui apparaissent dans les transcriptions automatiques conservent les mêmes catégories et les mêmes frontières que celles qui leur sont associées dans les corpus d'apprentissage. Par exemple, si le mot « hollande » apparaît systématiquement avec l'étiquette « personne » dans les corpus d'apprentissage, alors il conservera cette même étiquette dans les transcriptions automatiques.

L'avantage majeur d'inclure la REN directement au sein du système de transcription est d'éviter le post-traitement des transcriptions automatiques dont certains indices importants pour la REN sont absents tels que les majuscules et la ponctuation. Ces indices sont en revanche exploitables au niveau du système de transcription. En effet, les corpus d'apprentissage utilisés pour créer les modèles de langage sont généralement composés d'une grande quantité de dépêches et d'articles journalistiques et d'une quantité relativement faible de transcriptions manuelles d'émissions radiophoniques. La plupart des textes composant ces corpus est ainsi bien segmentée et les majuscules y sont présentes.

La méthode proposée consiste alors à exploiter les connaissances mobilisées pour la tâche de transcription en intégrant les informations relatives aux entités nommées dans les modèles de langage et dans le vocabulaire.

Cette méthode se décompose en quatre étapes : i) annotation des corpus d'apprentissage en entités nommées, ii) annotation du vocabulaire du système de transcription, iii) adaptation du dictionnaire phonétisé et iv) adaptation des modèles de langage.

Le système qui en résulte sera nommé SRAP_REN dans la suite de l'article.

Annotation des corpus d'apprentissage

Afin d'annoter en entités nommées les corpus d'apprentissage exploités pour créer les modèles de langage du système de transcription, nous avons utilisé le système LIA _NE. Nous avons commencé par annoter les corpus ayant servi à l'apprentissage des modèles de langage du système SRAP_SANS_REN. Pour ce faire, nous utilisons le corpus d'apprentissage de la campagne ESTER 2 augmenté de différentes ressources dont des articles du journal Le Monde, des dépêches de l'Agence France-Presse (AFP) et de l'*Associated Press Worldstream* (APW) ainsi que d'articles extraits de sites web notamment d'Afrik.com et de l'Humanité. La table 5.4 présente les caractéristiques des différents corpus (ceux-ci sont constitués de texte bien formé dont les phrases sont segmentées). Le corpus d'apprentissage a une taille de 1 082 000 742 mots dont 1 956 249 mots distincts. Dans ces corpus, la majuscule a été exploitée comme descripteur afin d'augmenter les performances du système LIA _NE et nous nous limitons aux 7 catégories principales d'entités nommées (personne, lieu, organisation, fonction, production humaine, montant et date et heure).

Après l'annotation des corpus d'apprentissage, nous avons transformé les annotation obtenues pour les mettre au format BI. L'encodage BI permet de préciser la catégorie ainsi que les frontières des entités nommées en indiquant, pour chaque mot appartenant à une entité nommée, s'il correspond au début (B) ou à l'intérieur de l'entité (I). Les mots n'appartenant à aucune entité nommée ne sont pas annotés. La figure 5.4 présente un exemple de texte annoté en entités nommées en utilisant l'encodage BI.

Corpus	Période	Nombre de mots
Corpus AFP	1994-2006	488 929 004
Corpus APW	1994-2006	173 598 873
Corpus LE MONDE	1994-2004	335 446 061
Corpus AFRIK	2007	6 319 708
Corpus L'HUMANITÉ	1990-2007	63 624 367
Corpus WEB	2007	9 617 468
Corpus ESTER 1	2007	3 249 228
Corpus d'apprentissage ESTER 2	1999-2004	1 216 033

TABLE 5.4 – *Corpus d'apprentissage pour la construction des modèles de langage.*

(a) le journal Denis**-pers-b** Astagneau**-pers-i** bonsoir Denis**-pers-b**
(b) son directeur**-fonc-b** général**-fonc-i** Philibeaux**-pers-b** Dillon**-pers-i** démissionne
(c) sur place pour France**-org-b** Inter**-org-i** Frédéric**-pers-b** Barrère**-pers-i**
(d) depuis vingt cinq**-time-b** ans**-time-i** dans le cinquième**-loc-b** arrondissement**-loc-i** Benoît**-pers-b** Collombat**-pers-i**
(e) autour de moins**-amount-b** trois**-amount-i** à moins**-amount-b** cinq**-amount-i** degrés**-amount-i** en général
(f) il est vingt**-time-b** heures**-time-i** à Johannesburg**-loc-b**
(g) deux**-amount-b** milliards**-amount-i** d'**-amount-i** euros**-amount-i** ont été mobilisés

FIGURE 5.4 – *Exemple de texte annoté au format BI.*

Annotation du vocabulaire du système de transcription

Après avoir réalisé l'annotation des corpus d'apprentissage, nous avons ajouté à chacun des mots du vocabulaire du système SRAP_SANS_REN la ou les étiquettes qui lui étaient associées dans le corpus annoté (cela inclut la catégorie de l'entité nommée ainsi que la position du mot dans l'entité nommée).Par exemple, les étiquettes « washington**-loc-b** », « washington**-org-b** », « washington**-org-i** », « washington**-pers-b** » et « washington**-pers-i** » sont associées au mot « washington ». Cela multiplie par un facteur quatre la taille du vocabulaire du système du SRAP_SANS_REN (503 192 mots par rapport à 122 981 mots initialement).

Adaptation du dictionnaire phonétisé

L'adaptation du dictionnaire phonétisé consiste à associer une ou plusieurs phonétisations aux mots étiquetés en entités nommées. L'ajout des étiquettes n'ayant aucun effet sur la phonétisation, les mots étiquetés conservent alors les mêmes phonétisations que celles attribuées aux mots sans étiquette dans le dictionnaire du système SRAP_SANS_REN. Par exemple, la phonétisation associée à « nantes**-loc-b** » et à « nantes**-org-b** » est « nn an tt ».

Nous supposons ici que la prononciation d'un mot n'est pas dépendante de la catégorie de l'entité nommée ; cette hypothèse est généralement vérifiée. On notera cependant que certains prénoms correspondant à des noms de ville peuvent avoir une prononciation différente de celle de

la ville. C'est le cas, par exemple, pour Paris Hilton ou encore pour Milan Milutinovi (président de la Serbie).

Adaptation des modèles de langage

Pour permettre au système de transcription de générer des sorties annotées en entités nommées, nous avons réappris les modèles de langage pour qu'ils prennent également en compte les étiquettes des entités nommées. Le rôle du modèle de langage est ainsi de contraindre le système de transcription à générer des transcriptions syntaxiquement correctes et annotées en entités nommées. Aucune coupe (*cut-off*) n'a été appliquée, c'est-à-dire que tous les n-grammes observés dans les corpus d'apprentissage sont pris en compte (même ceux n'apparaissant qu'une seule fois). Les deux modèles de langage exploités dans le système de transcription sont des modèles d'ordre 3 (trigramme) et 4 (quadrigramme), estimés avec l'outil SRILM (Stolcke 2002).

Un modèle trigramme et un modèle quadrigramme sont tout d'abord créés pour chaque corpus d'apprentissage. Ces modèles sont ensuite fusionnés par interpolation linéaire [6].

5.3 Expérimentations et résultats

Nous avons mené différents types d'expérimentations afin d'optimiser et de tester notre approche.

Comme nous l'avons vu dans la section précédente, la taille du vocabulaire du système SRAP_REN a considérablement augmenté. Dans un premier temps, nous avons mené une étude concernant l'impact de la taille de ce vocabulaire sur le WER et le SER du système. Les résultats ainsi obtenus nous ont permis de déterminer la taille optimale de ce vocabulaire afin de maximiser les performances du système SRAP_REN. Dans un second temps, nous avons évalué l'approche proposée, en termes de qualité de transcription et en termes de qualité de la REN sur les données de test. Nous présentons ensuite une analyse détaillée des erreurs commises par le système qui va permettre de mettre en évidence les qualités et les faiblesses de l'approche proposée. En dernier lieu, nous présentons une méthode afin d'améliorer la qualité de transcription des noms de personne.

6. La constitution des modèles globaux trigramme et quadrigramme est réalisée par interpolation linéaire de modèles individuels appris sur chaque corpus. Les coefficients d'interpolation ont été optimisés sur le corpus de développement *Dev1* afin de maximiser la perplexité du modèle global sur ce corpus.

5.3.1 Détermination de la taille optimale du vocabulaire

Constitution des différents ensembles de vocabulaire

Dans un premier temps, nous avons associé aux mots du vocabulaire du système SRAP_SANS_REN, toutes les étiquettes assignées dans les corpus annotés. Comme aucune coupe n'a été effectuée dans les modèles de langage, tout mot annoté apparaissant au moins une fois dans les corpus d'apprentissage est inclus dans le vocabulaire. Cela augmente la taille du vocabulaire de 122 981 à 503 192 mots. Il en résulte que certains mots sont annotés de manière erronée en raison des erreurs commises par le système LIA _NE lors de l'étiquetage des corpus d'apprentissage. Ces erreurs pourront alors se retrouver dans la sortie finale du système de transcription.

Afin de filtrer les étiquettes erronées ou rares (ce qui permettra également de diminuer la taille du vocabulaire), nous nous sommes appuyés sur l'hypothèse suivante : les mots incorrectement étiquetés apparaissent avec une fréquence beaucoup moins élevée que les mêmes mots correctement étiquetés. Par exemple, le mot « footballistique » apparaît 88 fois sans étiquette et une seule fois avec l'étiquette personne (« footballistique**-pers-b** »). Dans ce cas, nous considérons que le mot « footballistique**-pers-b** » a été incorrectement étiqueté.

La méthode utilisée pour réaliser le filtrage est inspirée de celle proposée par Allauzen et Gauvain (2004) dans le cadre de la détermination du vocabulaire d'un système de transcription. Elle consiste à :

1. entraîner un modèle unigramme pour chaque corpus d'apprentissage utilisé pour créer les modèles de langage, en utilisant le vocabulaire étiqueté (503 192 mots) ;
2. calculer les coefficients d'interpolation entre ces modèles unigrammes sur le corpus de développement *Dev1* ;
3. construire le modèle de langage global unigramme avec les coefficients d'interpolation calculés à l'étape 2 ;
4. extraire les N mots étiquetés les plus probables du modèle unigramme en fonction d'un seuil choisi.

Notre but étant d'étudier l'impact du vocabulaire sur le WER et le SER du système SRAP_REN, nous avons constitué 12 ensembles de mots. Chaque ensemble est construit en faisant varier le seuil relatif aux probabilités du modèle unigramme. Les seuils ont été choisis afin de déterminer des ensembles de mots (constituant ainsi un vocabulaire) dont la taille augmente de manière uniforme.

Il est à noter que, dans tous les cas, le vocabulaire sélectionné couvre l'ensemble des mots qui apparaissent dans le vocabulaire du système SRAP_SANS_REN (nous ajoutons ainsi, dans ces ensembles obtenus, les mots qui appartiennent au vocabulaire du système SRAP_SANS_REN s'ils n'y figuraient pas déjà).

Étude de l'influence de la taille du vocabulaire sur les performances du système SRAP_REN

Nous avons commencé par adapter les modèles de langage et le dictionnaire phonétisé, en fonction de chaque vocabulaire sélectionné. Nous obtenons ainsi 12 systèmes de transcription avec des vocabulaires différents. Ces 12 systèmes vont nous permettre d'évaluer l'influence du choix du vocabulaire sur la qualité de la transcription (WER) ainsi que sur la qualité de la REN (SER et F-mesure). Dans ces expérimentations, nous utilisons le corpus de développement *Dev2*.

Étude relative à la qualité de la transcription

Nous évaluons ici l'influence de la taille du vocabulaire du système de transcription sur la qualité des transcriptions. Pour les 12 systèmes de transcription créés (chacun ayant un vocabulaire différent), nous calculons le WER des transcriptions obtenues à partir des enregistrements sonores du corpus de développement *Dev2* (afin de calculer le WER, nous avons tout d'abord supprimé les étiquettes liées aux entités nommées dans les transcriptions).

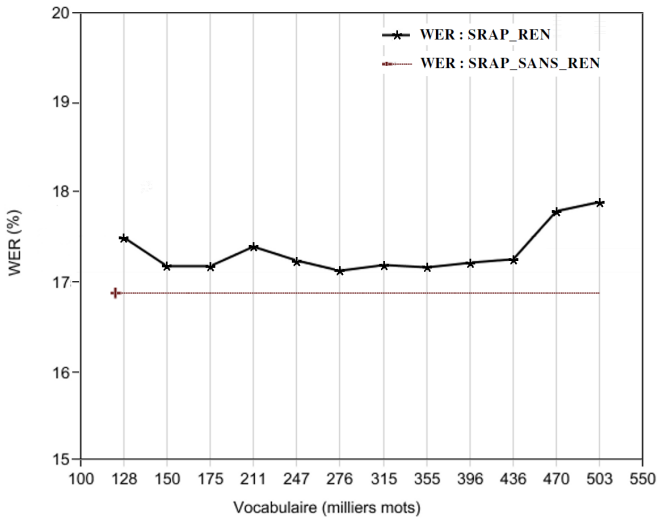

FIGURE 5.5 – *Influence sur le WER du choix du vocabulaire annoté du système de transcription (corpus de développement* Dev2*).*

La figure 5.5 montre la variation du WER en fonction de la taille N de chaque vocabulaire annoté sélectionné. Chaque point de la première courbe correspond ainsi à un système SRAP_REN avec un des vocabulaires annotés sélectionné. Il est à noter que le système SRAP_SANS_REN affiche un WER de 16,88 % sur le corpus *Dev2*, ce qui est représenté par la ligne droite sur la figure. Nous pouvons observer que, sans aucun filtrage du vocabulaire étiqueté ($N = 503\,192$ mots), l'intégration de la REN entraîne une augmentation du WER d'environ 1 %. Cela est principalement dû à l'augmentation de la taille de l'espace de recherche. Nous remarquons ensuite que la diminution de la taille du vocabulaire permet le plus souvent une amélioration du WER pour les différents N sélectionnés. Le

WER le plus proche de celui du système SRAP_SANS_REN est obtenu en utilisant un vocabulaire comportant $N = 276\,757$ mots (WER de 17,11 %, soit 0,23 % d'augmentation). Nous voyons donc que l'intégration de la REN au sein du système de transcription a un faible impact sur la qualité des transcriptions.

Étude relative à la qualité de la REN

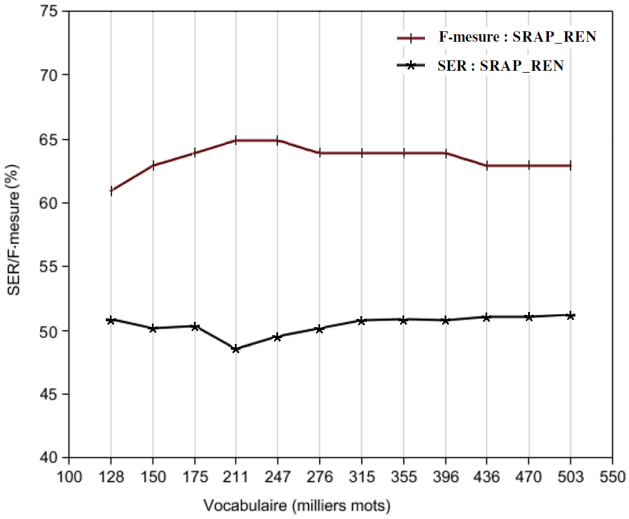

FIGURE 5.6 – *Influence, sur la qualité de la REN, du choix du vocabulaire annoté du système de transcription (corpus de développement* Dev2*).*

À partir des transcriptions annotées obtenues par chacun des 12 systèmes de transcription, nous avons calculé le SER et la F-mesure.

La figure 5.6 montre les différents scores obtenus. En utilisant l'ensemble du vocabulaire étiqueté ($N = 503\,192$ mots), le système affiche un SER de 51,23 % et une F-mesure de 63 %. La diminution de la taille du vocabulaire permet de constater une amélioration progressive du SER et de la F-mesure jusqu'à une certaine valeur de N, puis ensuite à une dégradation de ces mêmes indicateurs. L'amélioration s'explique par le fait que le filtrage de certaines étiquettes erronées permet de contrôler l'étiquetage de certains mots et donc d'éviter la reproduction de certaines erreurs commises par le système LIA _NE lors de l'annotation des corpus d'apprentissage. Par exemple, si le mot « DSK » apparaît dans le lexique comme étant une personne (sous la forme « DSK**-pers-b** »), il sera toujours étiqueté dans les transcriptions comme étant une personne et ce quel que soit son contexte ou sa position. Par contre, il existe une taille N du vocabulaire à partir de laquelle le nombre de mots annotés comme étant des entités nommées devient trop faible et dégrade la qualité des résultats.

Les meilleurs résultats sont obtenus en utilisant un vocabulaire de taille $N = 211\,576$ mots. Pour cette valeur, le SER est égal à 48,56 % et la F-mesure à 65 %.

Choix de la taille du vocabulaire

Nous venons d'étudier l'impact de la taille du vocabulaire sur la qualité de la transcription et de la REN. Nous avons vu que cette taille influence peu la qualité de la transcription (moins de 1 % du WER). En revanche, concernant la REN, nous avons clairement détecté un seuil qui correspond à des meilleurs résultats en termes de SER et de F-mesure. Pour fixer la taille du vocabulaire, nous privilégions ce critère qui nous permet d'obtenir un vocabulaire de taille $N = 211\,576$ mots. Dans ce contexte, le système affiche un WER de 17,38 %. Le système SRAP_REN que nous évaluerons dans la prochaine section disposera, en conséquence, de modèles de langages composés de 211 576 unigrammes, 30 010 819 bigrammes, 16 3047 041 trigrammes et 377 272 219 quadrigrammes.

5.3.2 Évaluation des performances du système SRAP_REN sur le corpus de test

Nous évaluons ici l'influence de l'intégration de la REN au sein du système de RAP sur la qualité de transcription et sur la REN, en utilisant le corpus de test.

Le corpus de test est constitué de 51 836 mots. Parmi ceux-ci, 6 158 mots appartiennent à des entités nommées (soit 8,41 % des mots de ce corpus), ce qui représente un total de 3 289 entités nommées.

Influence de l'intégration de la REN sur la qualité de transcription

Pour évaluer l'influence de l'intégration de la REN sur la qualité de transcription, nous avons comparé deux versions du système du LIUM. D'une part, nous avons utilisé le système SRAP_SANS_REN pour décoder les données du corpus de test. D'autre part, nous avons décodé ces mêmes données avec le système SRAP_REN puis nous avons supprimé les étiquettes d'entités nommées des transcriptions obtenues.

Système de RAP	WER (%)	WER entités nommées (%)
SRAP_SANS_REN	20,23	21,32
SRAP_REN	21,17	23,90

TABLE 5.5 – *Taux d'erreur sur les mots (WER) avant et après l'intégration de la REN dans le processus de RAP (corpus de test).*

La table 5.5 donne le WER obtenu pour les deux versions du décodeur.

Nous remarquons tout d'abord que le WER sur les mots correspondant à des entités nommées est plus élevé que le WER sur tous les mots, pour les deux versions du décodeur. Cette augmentation du WER peut provenir d'une part d'une faible représentation de la séquence de mots constituant les entités nommées dans les corpus d'apprentissage des modèles de langage. D'autre part, les noms de personnes, les noms d'organisation et plus

généralement les mots d'origine étrangère ont fréquemment des phonétisations erronées générant des erreurs lors du décodage. Nous constatons également que l'intégration de la REN au sein du décodeur entraîne une augmentation de 0,94 % pour le WER sur tous les mots des transcriptions et de 2,58 % pour le WER uniquement sur les mots correspondant à des entités nommées. Cela est dû à l'intégration des marqueurs d'entités nommées sur les mots de vocabulaire qui a pour effet d'augmenter le nombre de mots du vocabulaire mais aussi de modifier les probabilités des modèles de langage, ce qui les rend moins performants en termes de WER.

Catégorie	**Nombre de mots**	**WER (%) SRAP_SANS_REN**	**WER (%) SRAP_REN**
Personne	1 377	45,38	46,11
Organisation	1 480	19,12	21,48
Lieu	1 281	16,93	19,98
Fonction	572	11,88	12,23
Production humaine	82	13,41	14,63
Montant	943	8,03	8,51
Date et heure	423	12,19	15,80

TABLE 5.6 – *Nombre de mots et WER par catégorie d'entité nommée, avant et après l'intégration de la REN dans le processus de RAP (corpus de test).*

La table 5.6 présente le nombre de mots correspondant à des entités nommées, pour chaque catégorie d'entités nommées, ainsi que les performances obtenues par les deux versions du décodeur, pour chacune de ces catégories. Nous pouvons tout d'abord constater que la qualité de transcription se trouve légèrement dégradée pour les différentes catégories. Les résultats obtenus par les deux systèmes montrent également que les noms de personnes sont les plus difficiles à transcrire, avec un WER élevé d'environ 46 %. En effet, en dehors des journalistes qui sont présents à l'antenne quotidiennement, les noms des personnes sont généralement cités pour un fait d'actualité ponctuel. Ces noms correspondent fréquemment aux invités de l'émission ou aux personnes sujets de faits traités. La sélection d'un nombre limité d'entités de cette classe à inclure dans le vocabulaire du système de RAP est donc une tâche délicate (Dufour et al. 2012). En ce qui concerne les noms d'organisations, de lieux, de fonctions et de productions humaines, leur transcription s'avère satisfaisante. Ces entités évoluent moins rapidement dans le temps que les entités de type personne. La transcription des montants ainsi que des dates et heures est également satisfaisante. Cela est dû au fait que ces entités sont généralement composées de noms communs qui sont présents dans le vocabulaire du système de RAP.

Influence de l'intégration de la REN sur la qualité de la REN

Afin d'évaluer la contribution de l'intégration de la REN directement dans le processus de RAP, nous avons comparé deux versions du décodeur. D'une part, nous avons utilisé SRAP_SANS_REN pour décoder les données sonores du corpus de test puis nous avons ensuite utilisé le sys-

tème LIA _NE pour étiqueter en entités nommées les transcriptions résultantes. D'autre part, nous avons décodé les données de test avec le système SRAP_REN. Nous avons également utilisé SRAP_REN uniquement pour l'étape de transcription suivi de LIA _NE pour la tâche de REN.

Système de RAP	**SER (%)**	**F (%)**	**P (%)**	**R (%)**
SRAP_SANS_REN puis LIA _NE	54,01	58,00	64,50	52,76
SRAP_REN	49,22	63,00	70,32	57,59
SRAP_REN puis LIA _NE	54,12	58,00	66,44	50,84

TABLE 5.7 – *Performances de la REN avant et après l'intégration de la REN dans le processus de RAP (corpus de test) en termes de SER, F-mesure (F), Précision (P) et Rappel (R).*

Les deux premières lignes de table 5.7 montrent les résultats obtenus pour les deux versions du décodeur. Malgré la légère dégradation de la qualité de la transcription (et ce de manière plus marquée pour les entités nommées que pour les autres mots du vocabulaire), nous constatons que l'intégration de la REN au sein du système de RAP permet une amélioration importante du SER et de la F-mesure : une diminution de 4,79 % du SER et une augmentation de 5 % de F-mesure. En effet, nous cherchons une meilleure reconnaissance des entités nommées et non pas une meilleure transcription. Nous attribuons le gain obtenu au fait qu'assigner des étiquettes d'entités nommées au niveau du système de RAP permet de contrôler l'étiquetage des mots composant les entités nommées, notamment quand il y a des erreurs de transcription dans le contexte de ceux-ci. La dernière ligne de table 5.7 montre les résultats obtenus en utilisant SRAP_REN uniquement pour l'étape de transcription puis en utilisant LIA _NE pour la REN. Comme le système de reconnaissance utilisé a un moins bon WER que SRAP_SANS_REN, nous constatons que le taux de rappel relatif aux mots composant les entités est légèrement moins bon que pour l'exploitation de SRAP_SANS_REN et LIA _NE en cascade. Par contre, le taux de précision est légèrement meilleur. Le taux de SER est quant à lui stable. Il est à noter que la composition de ces deux derniers systèmes conduit aussi à de moins bonnes performances pour la REN que celles obtenues par SRAP_REN.

Catégorie	**SER SRAP_SANS_REN puis LIA _NE (%)**	**SER SRAP_REN (%)**
Personne	69,90	64,76
Organisation	76,64	71,56
Lieu	60,19	54,46
Fonction	65,33	50,17
Production humaine	112,50	97,92
Montant	53,21	51,28
Date et heure	57,44	55,38

TABLE 5.8 – *SER par catégorie d'entité nommée avant et après l'intégration de la REN dans le processus de RAP (corpus de test).*

La table 5.8 présente le nombre d'entités nommées de chaque catégorie, ainsi que les performances de la REN pour chacune des catégories, avant

et après l'intégration de la REN dans le processus de RAP. Nous pouvons tout d'abord remarquer une diminution du SER pour les différentes catégories. Cette diminution est de l'ordre de 5 % pour les personnes, les organisations et les lieux, qui correspondent aux entités nommées les plus représentées dans le corpus de test (environ 72 % des entités nommées). En ce qui concerne les fonctions et les productions humaines (dont le nombre est assez faible par rapport aux autres catégories), la baisse du SER est plus importante : -15,16 % et -14,58 %, respectivement. Pour expliquer cette baisse, nous pouvons regarder le détail des nombres d'erreurs de chaque type qui sont utilisés pour le calcul du SER.

	SRAP_SANS_REN puis LIA _NE				**SRAP_REN**			
Catégorie	**I**	**S**	**ET**	**EF**	**I**	**S**	**ET**	**EF**
Personne	118	177	1	210	107	165	2	217
Organisation	68	267	94	59	85	239	77	75
Lieu	71	144	30	92	72	139	35	79
Fonction	6	123	0	94	17	75	2	77
Production humaine	4	15	0	0	2	12	1	3
Montant	17	32	11	10	13	30	11	10
Date et heure	64	84	10	72	63	87	10	67

TABLE 5.9 – *Nombre d'erreurs pour les diffèrents types d'erreurs par catégorie d'entité nommée avant et après l'intégration de la REN dans le processus de RAP (corpus de test) : insertion (I), suppression (S), erreur de type (ET) et erreur de frontière (EF).*

La table 5.9 montre les différents types d'erreurs comptabilisés dans le SER. En utilisant SRAP_REN, nous constatons une diminution des suppressions pour les personnes et les organisations, une diminution des insertions pour les personnes ainsi qu'une diminution des erreurs de frontière pour les lieux. En ce qui concerne les fonctions, la baisse du SER est due aux suppressions qui sont presque deux fois moins présentes (123 suppressions pour SRAP_SANS_REN puis LIA _NE contre 75 pour SRAP_REN) ainsi qu'à une diminution des erreurs de frontière. Pour les productions humaines, la baisse du SER concerne principalement les suppressions et les insertions. Il y a en revanche une légère augmentation du nombre d'erreurs de frontière mais les entités nommées de type production humaine sont les plus longues avec une moyenne de 3,4 mots par entité nommée d'où une plus grande difficulté à correctement identifier leurs frontières. Finalement, pour les montants et les dates et heures, la baisse du SER est de l'ordre de 2 %. Le nombre d'erreurs de type et de frontière reste constant pour les montants mais ces entités nommées sont également parmi les plus longues avec en moyenne 2,7 mots par entité nommée. En ce qui concerne les dates et heures, la baisse porte essentiellement sur les erreurs de frontière.

Influence de la prise en compte des majuscules

Les performances de notre approche dépendent directement de la qualité d'annotation des corpus ayant servi à l'apprentissage des modèles

de langage. Cette annotation étant réalisée sur du texte bien formé, cela permet d'exploiter des traits discriminants pour la REN qui ne sont pas disponibles lorsque le système de REN est confronté à des transcriptions automatiques. La majuscule figure parmi les traits les plus importants. En effet, la prise en compte de ce trait a un impact direct sur les performances du système SRAP_REN. L'annotation des corpus d'apprentissage, en utilisant le système LIA _NE sans l'exploitation des informations concernant la majuscule, entraîne une dégradation des performances du système SRAP_REN d'environ 16 % en termes de SER sur le corpus *Dev2* (de 51,23 % pour le vocabulaire composé de 503 192 mots à 67,34 % pour un vocabulaire composé de 539 466 mots). Cela montre que l'efficacité de LIA _NE intégrant la majuscule comme trait est bien meilleure que la version de LIA _NE n'exploitant pas cette information.

5.3.3 Analyse des erreurs

Afin d'avoir une vue plus fine des problèmes liés à notre approche, nous avons réalisé une typologie des erreurs commises par notre système. Notre corpus de test de référence comprend 6 158 mots appartenant à des entités nommées. Parmi ces mots, 2 537 sont soit mal retranscrits, soit pas du tout étiquetés comme entité nommée ou alors mal étiquetés (mauvais typage) par notre système.

Nous avons analysé manuellement un échantillon de 450 mots choisis aléatoirement parmi ces 2 537 mots afin d'étudier les faiblesses du système. Nous avons ainsi défini quatre classes d'erreurs que nous explicitons dans les paragraphes suivants.

Une première classe d'erreurs est liée aux transcriptions. Ces erreurs (qui touchent 45,11 % des mots que nous étudions) représentent un problème majeur pour la REN dans le cadre des transcriptions automatiques, quelle que soit la méthode utilisée pour la REN. Par exemple, si une entité nommée prononcée est absente du vocabulaire du système de RAP, elle est alors remplacée par un ou plusieurs mots proches acoustiquement. Il en résulte qu'elle est considérée comme étant mal reconnue lors du calcul du SER, ce qui dégrade ce dernier. L'augmentation du taux de WER associée à ces entités nommées, que nous avons notée lors de l'étude précédente, est inhérente aux modèles de langage et aux décodages. En effet, une même séquence de mots est représentée par plusieurs n-grammes, chacun d'eux se distinguant par leurs étiquettes d'entité nommée. La multiplication des n-grammes pour une même séquence de mots augmente alors l'espace de recherche et la probabilité d'un n-gramme avec ses étiquettes est plus faible que la probabilité de la séquence de mots correspondante (c'est-à-dire du n-gramme obtenu avec un modèle de langage sans entités nommées). Le choix du vocabulaire adapté au domaine à traiter reste un point délicat comme dans tous les systèmes de transcription de la parole et de REN.

Une deuxième classe d'erreurs correspond à la reproduction d'erreurs commises par LIA _NE. Ces erreurs (qui touchent 20,44 % des mots que nous étudions) sont inhérentes à notre type d'approche. En effet, nous

avons annoté les corpus d'apprentissage à l'aide d'un outil automatique qui, malgré ses bonnes performances, commet des erreurs d'annotation qui sont alors réitérées par notre système. Par exemple, pour l'entité « ministère**-org-b** de**-org-i** la**-org-i** défense**-org-i** française », le mot « française » n'est pas inclus dans l'entité. La correction de ce type d'erreur nécessite son traitement en amont, c'est-à-dire au niveau de l'annotation initiale des corpus d'apprentissage. Pour atteindre cet objectif, un traitement manuel est bien sûr à exclure vue la taille des corpus d'apprentissage mais un outil plus performant de REN pourrait être mis en œuvre.

Une troisième classe d'erreurs est relative à l'absence de mots liés à une entité nommée dans les corpus d'apprentissage (en excluant les problèmes inhérents à LIA _NE). Ces erreurs (qui concernent 12,66 % des mots que nous étudions) sont liés aux mots qui appartiennent aux lexiques du système de RAP mais qui ne sont pas apparus dans le corpus d'apprentissage comme faisant partie d'une entité nommée ; ce phénomène concerne souvent les noms communs. Par exemple, pour l'entité « ministère**-org-b** des**-org-i** pme**-org-i** de l'économie sociale et de l'artisanat », seule la partie « ministère**-org-b** des**-org-i** pme**-org-i** » apparaît dans les corpus d'apprentissage. La correction de ce type d'erreur nécessite une mise à jour du vocabulaire étiqueté ou un traitement *a posteriori*.

Enfin, la dernière classe d'erreurs touche les derniers mots restants (21,79 % des mots que nous étudions). Elles concernent principalement la mauvaise résolution par le système de RAP de l'ambiguïté liée à des étiquettes multiples pour un même mot. Par exemple, « paris**-loc-b** » et « paris**-org-b** ». Pour ce type d'erreur, l'annotation en amont et la prise en compte des étiquettes dans les modèles de language du système n'ont pas suffi à résoudre ces problèmes d'ambiguïté.

5.3.4 Prise en compte des entités nommées multi-mots dans le lexique

Plusieurs travaux se sont intéressés à l'amélioration de la qualité de transcription des entités nommées, ce qui permet une meilleure reconnaissance de ces dernières. Certaines approches tentent de corriger *a posteriori* les transcriptions automatiques en recherchant les mots manquants sur le WEB (Oger et al. 2008, Dufour et al. 2012). D'autres adaptent le vocabulaire et les modèles du système de RAP en rajoutant de nouveaux lexiques. Ces derniers sont sélectionnés à partir des documents couvrant la même période que le domaine ciblé (Favre et al. 2005, Martins et al. 2007).

Nous évaluons ici une méthode permettant d'améliorer la qualité de transcriptions. Nous nous focalisons ici sur les noms de personne puisqu'ils ont obtenu les résultats les plus faibles en termes de WER (environ 46 %). Nous avons remarqué que, dans plusieurs cas, ces noms sont mal transcrits bien que les mots qui les forment soient présents dans le vocabulaire du système de RAP. Les erreurs de transcription peuvent toucher une partie ou la totalité des éléments constituants les entités nommées.

Partant du fait que le système de RAP peut privilégier une séquence composée de peu de mots longs, nous avons créé des entités multi-mots de

type personne en assemblant par juxtaposition les éléments formant ces entités. Ces entités nommées sont considérées comme étant un seul mot composé, à la fois dans le vocabulaire et dans les modèles de langage du décodeur. Cela devrait augmenter la probabilité que toute l'entité nommée soit transcrite correctement, en éliminant les homophones et en compensant les phonétisations et les prononciations hasardeuses. Cela pourra être le cas, par exemple, pour l'entité « sarah ghibaudo » qui a été transcrite en « sarah kimono ».

Système de RAP	**WER noms de personne (%)**	**SER noms de personne (%)**
SRAP_REN sans EN multi-mots	46,11	64,76
SRAP_REN avec EN multi-mots	46,04	64,41

TABLE 5.10 – *Performances en termes de WER et de SER avant et après la prise en compte des enités nommées multi-mots dans le lexique (corpus de test).*

Nous avons rajouté environ 1 000 noms de personne sous la forme « *prénom_nom***-pers-b** » au vocabulaire (par exemple, « juan_antonio_samaranch**-pers-b** »). Ces entités ont été sélectionnées à partir des corpus d'apprentissage (tous les mots composant ces entités existent déjà dans le vocabulaire du système SRAP_SANS_REN). Nous avons ensuite recréé les modèles de langage et adapté le dictionnaire phonétisé. La table 5.10 donne, sur le corpus de test, le WER et le SER concernant les noms de personne, avant et après la prise en compte des entités nommées multi-mots dans le lexique. Nous remarquons une faible amélioration de la qualité de transcription des noms de personne, ce qui a engendré une faible amélioration du SER. En regardant plus en détail les entités nommées multi-mots transcrites, nous trouvons que cette méthode a permis de corriger la transcription de seulement cinq entités. Cela reste insuffisant par rapport au pourcentage des entités mal transcrites.

5.4 Conclusion

Dans ce travail, nous nous sommes intéressés à la tâche de reconnaissance des entités nommées en parole. La reconnaissance des entités nommées pour ce type de modalité pose un certain nombre de difficultés inhérentes à ses caractéristiques intrinsèques.

Nous avons proposé de détourner les fonctionnalités de base d'un système de transcription de la parole pour en faire un système de reconnaissance des entités nommées (ou plus exactement en faire un système de transcription de la parole dont les sorties sont annotées en entités nommées). Ainsi, en mobilisant, pour la tâche de reconnaissance des entités nommées, les connaissances exploitées pour la tâche de reconnaissance de la parole, nous assurons une plus grande synergie entre ces deux tâches.

L'évaluation de notre méthode indique une augmentation significative

de la qualité de la reconnaissance des entités nommées d'environ 5 % en termes de SER comme de F-mesure, et ce par rapport aux résultats obtenus avec l'un des meilleurs systèmes de reconnaissance des entités nommées en parole lorsque celui-ci est utilisé en aval du système de transcription. Cette augmentation de la qualité de la reconnaissance des entités nommées se fait au détriment de la qualité de la transcription (moins de 3 % de perte de WER pour les entités nommées), quoique notre objectif ne soit pas d'en améliorer la transcription mais bien la reconnaissance. En outre, cette interaction entre système de RAP et de REN n'est pas très coûteuse en termes d'ingénierie logicielle puisqu'elle ne fait qu'enrichir les connaissances mobilisées pour la tâche de RAP, notamment les mots du vocabulaire et les modèles de langage.

Les deux éléments clés de notre système de reconnaissance des entités nommées sont le vocabulaire et les modèles de langage. Un avantage indéniable de cette approche provient du fait que le vocabulaire est unique et commun à l'ensemble de notre système, ce qui évite de créer d'éventuels effets de bord entre la transcription et la détection des entités nommées. En revanche, un mot hors-vocabulaire ne pourra ni être transcrit, ni étiqueté en entité nommée avec ce type d'approche. Le choix du vocabulaire adapté au domaine à traiter reste un point délicat comme dans tous les systèmes de transcription de la parole et de REN.

Les modèles de langage d'ordre 3 et 4 intégrant les étiquettes d'entités nommées sont appris à partir de textes annotés automatiquement en entités nommées. Cette solution, peu coûteuse par rapport à une annotation manuelle, est fonctionnelle quoique imparfaite. Une des sources d'erreurs provient ainsi de la reproduction des erreurs les plus fréquentes de LIA _NE par notre système. Bien entendu, la méthode proposée nécessite de disposer d'un système de REN mais celui-ci n'a pas besoin d'être spécifique au traitement de l'oral. La majorité des corpus d'apprentissage pour les modèles de langage sont ainsi des journaux ou des dépêches avec des marqueurs forts d'entités nommées comme les ponctuations et les majuscules.

Conclusion générale

Ce travail de thèse a été consacré à la tâche de REN pour la modalité orale. Nous commençons par rappeler les principales contributions apportées par notre travail, avant de présenter quelques perspectives pour de futurs travaux de recherche qui s'inscrivent dans la continuité de cette thèse.

Contributions

Dans un premier temps, nous avons présenté un état de l'art détaillé de la REN. Nous avons commencé par introduire les différents concepts permettant de cerner la notion d'entité nommée. Nous avons ensuite décrit quelques systèmes caractéristiques de la REN et les différentes approches associées.

Dans les chapitres 2 et 3 nous avons mis l'accent sur deux thèmes de recherche qui sont au cœur de nouveaux enjeux de TALN en général, et de REN en particulier : la multimodalité et le multilinguisme.

Dans le chapitre 2, nous avons mis en évidence les différentes difficultés auxquelles la REN doit faire face en fonction de la modalité traitée. Nous avons présenté des systèmes de REN caractéristiques des textes issus de différentes modalités : écrit, parole, vidéo, manuscrit. L'analyse des différents résultats montre que la REN peut être considérée comme un problème bien traité pour du texte bien formé, en revanche, cette tâche pose encore de nombreuses difficultés pour d'autres modalités comme la parole ou les textes manuscrits.

Le chapitre 3 s'est intéressé à la portabilité entre les langues des systèmes de REN. Nous avons donc étudié les différentes approches utilisées pour l'adaptation des systèmes de REN, que ce soit les systèmes symboliques ou à base d'apprentissage. Dans ce contexte, une piste a été explorée consistant à étudier l'intérêt et le coût associé pour porter un système existant de REN vers une autre langue. Nous avons proposé une méthode peu coûteuse pour porter un système symbolique dédié au français vers l'anglais. Pour ce faire, nous avons d'une part traduit automatiquement l'ensemble des lexiques de mots déclencheurs au moyen d'un dictionnaire bilingue. D'autre part, nous avons manuellement modifié quelques règles de manière à respecter la syntaxe de la langue anglaise. L'évaluation du système adapté montre des résultats satisfaisants pour la reconnaissance des noms de personne et de lieu. En revanche, les résultats restent insuffisants pour les noms d'organisation et d'entreprise. Un traitement plus approfondi est nécessaire pour ces deux catégories.

Dans le chapitre 4 nous avons étudié les spécificités de la REN en sortie du système de RAP. Nous avons mesuré l'impact des différents phénomènes spécifiques à la parole sur la qualité de REN : l'absence de la majuscule, la présence des disfluences et la présence des erreurs de transcription. Pour ce faire, nous avons développé un système de REN dédié aux transcriptions de la parole en adoptant une taxonomie hiérarchique et compositionnelle. L'approche proposée consiste à définir trois niveaux d'annotation. Le premier niveau s'intéresse à l'annotation des catégories. Le second niveau s'occupe de l'annotation des composants. Le dernier niveau permet l'imbrication des annotations. Les résultats obtenus sur les données de la campagne d'évaluation ETAPE montrent l'efficacité de notre approche. Ces résultats montrent l'impact de la majuscule, ainsi que l'effet direct des erreurs de transcription sur la qualité de REN.

Dans le dernier chapitre, nous avons proposé une approche alternative aux approches traditionnelles sur la REN en parole. Nous avons proposé de contourner les fonctionnalités de base d'un système de transcription de la parole pour en faire un système de REN. Cela permet de détourner un certain nombre des difficultés inhérentes à la parole telles que l'absence de majuscule et de ponctuation. Notre approche consiste à exploiter les connaissances mobilisées pour la tâche de transcription en intégrant les informations relatives aux entités nommées dans les modèles de langage et dans le vocabulaire. Nous avons donc adapté les modèles de langage et le vocabulaire de manière à ce que le système de RAP produise des transcriptions annotées en entités nommées. L'évaluation de notre méthode indique une augmentation significative de la qualité de la reconnaissance des entités nommées d'environ 5 % en termes de SER comme de F-mesure, par rapport aux résultats obtenus avec l'un des meilleurs systèmes de REN en parole lorsque celui-ci est utilisé en aval de la tâche de transcription de la parole. Nous avons aussi proposé dans ce chapitre une analyse des erreurs de REN, ce qui permet de mieux cerner les limites de notre approche.

Perspectives

Nous donnons ici quelques perspectives qui s'inscrivent dans la continuité directe de notre travail de thèse :

Intégrer des connaissances extraites à partir du système de RAP dans le système de REN

Notre système de REN, présenté dans le chapitre 4, n'exploite que des traits "traditionnelles" pour reconnaitre les entités nommées (étiquettes POS, lexiques, etc). Nous avons montré dans le chapitre 5 qu'une association entre les processus de RAP et de REN permet une amélioration des performances de la qualité de REN. Il serait donc intéressant d'incorporer des traits concernant le processus de transcription dans le système de REN. Certains traits peuvent être extraits à partir des ressources mobilisées par le système de transcription telles que les corpus d'apprentissage et le vocabulaire. D'autres traits peuvent être extraits à partir de graphe

de mot et de réseau de confusion. Par exemple, les scores de confiance concernant la transcription du mot.

Améliorer la sélection du vocabulaire annoté pour le système de RAP intégrant la REN

Notre approche qui permet d'effectuer conjointement la transcription et la détection des entités nommées dépend directement de la qualité d'annotation des corpus ayant servi à l'apprentissage des modèles de langage et de la sélection du vocabulaire annoté. L'amélioration de la qualité d'annotation des corpus d'apprentissage s'avère difficile vu la taille de ces corpus. Nous avons proposée une méthode pour filtrer les mots mal étiquetés à partir du vocabulaire annoté. Il serait intéressant de corriger l'annotation de ces mots au lieu de les supprimer en exploitant des ressources externes, notamment le Web.

Améliorer la qualité de transcription des entités nommées

La qualité de REN dépend de la qualité de transcription. En effet, si l'entité nommée n'est pas bien transcrite, elle ne peut pas être bien reconnue. La transcription des entités nommées reste une tâche difficile qui suscite encore de nombreuses recherches. La méthode présentée dans la section 5.3.4 concernant la prise en compte des entités nommées multimots dans le lexique nécessite une étude plus élaborée. Il serait donc intéressant de rajouter un plus grand nombre d'entités nommées juxtaposées couvrant les différentes catégories. D'autres pistes peuvent être explorées comme la correction *a posteriori* des transcriptions automatiques en recherchant les mots manquants ou l'adaptation du vocabulaire et des modèles du système de RAP en rajoutant de nouveaux lexiques.

Vers une optimisation conjointe de la qualité de transcription et de la qualité de REN ...

En intégrant la REN au processus de RAP, nous avons dégradé la qualité de la transcription par rapport au système ne l'intégrant pas. L'intégration des informations relatives aux entités nommées dans les modèles de langage et dans le vocabulaire est réalisée implicitement par suffixation des étiquettes concernant les entités nommées. Cela a engendré une augmentation de la taille du vocabulaire de système de RAP, ce qui a généré un effet négatif sur la qualité de transcription. Une autre piste de recherche serait d'optimiser conjointement la qualité de transcription et la qualité de REN. Cela revient à trouver la séquence de mots étiquetée en entités nommées la plus probable ($\hat{W}$, $\hat{e} = (w, e)_1, (w, e)_2, (w, e)_3, ..., (w, e)_k$) étant donné la séquence d'observations acoustiques ($X = x_1, x_2, x_3, ..., x_p$) :

$$\begin{aligned} \widehat{W}, \widehat{E} &= arg \max_{w,e} \ \mathbb{P}((W, E)|X) \\ &= arg \max_{w,e} \ \mathbb{P}(W, E)\mathbb{P}(X|W, E) \\ &= arg \max_{w,e} \ \mathbb{P}(W)\mathbb{P}(E|W)\mathbb{P}(X|W, E) \end{aligned} \tag{5.8}$$

Le modèle de langage se charge de calculer la valeur de $\mathbb{P}(W)$. Le modèle acoustique se charge de calculer la valeur de $\mathbb{P}(X|W, E)$. Il reste

donc à intégrer dans le système de RAP un nouveau modèle permettant de calculer la valeur de $\mathbb{P}(E|W)$.

Annexes

A

Sommaire

A.1 Participation à la campagne d'évaluation Etape

La table A.1 reporte les performances des systèmes participants à la campagne d'évaluation Etape après l'étape d'adjudication. Le SER a été utilisé afin d'évaluer les performances des différents systèmes (Galibert et al. 2011). La version du système développé dans le cadre de notre participation à cette campagne (système **LL_NE**) ne permet pas l'annotation des composants. Nous nous sommes intéressés seulement aux catégories (premier et troisième niveaux d'annotation). Cela a affecté les résultats globaux de notre système par rapport aux autres systèmes participants.

	Manuel	**Rover**	WER=23	WER=24	WER=25	WER=30	WER=35
S1	36,38	57,17	59,04	64,44	61,88	61,59	76,09
S2	37,31	67,72	67,87	67,44	70,69	69,88	84,92
S3	37,99	63,22	66,91	63,52	68,57	68,1	79,9
S4	39,16	64,28	68,92	65,38	69,66	69,02	86,09
S5	44,46	69,51	73,58	71,93	73,61	74,61	85,62
S6	49,96	87,3	97,73	76,29	91,9	93,85	98,65
S7	85,56	97,88	100,48	94,2	98,79	98,27	100,71
LL_NE	62,44	75,62	78,74	76,61	79,64	82,25	90,23

Table A.1 – *Performances en termes de SER des systèmes participants à la campagne d'évaluation Etape*

A.2 Évaluation sur les transcriptions automatiques

Nous reportons ici les résultats détaillés obtenus par notre système de REN (chapitre 4) pour les différents taux d'erreur de mots (WER) : WER=24, WER=25, WER=30, WER=35 et rover.

A.2.1 WER = 24

Catégorie	P (%)	R (%)	F (%)	Composant	P (%)	R (%)	F (%)
amount	56,19	46,00	50,58	kind	75,471	29,39	42,30
pers.ind	59,56	50,80	54,84	extractor	50,00	9,09	15,38
pers.coll	51,07	39,44	44,51	qualifier	52,17	22,36	31,30
pers.other	0	0	0	title	64,86	33,80	44,44
time.date.abs	46,90	25,11	32,71	unit	80,05	67,33	73,14
time.date.rel	72,30	47,69	57,47	object	51,36	42,80	46,69
time.hour.abs	58,62	30,35	40,00	range-mark	86,11	47,69	61,38
time.hour.rel	77,64	35,10	48,35	day	65,62	65,62	65,62
loc.oro	75,00	33,33	46,15	week	81,81	64,28	72,00
loc.fac	33,33	6,31	10,61	month	83,87	72,22	77,61
loc.add.phys	0	0	0	year	82,85	61,70	70,73
loc.add.elec	66,66	12,50	21,05	century	66,66	66,66	66,66
loc.adm.town	49,39	31,29	38,31	reference-era	0	0	0
loc.adm.reg	50,00	37,50	42,85	time-modifier	71,83	40,53	51,82
loc.adm.nat	78,67	82,30	80,45	award-cat	0	0	0
loc.adm.sup	69,56	47,05	56,14	demonym	64,02	43,84	52,04
loc.phys.geo	50,00	10,00	16,66	name	67,64	47,84	56,04
loc.phys.hydro	0	0	0	name.last	46,35	38,68	42,17
loc.phys.astro	0	0	0	name.first	61,95	47,27	53,63
prod.object	0	0	0	name.nickname	92,30	19,67	32,43
prod.art	0	0	0	val	75,00	64,34	69,26
prod.media	62,22	54,90	58,33				
prod.fin	76,92	10,98	19,23				
prod.soft	0	0	0				
prod.award	100	16,66	28,57				
prod.serv	0	0	0				
prod.doctr	100	40,00	57,14				
prod.rule	0	0	0				
prod.other	0	0	0				
prod.unk	0	0	0				
func.coll	73,01	27,87	40,35				
func.ind	68,53	40,87	51,20				
org.adm	71,52	35,17	47,16				
org.ent	38,43	31,47	34,60				
Total	60,09	41,16	48,86	Total	66,36	44,77	53,47

TABLE A.2 – *Performance du système par catégorie et par composant sur le corpus d'évaluation (WER=24). (F : F-mesure, P : précision, R : rappel).*

A.2.2 WER = 25

Catégorie	P (%)	R (%)	F (%)	Composant	P (%)	R (%)	F (%)
amount	51,24	46,60	48,81	kind	71,83	27,27	39,53
pers.ind	62,99	48,52	54,81	extractor	0	0	0
pers.coll	49,31	38,91	43,50	qualifier	46,75	22,36	30,25
pers.other	0	0	0	title	56,75	28,37	37,83
time.date.abs	57,14	32,07	41,08	unit	76,52	68,94	72,53
time.date.rel	74,48	47,53	58,03	object	48,97	44,44	46,60
time.hour.abs	62,06	31,03	41,37	range-mark	78,78	37,68	50,98
time.hour.rel	79,16	30,31	43,84	day	73,33	70,967	72,13
loc.oro	50,00	11,11	18,18	week	84,84	66,66	74,66
loc.fac	55,55	05,31	09,70	month	82,25	71,83	76,69
loc.add.phys	0	0	0	year	90,4109589	70,96	79,51
loc.add.elec	86,66	76,47	81,25	century	66,66	66,66	66,66
loc.adm.town	44,84	33,20	38,15	reference-era	0	0	0
loc.adm.reg	50,00	31,91	38,96	time-modifier	73,28	37,74	49,82
loc.adm.nat	79,044	82,06	80,52	award-cat	0	0	0
loc.adm.sup	76,00	54,28	63,33	demonym	66,66	40,19	50,14
loc.phys.geo	0	0	0	name	66,37	45,41	53,93
loc.phys.hydro	0	0	0	name.last	48,13	36,35	41,42
loc.phys.astro	0	0	0	name.first	66,83	42,16	51,70
prod.object	0	0	0	name.nickname	78,57	18,33	29,72
prod.art	0	0	0	val	68,21	63,19	65,61
prod.media	72,72	47,05	57,14				
prod.fin	55,55	05,55	10,10				
prod.soft	0	0	0				
prod.serv	0	0	0				
prod.doctr	0	0	0				
prod.rule		0	0				
prod.other	0	0	0				
prod.unk	0	0	0				
func.coll	62,06	27,27	37,89				
func.ind	70,25	41,58	52,24				
org.adm	76,15	36,85	49,67				
org.ent	38,83	30,45	34,13				
org.other	0	0	0				
Total	61,00	40,68	48,81	Total	65,46	42,73	51,71

TABLE A.3 – *Performance du système par catégorie et par composant sur le corpus d'évaluation (WER=25). (F : F-mesure, P : précision, R : rappel).*

A.2.3 WER = 30

Catégorie	P (%)	R (%)	F (%)	Composant	P (%)	R (%)	F (%)
amount	55,85	47,95	51,60	kind	72,90	27,83	40,28
pers.ind	61,49	48,99	54,53	extractor	100	10,00	18,18
pers.coll	55,03	39,88	46,25	qualifier	48,63	22,75	31,00
pers.other	0	0	0	title	55,00	32,83	41,12
time.date.abs	51,40	26,69	35,14	unit	79,43	72,87	76,01
time.date.rel	57,33	24,50	34,34	object	52,65	47,60	50,00
time.hour.abs	44,82	25,00	32,09	range-mark	87,87	44,61	59,18
time.hour.rel	72,46	27,93	40,32	day	74,19	82,14	77,96
loc.oro	100	22,22	36,36	week	70,00	66,66	68,29
loc.fac	55,55	5,49	10,00	month	85,71	68,57	76,19
loc.add.phy	0	0	0	year	86,56	61,70	72,04
loc.add.elec	66,66	53,33	59,25	century	100	66,66	80,00
loc.adm.town	47,15	34,43	39,80	reference-era	0	0	0
loc.adm.reg	57,89	45,83	51,16	time-modifier	74,82	37,88	50,30
loc.adm.nat	77,40	81,32	79,31	award-cat	0	0	0
loc.adm.sup	71,42	44,11	54,54	demonym	72,41	41,37	52,66
loc.phys.geo	0	0	0	name	64,15	39,46	48,86
loc.phys.hydro	0	0	0	name.last	46,50	35,76	40,43
loc.phys.astro	0	0	0	name.first	65,68	42,95	51,94
prod.object	0	0	0	name.nickname	93,75	26,78	41,66
prod.art	0	0	0	val	74,37	66,84	70,40
prod.media	78,72	49,00	60,40				
prod.fin	85,71	06,89	12,76				
prod.soft	0	0	0				
prod.award	100	25,00	40,00				
prod.serv	0	0	0				
prod.doctr	25,00	20,00	22,22				
prod.rule	0	0	0				
prod.other	0	0	0				
func.ind	70,04	42,51	52,91				
func.coll	64,17	26,54	37,55				
org.adm	66,66	36,72	47,35				
org.ent	43,03	28,73	34,45				
Total	60,24	38,56	47,02	Total	66,25	42,17	51,54

TABLE A.4 – *Performance du système par catégorie et par composant sur le corpus d'évaluation (WER=30). (F : F-mesure, P : précision, R : rappel).*

A.2.4 WER = 35

Catégorie	P (%)	R (%)	F (%)	Composant	P (%)	R (%)	F (%)
amount	46,41	45,44	45,92	kind	59,76	22,21	32,38
pers.coll	44,73	36,95	40,47	extractor	22,22	18,18	20,00
pers.ind	45,22	46,60	45,90	qualifier	44,44	20,60	28,15
pers.other	0	0	0	title	43,24	20,77	28,07
time.date.abs	49,16	27,69	35,43	unit	68,70	63,34	65,91
time.date.rel	43,63	17,97	25,46	object	39,73	43,54	41,54
time.hour.abs	36,11	22,03	27,36	range-mark	81,81	39,13	52,94
time.hour.rel	79,16	30,15	43,67	day	64,70	68,75	66,66
loc.oro	66,66	22,22	33,33	week	53,70	67,44	59,79
loc.fac	33,33	04,16	07,40	month	75,40	65,71	70,22
loc.add.phys	0	0	0	year	72,60	55,20	62,72
loc.add.elec	54,54	35,29	42,85	century	100,00	66,66	80,00
loc.adm.town	29,92	29,47	29,69	reference-era	0	0	0
loc.adm.reg	4	36,73	38,29	time-modifier	67,51	36,80	47,64
loc.adm.nat	66,77	74,62	70,48	award-cat	0	0	0
loc.adm.sup	58,33	40,00	47,45	demonym	61,78	36,53	45,92
loc.phys.geo	100,00	03,22	6,25	name	49,45	35,14	41,08
loc.phys.hydro	0	0	0	name.first	51,06	40,88	45,41
loc.phys.astro	0	0	0	name.last	35,44	36,96	36,18
prod.object	12,50	01,66	02,94	name.nickname	81,25	20,96	33,33
prod.art	0	0	0	val	60,12	59,52	59,82
prod.media	60,52	44,80	51,49				
prod.fin	61,53	08,60	15,09				
prod.soft	0	0	0				
prod.award	100,00	16,66	28,57				
prod.serv	0	0	0				
prod.doctr	50,000	20,00	28,57				
prod.rule	0	0	0				
prod.other	0	0	0				
prod.unk	0	0	0				
func.coll	55,47	24,39	33,89				
func.ind	56,44	32,07	40,90				
org.adm	50,00	26,75	34,85				
org.ent	28,65	26,27	27,41				
Total	46,55	34,82	39,84	Total	52,60	38,13	44,21

TABLE A.5 – *Performance du système par catégorie et par composant sur le corpus d'évaluation (WER=35). (F : F-mesure, P : précision, R : rappel).*

A.2.5 Rover

Catégorie	P (%)	R (%)	F (%)	Composant	P (%)	R (%)	F (%)
amount	61,83	48,84	54,57	kind	73,91	27,21	39,77
pers.coll	51,51	38,85	44,29	extractor	100	12,50	22,22
pers.ind	70,25	53,66	60,84	qualifier	49,24	21,17	29,61
pers.other	0	0	0	title	61,11	32,83	42,71
time.date.abs	53,21	27,75	36,47	unit	86,81	70,62	77,89
time.date.rel	71,08	40,39	51,51	object	57,91	50,39	53,89
time.hour.abs	60,86	26,41	36,84	range-mark	87,09	45,00	59,34
time.hour.rel	88,88	30,93	45,90	day	68,75	68,75	68,75
loc.oro	50,00	11,11	18,18	week	78,37	70,73	74,35
loc.fac	55,55	05,55	10,10	month	81,66	68,05	74,24
loc.add.phys	0	0	0	year	92,64	67,02	77,77
loc.add.elec	81,81	60,00	69,23	century	100	66,66	80,00
loc.adm.town	61,53	36,36	45,71	reference-era	0	0	0
loc.adm.reg	52,77	39,58	45,23	time-modifier	76,77	37,82	50,67
loc.adm.nat	82,23	85,20	83,69	award-cat	0	0	0
loc.adm.sup	76,19	48,48	59,25	demonym	69,42	41,79	52,17
loc.phys.geo	100	3,84	7,40	name	72,28	45,24	55,65
loc.phys.hydro	0	0	0	name.first	72,99	48,32	58,15
loc.phys.astro	0	0	0	name.last	53,37	40,98	46,36
prod.object	0	0	0	name.nickname	88,23	26,31	40,54
prod.art	0	0	0	val	80,48	66,79	73,00
prod.media	77,52	45,69	57,50				
prod.fin	85,71	06,89	12,76				
prod.soft	0	0	0				
prod.award	100	16,66	28,57				
prod.serv	0	0	0				
prod.doctr	50,00	20,00	28,57				
prod.rule	0	0	0				
prod.other	0	0	0				
prod.unk	0	0	0				
func.coll	68,75	27,24	39,02				
func.ind	68,63	40,26	50,75				
org.adm	69,34	31,14	42,98				
org.ent	46,12	32,19	37,91				
Total	66,38	41,22	50,86	Total	71,62	44,63	55,00

TABLE A.6 – *Performance du système par catégorie et par composant sur le corpus d'évaluation (rover). (F : F-mesure, P : précision, R : rappel).*

A.3 Exemple de transcription manuelle balisée d'ESTER 2

```
<Section type="filler" startTime="11.466" endTime="16.226">
    <Turn startTime="11.466" endTime="16.226">
        <Sync time="11.466"/>
        <Background time="11.466" type="music" level="high"/>
        <Event desc="indicatif" type="noise" extent="instantaneous"/>
    </Turn>
</Section>
<Section type="report" startTime="16.226" endTime="90.466">
    <Turn speaker="spk1" startTime="16.226" endTime="21.795">
        <Sync time="16.226"/>
        <Event desc="time.hour" type="entities" extent="begin"/>
        7 heures
        <Event desc="time.hour" type="entities" extent="end"/>
        à
        <Event desc="loc.admi" type="entities" extent="begin"/>
        Paris
        <Event desc="loc.admi" type="entities" extent="end"/>
        ,
        <Background time="17.294" type="other" level="off"/>
        nous sommes
        <Event desc="time.date.abs" type="entities" extent="begin"/>
        le samedi 7 juillet 2007
        <Event desc="time.date.abs" type="entities" extent="end"/>
        .
        <Event desc="time.hour" type="entities" extent="begin"/>
        7 heures
        <Event desc="time.hour" type="entities" extent="end"/>
        ,
        <Event desc="loc.admi" type="entities" extent="begin"/>
        Paris
        <Event desc="loc.admi" type="entities" extent="end"/>
        , le journal ,
        <Event desc="pers.hum" type="entities" extent="begin"/>
        Sylvie Berruet
        <Event desc="pers.hum" type="entities" extent="end"/>
        , bonjour .
    </Turn>
    <Turn speaker="spk2" startTime="21.795" endTime="22.625">
        <Sync time="21.795"/>
        bonjour .
    </Turn>
    <Turn startTime="22.625" endTime="24.163">
        <Sync time="22.625"/>
        <Background time="22.625" type="music" level="high"/>
        <Event desc="musique" type="noise" extent="instantaneous"/>
    </Turn>
```

<Turn speaker="spk2" startTime="24.163" endTime="90.466">
<Sync time="24.163"/>
8 concerts sur 5 continents pour dénoncer le réchauffement climatique ,
<Event desc="org.gsp" type="entities" extent="begin"/>
Sydney
<Event desc="org.gsp" type="entities" extent="end"/>
a donné le coup d'envoi du
<Event desc="en" type="language" extent="begin"/>
Live Earth
<Event desc="en" type="language" extent="end"/>
<Event desc="rev" type="lexical" extent="begin"/>
une ma(nifestation)
<Event desc="rev" type="lexical" extent="end"/>
une initiative lancée par
<Event desc="pers.hum" type="entities" extent="begin"/>
<Event desc="pers.hum" type="entities" extent="begin"/>
Al Gore
<Event desc="pers.hum" type="entities" extent="end"/>
,
<Event desc="r" type="noise" extent="instantaneous"/>
l'ancien
<Event desc="fonc.pol" type="entities" extent="begin"/>
vice-président
<Event desc="fonc.pol" type="entities" extent="end"/>
des
<Event desc="loc.admi" type="entities" extent="begin"/>
États-Unis
<Event desc="loc.admi" type="entities" extent="end"/>
<Event desc="pers.hum" type="entities" extent="end"/>
.
<Sync time="35.389"/>
<Event desc="r" type="noise" extent="instantaneous"/>
de
<Event desc="loc.admi" type="entities" extent="begin"/>
Shanghai
<Event desc="loc.admi" type="entities" extent="end"/>
à
<Event desc="loc.admi" type="entities" extent="begin"/>
Londres
<Event desc="loc.admi" type="entities" extent="end"/>
<Event desc="en" type="language" extent="begin"/>
Live Earth
<Event desc="en" type="language" extent="end"/>
<Event desc="musique" type="noise" extent="begin"/>
s'achèvera au
<Event desc="loc.admi" type="entities" extent="begin"/>
Brésil
<Event desc="loc.admi" type="entities" extent="end"/>
, à
<Event desc="loc.admi" type="entities" extent="begin"/>

Rio
<Event desc="loc.admi" type="entities" extent="end"/>
. 2 milliards de spectateurs
<Event desc="r" type="noise" extent="instantaneous"/>
devraient suivre l'événement à la télévision et sur internet .
<Event desc="musique" type="noise" extent="end"/>
<Sync time="45.522"/>
<Event desc="r" type="noise" extent="instantaneous"/>
première inculpation dans les attentats manqués de
<Event desc="loc.admi" type="entities" extent="begin"/>
Londres
<Event desc="loc.admi" type="entities" extent="end"/>
et de
<Event desc="loc.admi" type="entities" extent="begin"/>
Glasgow
<Event desc="loc.admi" type="entities" extent="end"/>
, un médecin iraquien qui figurait parmi les 8 suspects
<Event desc="r" type="noise" extent="instantaneous"/>
a été mis en examen
<Event desc="time.date.rel" type="entities" extent="begin"/>
hier soir
<Event desc="time.date.rel" type="entities" extent="end"/>
.
<Sync time="53.767"/>
<Event desc="r" type="noise" extent="instantaneous"/>
en
<Event desc="loc.admi" type="entities" extent="begin"/>
France
<Event desc="loc.admi" type="entities" extent="end"/>
, 2 perquisitions
<Event desc="amount.phy.dur" type="entities" extent="begin"/>
en 2 jours
<Event desc="amount.phy.dur" type="entities" extent="end"/>
. les juges de l'affaire
<Event desc="org.com" type="entities" extent="begin"/>
Clearstream
<Event desc="org.com" type="entities" extent="end"/>
ont perquisitionné
<Event desc="time.date.rel" type="entities" extent="begin"/>
hier
<Event desc="time.date.rel" type="entities" extent="end"/>
le bureau de
<Event desc="pers.hum" type="entities" extent="begin"/>
<Event desc="pers.hum" type="entities" extent="begin"/>
Dominique de Villepin
<Event desc="pers.hum" type="entities" extent="end"/>
,
<Event desc="r" type="noise" extent="instantaneous"/>
l'ancien
<Event desc="fonc.pol" type="entities" extent="begin"/>

Premier ministre
<Event desc="fonc.pol" type="entities" extent="end"/>
<Event desc="pers.hum" type="entities" extent="end"/>
.
<Sync time="62.327"/>
<Event desc="r" type="noise" extent="instantaneous"/>
les sports avec le départ du quatre-vingt-quatorzième
<Event desc="org.div" type="entities" extent="begin"/>
Tour de
<Event desc="loc.admi" type="entities" extent="begin"/>
France
<Event desc="loc.admi" type="entities" extent="end"/>
<Event desc="org.div" type="entities" extent="end"/>
. le prologue s'élance à
<Event desc="loc.admi" type="entities" extent="begin"/>
Londres
<Event desc="loc.admi" type="entities" extent="end"/>
<Event desc="time.date.rel" type="entities" extent="begin"/>
cet après-midi
<Event desc="time.date.rel" type="entities" extent="end"/>

A.4 Exemple de transcription manuelle balisée d'ETAPE

bonsoir . bonsoir à toutes et tous . merci d' être fidèles au rendez-vous de **<org.ent> <name>** BFM TV **</name> </org.ent>** , première chaîne info de **<loc.adm.nat> <name>** France **</name> </loc.adm.nat>** . c' est l' heure de **<prod.media> <name>** BFM Story **</name> </prod.media>** . nous sommes ensemble **<amount>** pour **<val>** 60 **</val> <unit>** minutes **</unit> </amount>** , **<amount> <val>** une **</val> <unit>** heure **</unit> </amount>** , au coeur de tout ce qui fait l' actualité . l' actualité , c' est la sécheresse . il faudra **<amount> <val><amount> <val>** des centaines de millions **</val>** d' **<unit>** euros **</unit> </amount>** pour indemniser les **<func.coll> <kind>** éleveurs **</kind> </func.coll>** . c' est ce qu' a dit le **<func.ind> <kind>** ministre **</kind> </func.ind> <pers.ind> <name.first>** Bruno **</name.first> <name.last>** Le Maire **</name.last> </pers.ind>** , **<func.ind> <kind>** ministre **</kind>** de l' **<org.adm> <name>** agriculture **</name> </org.adm> </func.ind>** . nous , nous serons en direct avec le **<func.ind> <kind>** président **</kind> </func.ind>** d' **<amount> <extractor>** un **</extractor>** des **<object>** départements **</object> </amount>** les plus touchés hm par cette sécheresse , c' est la **<loc.adm.reg> <name>** Charente-Maritime **</name> </loc.adm.reg>** , et son **<func.ind> <kind>** président **</kind> </func.ind>** , c' est l' **<func.ind> <qualifier>** ancien **</qualifier> <kind>** ministre **</kind> </func.ind> <pers.ind> <name.first>** Dominique **</name.first> <name.last>** Bussereau **</name.last> </pers.ind>** . nous parlerons également de ce **<time.date.abs> <name>** printemps **</name> <year>** 2011 **</year> </time.date.abs>** , qui est d' ores et déjà le plus chaud **<time.date.abs> <time-modifier>** depuis le début **</time-modifier>** du **<century>** vingtième siècle **</century> </time.date.abs>** . nous avons invité un **<func.ind> <kind>** climatologue **</kind> </func.ind>** pour en parler . nous avons également invité un **<func.ind> <kind>** éleveur **</kind> <qualifier>** laitier **</qualifier> </func.ind>** , pour parler des conséquences de cette sécheresse . et puis , l' actualité , c' est aussi ce concombre contaminé , qui a déjà fait **<amount> <val>** 16 **</val> <object>** morts **</object> </amount>** . d' où vient cette bactérie tueuse ? nous en parlerons avec

un spécialiste, avant d' évoquer l' actualité politique . c' était **<time.hour.rel> <time-modifier>** ce **</time-modifier> <name>** matin **</name> </time.hour.rel>** , sur **<prod.media> <name>** BFM TV **</name> </prod.media>** , **<pers.ind> <name.first>** Jean-Louis **</name.first> <name.last>** Borloo **</name.last> </pers.ind>** , presque **<func.ind> <kind>** candidat **</kind> </func.ind>** . analyse avec **<pers.ind> <name.first>** Olivier **</name.first> <name.last>** Mazerolle **</name.last> </pers.ind>** , **<time.hour.rel> <time-modifier>** ce **</time-modifier> <name>** soir **</name> </time.hour.rel>** . et on en reparle bien sûr dans le journal . et dans ce journal , nous irons aussi à **<loc.fac> <name>** Roland Garros **</name> </loc.fac>** , voir où en est **<time.hour.rel> <time-modifier>** en ce **</time-modifier> <name>** moment **</name> </time.hour.rel> <pers.ind> <name.first>** Gaël </name.first> <name.last> Monfils **</name.last> </pers.ind>** , face à à **<pers.ind> <name.first>** Roger **</name.first> <name.last>** Federer **</name.last> </pers.ind>** . et **<amount> <val>** 2 **</val> <object>** enfants **</object> </amount>** restent **<time.hour.rel> <time-modifier>** ce **</time-modifier> <name>** soir **</name> </time.hour.rel> </name> </time.hour.rel>** dans un état critique . ils font partie du groupe d' élèves blessés **<time.date.rel> <name>** hier **</name> </time.date.rel>** , à un rond-point , par une camionnette de gendarmerie . une fillette est décédée **<time.hour.rel> <time-modifier>** dans la **</time-modifier> <name>** soirée **</name> </time.hour.rel>**, c' était **<time.date.rel> <name>** hier **</name> </time.date.rel>** , **<amount> <val>** 6 **</val> <qualifier>** autres **</qualifier> <object>** enfants **</object> </amount>** sont grièvement touchés . **<pers.ind> <name.first>** Benoît **</name.first> <name.last>** Petit **</name.last> </pers.ind>** , vous êtes avec euh **<pers.ind> <name.first>** Charles-wood **</name.first>** euh **<name.last>** Longchamps **</name.last> </pers.ind>** . euh vous êtes retourné sur les lieux du drame , à **<loc.adm.town> <name>** Joué-les-Tours **</name> </loc.adm.town>** , le **<func.ind> <kind>** procureur **</kind> </func.ind>** va normalement mettre le conducteur en examen . **<pers.ind> <name.first>** Benoît **</name.first> </pers.ind>** . oui , une mise en examen pour euh homicide involontaire , blessures involontaires et défaut de maîtrise du véhicule . pour l' ensemble de ces faits , euh le **<func.ind> <kind>** gendarme **</kind> </func.ind>** mis en cause , âgé euh de **<amount> <val>** 34 **</val> <unit>** ans **</unit> </amount>** , marié , père de famille , risque **<amount> <qualifier>** jusqu'à **</qualifier> <val>** 5 **</val> <unit>** ans **</unit> </amount>** de prison , mais il devrait être remis en liberté euh **<time.hour.rel> <time-modifier>** dès ce **</time-modifier> <name>** soir **</name> </time.hour.rel>** . une mise en liberté assortie d' un contrôle judiciaire strict . et euh **<pers.ind> <name.first>** Benoît **</name.first> <name.last>** Petit **</name.last> </pers.ind>** , alors sait-on **<time.date.rel> <name>** maintenant **</name> </time.date.rel>** pourquoi le conducteur , ce **<func.ind> <kind>** gendarme **</kind> </func.ind>** , a perdu le contrôle de son véhicule , **<time.date.rel> <name>** hier **</name> </time.date.rel>**? alors , les circonstances de ce drame restent **<time.hour.rel> <time-modifier>** pour l' **</time-modifier> <name>** instant **</name> </time.hour.rel>** assez floues . en fait , le **<func.ind> <kind>** procureur **</kind>** de la **<org.adm> <kind>** république **</kind> </org.adm>** de **<loc.adm.town> <name>** Tours **</name></loc.adm.town> </func.ind>** a surtout euh fermé des portes , écarté euh des pistes . il a d' abord rappelé que le **<func.ind> <kind>** gendarme **</kind> </func.ind>** eh bien maintient euh ses déclarations euh des **<time.date.rel> <time-modifier>** premières **</time-modifier> <kind>** heures **</kind> </time.date.rel>** . à la sortie euh du rond-point , il aurait perdu le la maîtrise de son véhicule . je vous propose d' écouter les précisions du **<func.ind> <kind>** procureur **</kind> </func.ind>** . le conducteur de du fourgon d' **<loc.adm.town> <name>** Auxerre **</name> </loc.adm.town>** explique , quant à lui , que lorsqu' il s' est engagé dans le rond-point , le pneu arrière droit du fourgon est monté sur le trottoir , ou a touché le trottoir , ce qui a déporté le fourgon vers la gauche , et qu' **<time.hour.rel> <time-modifier>** à partir de ce **</time-modifier> <name>** moment **</name> <time-modifier>** -là **</time-modifier> </time.hour.rel>** , il a complètement perdu le contrôle du véhicule . le **<func.ind> <kind>** gendarme **</kind> </func.ind>** a fait **<amount> <val>** 2 **</val> <object>** évaluations **</object> </amount>** . il prétend qu' il s' est arrêté à l' entrée du rond-point , et euh il donne une évaluation de l' ordre euh de euh **<amount> <val>** 30 **</val> </amount>** , **<amount> <val>** 40 **</val> <unit>** kilomètres heure **</unit> </amount>** . l' enquête a déjà permis de déterminer que le

<func.ind> <kind> gendarme **</kind> </func.ind>** ne roulait ni sous l' emprise de l' alcool , ni sous l' emprise euh de substances stupéfiantes .

A.5 Exemple de transcription automatique balisée d'ETAPE

le rendez vous de <annotation id=0 type="org.ent" ftype="s" depth=0/> <annotation id=1 type="name" ftype="s" depth=1 parent=0/> **bfm** <annotation id=0 type="org.ent" ftype="s" depth=0/> <annotation id=1 type="name" ftype="s" depth=1 parent=0/> **tv** <annotation id=0 type="org.ent" ftype="e" depth=0/> <annotation id=1 type="name" ftype="e" depth=1 parent=0/> **première chaîne info de** <annotation id=2 type="loc.adm.nat" ftype="s" depth=0/> <annotation id=3 type="name" ftype="s" depth=1 parent=2/> **france** <annotation id=2 type="loc.adm.nat" ftype="e" depth=0/> <annotation id=3 type="name" ftype="e" depth=1 parent=2/> **c' est l' heure de** <annotation id=4 type="prod.media" ftype="s" depth=0/> <annotation id=5 type="name" ftype="s" depth=1 parent=4/> **bfm** <annotation id=4 type="prod.media" ftype="s" depth=0/> <annotation id=5 type="name" ftype="s" depth=1 parent=4/> **story** <annotation id=4 type="prod.media" ftype="e" depth=0/> <annotation id=5 type="name" ftype="e" depth=1 parent=4/> **nous sommes ensemble** <annotation id=6 type="amount" ftype="s" depth=0/> **pour** <annotation id=7 type="val" ftype="s" depth=1 parent=6/> **soixante** <annotation id=7 type="val" ftype="e" depth=1 parent=6/> <annotation id=8 type="unit" ftype="s" depth=1 parent=6/> **minutes** <annotation id=6 type="amount" ftype="e" depth=0/> <annotation id=8 type="unit" ftype="e" depth=1 parent=6/> <annotation id=9 type="amount" ftype="s" depth=0/> <annotation id=10 type="val" ftype="s" depth=1 parent=9/> **une** <annotation id=10 type="val" ftype="e" depth=1 parent=9/> <annotation id=11 type="unit" ftype="s" depth=1 parent=9/> **heure** <annotation id=9 type="amount" ftype="e" depth=0/> <annotation id=11 type="unit" ftype="e" depth=1 parent=9/> **au coeur de tout ce qui fait l' actualité c' est la sécheresse il faudra** <annotation id=12 type="amount" ftype="s" depth=0/> <annotation id=13 type="val" ftype="s" depth=1 parent=12/> **des centaines de millions** <annotation id=13 type="val" ftype="e" depth=1 parent=12/> **d'** <annotation id=14 type="unit" ftype="s" depth=1 parent=12/> **euros** <annotation id=12 type="amount" ftype="e" depth=0/> <annotation id=14 type="unit" ftype="e" depth=1 parent=12/> **pour indemniser les** <annotation id=15 type="func.coll" ftype="s" depth=0/> <annotation id=16 type="kind" ftype="s" depth=1 parent=15/> **éleveurs** <annotation id=15 type="func.coll" ftype="e" depth=0/> <annotation id=16 type="kind" ftype="e" depth=1 parent=15/> **c'** <annotation id=15 type="func.coll" ftype="e" depth=0/> <annotation id=16 type="kind" ftype="e" depth=1 parent=15/> **est ce qu' a dit** <annotation id=17 type="pers.ind" ftype="s" depth=0/> <annotation id=18 type="name.last" ftype="s" depth=1 parent=17/> **la** <annotation id=17 type="pers.ind" ftype="e" depth=0/> <annotation id=18 type="name.last" ftype="e" depth=1 parent=17/> <annotation id=19 type="func.ind" ftype="s" depth=0/> <annotation id=20 type="kind" ftype="s" depth=1 parent=19/> **ministre** <annotation id=20 type="kind" ftype="e" depth=1 parent=19/> **de l'** <annotation id=21 type="org.adm" ftype="s" depth=1 parent=19/> <annotation id=22 type="name" ftype="s" depth=2 parent=21/> **agriculture** <annotation id=19 type="func.ind" ftype="e" depth=0/> <annotation id=21 type="org.adm" ftype="e" depth=1 parent=19/> <annotation id=22 type="name" ftype="e" depth=2 parent=21/> **nous nous serons en direct avec le** <annotation id=23 type="func.ind" ftype="s" depth=0/> <annotation id=24 type="kind" ftype="s" depth=1 parent=23/> **président** <annotation id=23 type="func.ind" ftype="e" depth=0/> <annotation id=24 type="kind" ftype="e" depth=1 parent=23/> **d'** <annotation id=23 type="func.ind" ftype="e" depth=0/> <annotation id=24 type="kind" ftype="e" depth=1 parent=23/> <annotation id=25 type="amount" ftype="s" depth=0/> <annotation id=26 type="extractor" ftype="s" depth=1 parent=25/> **un** <annotation id=26 type="extractor" ftype="e" depth=1 parent=25/> <annotation

id=27 type="object" ftype="s" depth=1 parent=25/> **département** <annotation id=25 type="amount" ftype="e" depth=0/> <annotation id=27 type="object" ftype="e" depth=1 parent=25/> **les plus touchés par cette sécheresse c' est la** <annotation id=28 type="loc.adm.reg" ftype="s" depth=0/> <annotation id=29 type="name" ftype="s" depth=1 parent=28/> **charente maritime** <annotation id=28 type="loc.adm.reg" ftype="e" depth=0/> <annotation id=29 type="name" ftype="e" depth=1 parent=28/> **et** <annotation id=28 type="loc.adm.reg" ftype="e" depth=0/> <annotation id=29 type="name" ftype="e" depth=1 parent=28/> **son** <annotation id=30 type="func.ind" ftype="s" depth=0/> <annotation id=31 type="kind" ftype="s" depth=1 parent=30/> **président** <annotation id=30 type="func.ind" ftype="e" depth=0/> <annotation id=31 type="kind" ftype="e" depth=1 parent=30/> **c' est** <annotation id=32 type="func.ind" ftype="s" depth=0/> <annotation id=33 type="qualifier" ftype="s" depth=1 parent=32/> **l'** <annotation id=32 type="func.ind" ftype="s" depth=0/> <annotation id=33 type="qualifier" ftype="s" depth=1 parent=32/> **ancien** <annotation id=33 type="qualifier" ftype="e" depth=1 parent=32/> <annotation id=34 type="kind" ftype="s" depth=1 parent=32/> **ministre** <annotation id=32 type="func.ind" ftype="e" depth=0/> <annotation id=34 type="kind" ftype="e" depth=1 parent=32/> <annotation id=35 type="pers.ind" ftype="s" depth=0/> <annotation id=36 type="name.first" ftype="s" depth=1 parent=35/> **dominique** <annotation id=36 type="name.first" ftype="e" depth=1 parent=35/> <annotation id=37 type="name.last" ftype="s" depth=1 parent=35/> **bussereau** <annotation id=35 type="pers.ind" ftype="e" depth=0/> <annotation id=37 type="name.last" ftype="e" depth=1 parent=35/> **nous parlerons également de ce** <annotation id=38 type="time.date.abs" ftype="s" depth=0/> <annotation id=39 type="name" ftype="s" depth=1 parent=38/> **printemps** <annotation id=39 type="name" ftype="e" depth=1 parent=38/> <annotation id=40 type="year" ftype="s" depth=1 parent=38/> **deux mille onze** <annotation id=38 type="time.date.abs" ftype="e" depth=0/> <annotation id=40 type="year" ftype="e" depth=1 parent=38/> **qui est d' ores et déjà le plus chaud** <annotation id=41 type="time.date.abs" ftype="s" depth=0/> <annotation id=42 type="time-modifier" ftype="s" depth=1 parent=41/> **depuis le début** <annotation id=42 type="time-modifier" ftype="e" depth=1 parent=41/> **du** <annotation id=43 type="century" ftype="s" depth=1 parent=41/> **vingtième siècle** <annotation id=41 type="time.date.abs" ftype="e" depth=0/> <annotation id=43 type="century" ftype="e" depth=1 parent=41/> **nous avons invité** <annotation id=44 type="func.ind" ftype="s" depth=0/> <annotation id=45 type="kind" ftype="s" depth=1 parent=44/> **climat** <annotation id=44 type="func.ind" ftype="e" depth=0/> <annotation id=45 type="kind" ftype="e" depth=1 parent=44/> **pour en parler nous avons également invité à l'** <annotation id=46 type="func.ind" ftype="s" depth=0/> <annotation id=47 type="kind" ftype="s" depth=1 parent=46/> **éleveur** <annotation id=47 type="kind" ftype="e" depth=1 parent=46/> <annotation id=48 type="qualifier" ftype="s" depth=1 parent=46/> **laitier** <annotation id=46 type="func.ind" ftype="e" depth=0/> <annotation id=48 type="qualifier" ftype="e" depth=1 parent=46/> **pour parler des conséquences de cette sécheresse et puis l' actualité c' est aussi sûr concombres contaminés qui a déjà fait** <annotation id=49 type="amount" ftype="s" depth=0/> <annotation id=50 type="val" ftype="s" depth=1 parent=49/> **seize** <annotation id=50 type="val" ftype="e" depth=1 parent=49/> <annotation id=51 type="object" ftype="s" depth=1 parent=49/> **morts** <annotation id=49 type="amount" ftype="e" depth=0/> <annotation id=51 type="object" ftype="e" depth=1 parent=49/> **nous en parlerons avec un spécialiste avant d' évoquer l' actualité politique c' était** <annotation id=52 type="time.hour.rel" ftype="s" depth=0/> <annotation id=53 type="time-modifier" ftype="s" depth=1 parent=52/> **ce** <annotation id=53 type="time-modifier" ftype="e" depth=1 parent=52/> <annotation id=54 type="name" ftype="s" depth=1 parent=52/> **matin** <annotation id=52 type="time.hour.rel" ftype="e" depth=0/> <annotation id=54 type="name" ftype="e" depth=1 parent=52/> **sur** <annotation id=55 type="prod.media" ftype="s" depth=0/> <annotation id=56 type="name" ftype="s" depth=1 parent=55/> **bfmtv** <annotation id=55 type="prod.media" ftype="e" depth=0/> <annotation id=56 type="name" ftype="e" depth=1 parent=55/> <annotation id=57 type="pers.ind" ftype="s" depth=0/> <annotation id=58 type="name.first" ftype="s" depth=1 parent=57/> **jean louis** <anno-

tation id=58 type="name.first" ftype="e" depth=1 parent=57/> <annotation id=59 type="name.last" ftype="s" depth=1 parent=57/> **borloo** <annotation id=57 type="pers.ind" ftype="e" depth=0/> <annotation id=59 type="name.last" ftype="e" depth=1 parent=57/> **presque** <annotation id=60 type="func.ind" ftype="s" depth=0/> <annotation id=61 type="kind" ftype="s" depth=1 parent=60/> **candidat** <annotation id=60 type="func.ind" ftype="e" depth=0/> <annotation id=61 type="kind" ftype="e" depth=1 parent=60/> **analyse avec** <annotation id=62 type="pers.ind" ftype="s" depth=0/> <annotation id=63 type="name.first" ftype="s" depth=1 parent=62/> **olivier** <annotation id=63 type="name.first" ftype="e" depth=1 parent=62/> <annotation id=64 type="name.last" ftype="s" depth=1 parent=62/> **mazerolle** <annotation id=62 type="pers.ind" ftype="e" depth=0/> <annotation id=64 type="name.last" ftype="e" depth=1 parent=62/> <annotation id=65 type="time.hour.rel" ftype="s" depth=0/> <annotation id=66 type="time-modifier" ftype="s" depth=1 parent=65/> **ce** <annotation id=66 type="time-modifier" ftype="e" depth=1 parent=65/> <annotation id=67 type="name" ftype="s" depth=1 parent=65/> **soir** <annotation id=65 type="time.hour.rel" ftype="e" depth=0/> <annotation id=67 type="name" ftype="e" depth=1 parent=65/> **bien sûr dans le journal dans ce journal nous irons aussi** <annotation id=68 type="loc.fac" ftype="s" depth=0/> <annotation id=69 type="name" ftype="s" depth=1 parent=68/> **à** <annotation id=68 type="loc.fac" ftype="s" depth=0/> <annotation id=69 type="name" ftype="s" depth=1 parent=68/> **roland garros** <annotation id=68 type="loc.fac" ftype="e" depth=0/> <annotation id=69 type="name" ftype="e" depth=1 parent=68/> **warhouver** <annotation id=70 type="time.hour.rel" ftype="s" depth=0/> <annotation id=71 type="time-modifier" ftype="s" depth=1 parent=70/> **en ce** <annotation id=71 type="time-modifier" ftype="e" depth=1 parent=70/> <annotation id=72 type="name" ftype="s" depth=1 parent=70/> **moment** <annotation id=70 type="time.hour.rel" ftype="e" depth=0/> <annotation id=72 type="name" ftype="e" depth=1 parent=70/> <annotation id=73 type="pers.ind" ftype="s" depth=0/> <annotation id=74 type="name.first" ftype="s" depth=1 parent=73/> **gaël** <annotation id=74 type="name.first" ftype="e" depth=1 parent=73/> <annotation id=75 type="name.last" ftype="s" depth=1 parent=73/> **monfils** <annotation id=73 type="pers.ind" ftype="e" depth=0/> <annotation id=75 type="name.last" ftype="e" depth=1 parent=73/> **face à** <annotation id=76 type="pers.ind" ftype="s" depth=0/> <annotation id=77 type="name.first" ftype="s" depth=1 parent=76/> **roger** <annotation id=77 type="name.first" ftype="e" depth=1 parent=76/> <annotation id=78 type="name.last" ftype="s" depth=1 parent=76/> **federer** <annotation id=76 type="pers.ind" ftype="e" depth=0/> <annotation id=78 type="name.last" ftype="e" depth=1 parent=76/>

Bibliographie

Jens Allwood. Multimodal Corpora. *Lüdeling, Anke Kytö, Merja (eds.) Corpus Linguistics. An International Handbook. Mouton de Gruyter*, pages 207–225, 2008. (Cité page 29.)

Mikheev Andrei, Grover Claire, et Moens Marc. Description of the LTG system used for MUC-7. Dans *Proceedings of the Seventh Message Understanding Conference (MUC-7)*, pages 1–12, Fairfax, VA, USA, 1998. (Cité pages 25, 26 et 31.)

Borthwick Andrew, Sterling John, Agichtein Eugene, et Grishman Ralph. NYU : Description of the MENE Named Entity System as Used in MUC-7. Dans *Proceedings of the Seventh Message Understanding Conference (MUC-7)*, pages 1–6, Fairfax, VA, USA, 1998. (Cité pages 23 et 31.)

Frédéric Béchet. *LIA_PHON* : un système complet de phonétisation de textes. *Traitement Automatique des Langues (TAL)*, 42 :47–67, 2001. (Cité page 89.)

Frederic Bechet, Remi Auguste, Stephane Ayache, Delphine Charlet, Geraldine Damnati, Benoit Favre, Corinne Fredouille, Christophe Levy, Georges Linares, et Jean Martinet. Percolo ? un système multimodal de détection de personnes dans des documents vidéo. Dans *Proceedings of the Joint Conference JEP-TALN-RECITAL 2012, volume 1 : JEP*, pages 553–560, Grenoble, France, 2012. (Cité page 38.)

Benjamin Bigot, Grégory Senay, Georges Linarès, Corinne Fredouille, et Richard Dufour. Person name recognition in ASR outputs using continuous context models. Dans *Proceedings of the International Conference on Acoustic Speech and Signal Processing (ICASSP'13)*, pages 8470–8474, Vancouver, Canada, 2013. (Cité page 36.)

Daniel M. Bikel, Scott Miller, Richard Schwartz, et Ralph Weischedel. Nymble : a high-performance learning name-finder. Dans *Proceedings of the 5th Conference on Applied Natural Language Processing (ANLP'97)*, pages 194–201, Washington, DC, USA, 1997. (Cité page 23.)

Caroline Brun et Maud Ehrmann. Un système de détection d'entités nommées adapté pour la campagne d'évaluation ESTER 2. Dans *Actes de la 17ème conférence sur le Traitement Automatique des Langues Naturelles (TALN'10)*, Montréal, Canada, 2010. (Cité page 21.)

David Burkett, Slav Petrov, John Blitzer, et Dan Klein. Learning better monolingual models with unannotated bilingual text. Dans *Proceedings of the 40th Conference on Computational Natural Language Learning (ConLL'10)*, pages 46–54, Uppsala, Sweden, 2010. (Cité page 41.)

Frédéric Béchet et Eric Charton. Unsupervised knowledge acquisition for Extracting Named Entities from speech. Dans *Proceedings of the 35th IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'10)*, pages 5338–5341, Dallas, TX, USA, 2010. (Cité pages 17, 23, 33, 87 et 90.)

Wanxiang Che, Mengqiu Wang, Christopher D. Manning, et Ting Liu. Named entity recognition with bilingual constraints. Dans *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies (NAACL-HLT'13)*, pages 52–62, Atlanta, GA, USA, 2013. (Cité pages 48 et 49.)

Stanley F. Chen et Joshua Goodman. An empirical study of smoothing techniques for language modeling. Dans *Proceedings of the 34th annual meeting on Association for Computational Linguistics (ACL'96)*, pages 310–318, Santa Cruz, CA, USA, 1996. (Cité page 88.)

Nancy Chinchor. Overview of MUC-7/MET-2. Dans *Proceedings of the Seventh Message Understanding Conference (MUC-7)*, pages 1–4, Fairfax, VA, USA, 1998. (Cité page 10.)

Béatrice Daille, Nordine Fourour, et Emmanuel Morin. Catégorisation des noms propres : une étude en corpus. *Cahiers de Grammaire*, 25 :61–73, 2000. (Cité pages x, 13, 14 et 15.)

Kareem Darwish. Named entity recognition using cross-lingual resources : Arabic as an example. Dans *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL'13)*, pages 1558–1567, Sofia, Bulgaria, 2013. (Cité page 41.)

Anoop Deoras, Gökhan Tür, Ruhi Sarikaya, et Dilek Z. Hakkani-Tür. Joint discriminative decoding of words and semantic tags for spoken language understanding. *IEEE Transactions on Audio, Speech & Language Processing*, 21(8) :1612–1621, 2013. (Cité pages 34 et 35.)

Marco Dinarelli et Sophie Rosset. Models cascade for tree-structured named entity detection. Dans *Proceedings of 5th International Joint Conference on Natural Language Processing (IJCNLP'11)*, pages 1269–1278, Chiang Mai, Thailand, 2011. (Cité pages 23 et 78.)

Marco Dinarelli et Sophie Rosset. Tree-structured named entity recognition on ocr data : Analysis, processing and results. Dans *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'12)*, pages 1266–1272, Istanbul, Turkey, 2012. (Cité page 37.)

George R. Doddington, Alexis Mitchell, Mark A. Przybocki, Lance A. Ramshaw, Stephanie Strassel, et Ralph M. Weischedel. The Automatic Content Extraction (ACE) Program Tasks, Data, and Evaluation. Dans *Proceedings of the 5th Conference on Language Resources and Evaluation (LREC 2004)*, pages 837–840, Lisbon, Portugal, 2004. (Cité page 10.)

Richard Dufour. *Transcription automatique de la parole spontanée*. Thèse de doctorat, Université du Maine, Le Mans, France, 2010. (Cité page 89.)

Richard Dufour, Géraldine Damnati, Delphine Charlet, et Frédéric Béchet. Automatic transcription error recovery for person name recognition. Dans *Proceedings of the 13th Annual Conference of the International Speech Communication Association (Interspeech'12)*, Portland, OR, USA, 2012. (Cité pages 100 et 104.)

Maud Ehrmann. *Les Entités Nommées, de la Linguistique au TAL - Statut Théorique et Méthodes de Désambiguïsation*. Thèse de doctorat, Université Paris 7, France, 2008. (Cité pages 9 et 12.)

Maud Ehrmann et Guillaume Jacquet. Vers une double annotation des entités nommées. *Traitement automatique des langues (TAL)*, 42 :63–88, 2006. (Cité page 19.)

Yannick Estève. *Traitement automatique de la parole : contributions*. Habilitation à Diriger des Recherches (HDR), LIUM, Université du Maine, France, 2009. (Cité pages ix, 89 et 91.)

Benoît Favre, Frédéric Béchet, et Pascal Nocéra. Robust named entity extraction from large spoken archives. Dans *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing (HLT'05)*, pages 491–498, Vancouver, British Columbia, Canada, 2005. (Cité pages 34, 35, 82 et 104.)

Jenny Rose Finkel, Trond Grenager, et Christopher Manning. Incorporating non-local information into information extraction systems by Gibbs sampling. Dans *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics (ACL'05)*, pages 363–370, Ann Arbor, MI, USA, 2005. (Cité pages 46 et 47.)

Nordine Fourour. Nemesis, un système de reconnaissance incrémentielle des entités nommées pour le français. Dans *Actes de la 9ème conférence sur le Traitement Automatique des Langues Naturelles (TALN'02)*, pages 265–274, Nancy, France, 2002. (Cité pages 22, 43 et 67.)

Nordine Fourour. *Identification et catégorisation automatique des entités nommées dans les textes français*. Thèse de doctorat, Université de Nantes, France, 2004. (Cité page 21.)

Nordine Fourour et Emmanuel Morin. Apport du Web dans la reconnaissance des entités nommées. *Revue Québécoise de Linguistique (RQL)*, pages 41–60, 2003. (Cité page 23.)

Nathalie Friburger. *Reconnaissance automatique des noms propres : application à la classification automatique de textes journalistiques*. Thèse de doctorat, Université François Rabelais, Tours, France, 2002. (Cité pages 21 et 78.)

Ruiji Fu, Bing Qin, et Ting Liu. Generating chinese named entity data from a parallel corpus. Dans *Proceedings of the 5th International Joint Conference on Natural Language Processing (IJCNLP'11)*, pages 264–272, Chiang Mai, Thailand, 2011. (Cité page 48.)

O. Galibert, S. Rosset, C. Grouin, et P Zweigenbaum. Structured and extended named entity evaluation in automatic speech transcriptions. Dans

Proceedings of the International Joint Conference on Natural Language Processing (IJCNLP'11), pages 518–526, Chiang Mai, Thailand, 2011. (Cité pages 61 et 113.)

Sylvain Galliano, Edouard Geoffrois, Djamel Mostefa, Khalid Choukri, Jean-François Bonastre, et Guillaume Gravier. The ESTER Phase II Evaluation Campaign for the Rich Transcription of French Broadcast News. Dans *Proceedings of the 9th European Conference on Speech Communication and Technology (InterSpeech'05)*, pages 1149–1152, Lisbon, Portugal, 2005. (Cité page 90.)

Sylvain Galliano, Guillaume Gravier, et Laura Chaubard. The ester 2 evaluation campaign for the rich transcription of French radio broadcasts. Dans *Proceedings of the 10th Annual Conference of the International Speech Communication Association (Interspeech'09)*, pages 2583–2586, Brighton, UK, 2009. (Cité pages 11, 21, 33, 55 et 83.)

Michael Gamon, Carmen Lozano, Jessie Pinkham, et Tom Reutter. Practical Experience with Grammar Sharing in Multilingual NLP. Dans *Workshop From research to commercial applications : making NLP work in practice*, pages 49–56, Madrid, Spain, 1997. (Cité page 43.)

Bauer Gerhard. Namenkunde des Deuschen. *Germanistische Lehrbuchsammlung*, 1985. (Cité pages x, 13 et 15.)

Agustín Gravano, Martin Jansche, et Michiel Bacchiani. Restoring punctuation and capitalization in transcribed speech. Dans *Proceedings of the 34th IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'09)*, pages 4741–4744, Taipei, Taiwan, 2009. (Cité page 39.)

Guillaume Gravier, Gilles Adda, Niklas Paulsson, Matthieu Carré, Aude Giraudel, et Olivier Galibert. The etape corpus for the evaluation of speech-based tv content processing in the french language. Dans *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'12)*, pages 114–118, Istanbul, Turkey, 2012. (Cité pages x, 55 et 59.)

Spence Green, Nicholas Andrews, Matthew R. Gormley, Mark Dredze, et Christopher D. Manning. Entity clustering across languages. Dans *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies (NAACL-HLT'12)*, pages 60–69, Montreal, Canada, 2012. (Cité page 41.)

Ralph Grishman et Beth Sundheim. Message Understanding Conference - 6 : A Brief History. Dans *Proceedings of the 16th conference on Computational linguistics (COLING'96)*, pages 466–471, Copenhagen, Denmark, 1996. (Cité pages 9 et 10.)

Claire Grover, Sharon Givon, Richard Tobin, et Julian Ball. Named Entity Recognition for Digitised Historical Texts. Dans *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, 2008. (Cité page 37.)

Mohamed Hatmi. Adaptation d'un système de reconnaissance d'entités nommées pour le français à l'anglais à moindre coût. Dans *Actes de la 15ième Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RECITAL'12)*, pages 151–161, Grenoble, France, 2012. (Cité page 43.)

Mohamed Hatmi, Christine Jacquin, Emmanuel Morin, et Sylvain Meignier. Named entity recognition in speech transcripts following an extended taxonomy. Dans *Proceedings of the First Workshop on Speech, Language and Audio in Multimedia (SLAM'13)*, pages 61–65, Marseille, France, 2013. (Cité page 23.)

Marti A. Hearst. Automatic acquisition of hyponyms from large text corpora. Dans *Proceedings of the 14th conference on Computational linguistics (COLING'92)*, pages 539–545, Nantes, France, 1992. (Cité page 25.)

Akaaki Hori et Nakamura Atsushi. An extremely-large-vocabulary approach to named entity extraction from speech. Dans *Proceedings of the 31st IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'06)*, pages 973–976, Toulouse, France, 2006. (Cité page 35.)

Fei Huang. *Multilingual named entity extraction and translation from text and speech*. Thèse de doctorat, Carnegie Mellon University, USA, 2006. (Cité page 47.)

Xuedong Huang, Fileno Alleva, Mei-Yuh Hwang, et Ronald Rosenfeld. An overview of the sphinx-ii speech recognition system. Dans *Proceedings of the workshop on Human Language Technology (HLT'93)*, pages 81–86, Princeton, NJ, USA, 1993. (Cité pages 87 et 89.)

Hideki Isozaki. Japanese named entity recognition based on a simple rule generator and decision tree learning. Dans *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics (ACL'01)*, pages 314–321, Toulouse, France, 2001. (Cité page 23.)

Hideki Isozaki et Hideto Kazawa. Efficient Support Vector Classifiers for Named Entity Recognition. Dans *Proceedings of the 19th International Conference on Computational Linguistics (COLING'02)*, pages 390–396, Morristown, NJ, USA, 2002. (Cité page 23.)

Fred Jelinek. Continuous speech recognition by statistical methods. *Proceedings of the IEEE, Volume 64*, pages 532–556, 1976. (Cité page 88.)

Vincent Jousse. *Identification nommée du locuteur : exploitation conjointe du signal sonore et de sa transcription*. Thèse de doctorat, Université du Maine, France, 2011. (Cité page 36.)

Mahboob Alam Khalid, Valentin Jijkoun, et Maarten De Rijke. The impact of named entity normalization on information retrieval for question answering. Dans *Proceedings of the IR research, 30th European conference on Advances in information retrieval (ECIR'08)*, pages 705–710, Glasgow, UK, 2008. (Cité page 32.)

Ji-hwan Kim et P.C. Woodland. A Rule-Based Named Entity Recognition System for Speech Input. Dans *Proceedings of the International Conference on Spoken Language Processing (ICSLP'00)*, pages 521–524, Beijing, China, 2000. (Cité page 21.)

Sungchul Kim, Kristina Toutanova, et Hwanjo Yu. Multilingual Named Entity Recognition using Parallel Data and Metadata from Wikipedia. Dans *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL'12), volume 1*, pages 694–702, Jeju, Korea, 2012. (Cité page 48.)

Alexandre Klementiev et Dan Roth. Named entity transliteration and discovery from multilingual comparable corpora. Dans *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics (HLT-NAACL'06)*, pages 82–88, New York, NY, USA, 2006. (Cité pages 48 et 49.)

George R. Krupka et Kevin Hausman. IsoQuest Inc. : Description Of The NetOwl (TM) Extractor System As Used For MUC-7. Dans *Proceedings of the Seventh Message Understanding Conference (MUC-7)*, Fairfax, VA, USA, 1998. (Cité page 31.)

Gakuto Kurata, Nobuyasu Itoh, Masafumi Nishimura, Abhinav Sethy, et Bhuvana Ramabhadran. Leveraging word confusion networks for named entity modeling and detection from conversational telephone speech. *Speech Communication*, 54(3) :491–502, 2012. (Cité page 34.)

John D. Lafferty, Andrew McCallum, et Fernando C. N. Pereira. Conditional random fields : Probabilistic models for segmenting and labeling sequence data. Dans *Proceedings of the 18th International Conference on Machine Learning (ICML'01)*, pages 282–289, Williamstown, MA, USA, 2001. (Cité page 62.)

Céline Le Meur, Sylvain Galliano, et Edouard Geoffrois. Conventions d'annotations en Entités Nommées - ESTER -. 2004. (Cité page 11.)

K.-F. Lee, H.-W. Hon, et Raj Reddy. An overview of the SPHINX speech recognition system. Dans *Proceedings of the IEEE Transactions on Acoustics, Speech and Signal (ITASS'90), Volume 38*, pages 35 – 45, 1990. (Cité pages 87 et 89.)

Xiaohua Liu, Shaodian Zhang, Furu Wei, et Ming Zhou. Recognizing named entities in tweets. Dans *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics : Human Language Technologies (HLT'11)*, pages 359–367, Portland, OR, USA, 2011. (Cité page 32.)

Xiaohua Liu, Ming Zhou, Furu Wei, Zhongyang Fu, et Xiangyang Zhou. Joint inference of named entity recognition and normalization for tweets. Dans *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL'12)*, pages 526–535, Jeju Island, Korea, 2012. (Cité page 32.)

J. Makhoul, F. Kubala, R. Schwartz, et R. Weischedel. Performance measures for information extraction. Dans *Proceedings of DARPA Broadcast News*, pages 249–252, Herndon, VA, USA, 1999. (Cité pages 20 et 61.)

Ciro Martins, António Teixeira, et Joao Neto. Dynamic language modeling for a daily broadcast news transcription system. Dans *Proceedings of the 9th Automatic Speech Recognition and Understanding Workshop (ASRU'05)*, pages 165–170, Kyoto, Japan, 2007. (Cité page 104.)

Denis Maurel, Nathalie Friburger, Jean-Yves Antoine, Iris Eshkol-Taravella, et Damien Nouvel. Cascades de transducteurs autour de la reconnaissance des entités nommées. *Traitement Automatique des Langues (TAL)*, 52 :69–96, 2011. (Cité page 21.)

David D. McDonald. Internal and external evidence in the identification and semantic categorization of proper names. *Corpus processing for lexical acquisition, Chapter 2*, pages 61–76, 1996. (Cité pages 14, 17, 18, 21 et 22.)

Collins Michael et Singer Yoram. Unsupervised models for named entity classification. Dans *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (SIGDA'99)*, pages 100–110, 1999. (Cité page 25.)

Andrei Mikheev. A knowledge-free method for capitalized word disambiguation. Dans *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics (ACL'99)*, pages 159–166, Maryland, MD, USA, 1999. (Cité page 17.)

David Miller, Sean Boisen, Richard Schwartz, Rebecca Stone, et Ralph Weischedel. Named entity extraction from noisy input : Speech and OCR. Dans *Proceedings of the 6th Applied Natural Language Processing Conference (ANLC'oo)*, pages 316–324, Seattle, WA, USA, 2000. (Cité pages 37 et 76.)

David Nadeau et Satoshi Sekine. A survey of named entity recognition and classification. *Linguisticae Investigationes*, 30 :3–26, January 2007. (Cité page 21.)

David Nadeau, Peter D. Turney, et Stan Matwin. Unsupervised named-entity recognition : generating gazetteers and resolving ambiguity. Dans *Proceedings of the 19th international conference on Advances in Artificial Intelligence : Canadian Society for Computational Studies of Intelligence (AI'06)*, pages 266–277, Qébec City, Québec, Canada, 2006. (Cité page 31.)

Ajay Nagesh, Ganesh Ramakrishnan, Laura Chiticariu, Rajasekar Krishnamurthy, Ankush Dharkar, et Pushpak Bhattacharyya. Towards efficient named-entity rule induction for customizability. Dans *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL'12)*, pages 128–138, Jeju Island, Korea, 2012. (Cité page 25.)

Damien Nouvel. *Reconnaissance des entités nommées par exploration de règles d'annotation*. Thèse de doctorat, Université François Rabelais Tours, France, 2012. (Cité page 78.)

Damien Nouvel, Jean-Yves Antoine, Nathalie Friburger, et Arnaud Soulet. Fouille de règles d'annotation partielles pour la reconnaissance des entités nommées. Dans *Actes de la 20ème conférence sur le Traitement Automatique des Langues Naturelles (TALN'13)*, pages 421–434, Les Sables d'Olonne, France, 2013. (Cité page 23.)

Stanislas Oger, Georges Linarès, Frédéric Béchet, et Pascal Nocéra. Enrichissement dynamique du vocabulaire à partir du web. Dans *Actes des 27ème Actes des Journées d'Étude sur le Parole (JEP'08)*, pages 341–344, Avignon, France, 2008. (Cité page 104.)

Mai Oudah et Khaled F. Shaalan. A pipeline arabic named entity recognition using a hybrid approach. Dans *Proceedings of the 24th International Conference on Computational Linguistics (COLING'12)*, pages 2159–2176, Mumbai, India, 2012. (Cité page 25.)

Thomas L. Packer, Joshua F. Lutes, Aaron P. Stewart, David W. Embley, Eric K. Ringger, Kevin D. Seppi, et Lee S. Jensen. Extracting person names from diverse and noisy OCR text. Dans *Proceedings of the 4th workshop on Analytics for noisy unstructured text data (AND'10)*, pages 19–26, Toronto, ON, Canada, 2010. (Cité page 38.)

Woojin Paik, Elizabeth D. Liddy, Edmund Yu, et Mary McKenna. Categorizing and standardizing proper nouns for efficient information retrieval. *Corpus processing for lexical acquisition*, pages 61–73, 1996. (Cité page 13.)

Carolina Parada, Mark Dredze, et Frederick Jelinek. OOV Sensitive Named-Entity Recognition in Speech. Dans *Proceedings of the 12th Annual Conference of the International Speech Communication Association (Interspeech'11)*, pages 2085–2088, Florence, Italy, 2011. (Cité pages 34 et 82.)

Thierry Poibeau. « Deconstructing Harry », une évaluation des systèmes de repérage d'entités nommées. *Revue de la société d'électronique, d'électricité et de traitement de l'information*, pages 1265–6534, 2001. (Cité page 21.)

Thierry Poibeau. The multilingual named entity recognition framework. Dans *Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics - Volume 2 (EACL'03)*, pages 155–158, Budapest, Hungary, 2003. (Cité page 43.)

Lawrence R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. Dans *Proceedings of the IEEE, Volume 77*, pages 257–286, 1989. (Cité page 88.)

Lance A Ramshaw et Mitchell P Marcus. Text chunking using transformation-based learning. Dans *Proceedings of the Third ACL Workshop on Very Large Corpora (WVLC'95)*, pages 82–94, Boston, MA, USA, 1995. (Cité page 63.)

Christian Raymond. Robust tree-structured named entities recognition from speech. Dans *Proceedings of the International Conference on Acoustic Speech and Signal Processing (ICASSP'13)*, Vancouver, Canada, 2013. (Cité pages 23 et 78.)

Christian Raymond et Julien Fayolle. Reconnaissance robuste d'entités nommées sur de la parole transcrite automatiquement. Dans *Actes de la 17ème conférence sur le Traitement Automatique des Langues Naturelles (TALN'10)*, Montréal, Canada, 2010. (Cité page 24.)

Christian Raymond et Giuseppe Riccardi. Generative and discriminative algorithms for spoken language understanding. Dans *Proceedings of the 8th Annual Conference of the International Speech Communication Association (Interspeech'07)*, pages 1605–1608, Antwerp, Belgium, 2007. (Cité page 62.)

Evans Richard. A framework for named entity recognition in the open domain. Dans *Proceedings of the Recent Advances in Natural Language Processing (RANLP'03)*, pages 137–144, Borovets, Bulgaria, 2003. (Cité page 25.)

Alexander Richman et Patrick Schone. Mining wiki resources for multilingual named entity recognition. Dans *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL'08)*, pages 1–9, Columbus, OH, USA, 2008. (Cité page 48.)

C. J. Van Rijsbergen. *Information Retrieval*. Butterworth-Heinemann, Newton, MA, USA, 1979. (Cité page 20.)

Alan Ritter, Sam Clark, Mausam, et Oren Etzioni. Named Entity Recognition in Tweets : An Experimental Study. Dans *Proceedings of the 2011 Conference on Empirical Methods in Natural Language (EMNL'11)*, pages 1524–1534, Edinburgh, Scotland, UK., 2011. (Cité page 32.)

Sophie Rosset, Cyril Grouin, Olivier Galibert, Pierre Zweigenbaum, Karën Fort, et Ludovic Quintard. Les entités nommées dans le programme quaero. Dans *Journée Atala 2011 : Reconnaissance d'Entités Nommées Nouvelles Frontières et Nouvelles Approches*, Paries, France, 2011a. (Cité page 61.)

Sophie Rosset, Cyril Grouin, et Pierre Zweigenbaum. Entités nommées structurées : guide d'annotation Quaero. 2011b. (Cité pages ix, 12, 13 et 55.)

Satoshi Sekine, Kiyoshi Sudo, et Chikashi Nobata. Extended Named Entity Hierarchy. Dans *Proceedings of 3rd International Conference on Language Resources and Evaluation (LREC'02)*, pages 1818–1824, Las Palmas, Canary Islands, Spain, 2002. (Cité pages x, 13, 14, 16 et 55.)

Christophe Servan, Christian Raymond, Frédéric Béchet, et Pascal Nocéra. Conceptual decoding from word lattices : application to the spoken dialogue corpus media. Dans *Proceedings of the 7th Annual Conference of the International Speech Communication Association (Interspeech'06)*, pages 1614–1617, Pittsburgh, PA, USA, 2006. (Cité page 34.)

Fei Sha et Fernando Pereira. Shallow parsing with conditional random fields. Dans *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology (NAACL'03)*, pages 134–141, Edmonton, Canada, 2003. (Cité page 62.)

Cucerzan Silviu et Yarowsky David. Language Independent Named Entity Recognition Combining Morphological and Contextual Evidence. Dans *Proceedings of Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (SIGDAT'99)*, pages 90–99, Maryland, MD, USA, 1999. (Cité page 17.)

Rosa Stern et Benoît Sagot. Détection et résolution d'entités nommées dans des dépêches d'agence. Dans *Actes de la 17ème conférence sur le Traitement Automatique des Langues Naturelles (TALN'10)*, Montréal, Canada, 2010. (Cité pages 21 et 45.)

Andreas Stolcke. SRILM - An Extensible Language Modeling Toolkit. Dans *Proceedings of 9th International Conference on Spoken Language (ICSLP'02)*, pages 901–904, Denver, Colorado, USA, 2002. (Cité pages 88 et 95.)

Krishna Subramanian, Rohit Prasad, et Prem Natarajan. Robust named entity detection from optical character recognition output. *International Journal on Document Analysis and Recognition*, pages 189–200, 2011. (Cité page 38.)

Katsuhito Sudoh, Hajime Tsukada, et Hideki Isozaki. Incorporating speech recognition confidence into discriminative named entity recognition of speech data. Dans *Proceedings of the 44th Annual Meeting on Association for Computational Linguistics (ACL'06)*, pages 617–624, Sydney, Australia, 2006. (Cité pages 39 et 82.)

Nina Tahmasebi, Gerhard Gossen, Nattiya Kanhabua, Helge Holzmann, et Thomas Risse. Neer : An unsupervised method for named entity evolution recognition. Dans *Proceedings of the 24th International Conference on Computational Linguistics (COLING'12)*, pages 2553–2568, Mumbai, India, 2012. (Cité page 25.)

Erik F. Tjong Kim Sang et Fien De Meulder. Introduction to the CoNLL-2003 shared task : language-independent named entity recognition. Dans *Proceedings of the 7th conference on Natural language learning (NLL'03)*, pages 155–158, Edmonton, Canada, 2003. (Cité page 10.)

Mickaël Tran. *Prolexbase. Un dictionnaire relationnel multilingue de noms propres : conception implantation et gestion en ligne*. Thèse de doctorat, Université François Rabelais de Tours, France, 2006. (Cité page 12.)

Karkaletsis Vangelis, D. Spyropoulos Constantine, et Petasis George. Named Entity Recognition from Greek Texts : the GIE Project. *the GIE Project'. In : S.Tzafestas (ed.) : Advances in Intelligent Systems : Concepts, Tools and Applications*, pages 131–142, 1998. (Cité page 17.)

Nina Wacholder, Yael Ravin, et Misook Choi. Disambiguation of proper names in text. Dans *Proceedings of the 5th conference on Applied natural language processing (ANLC'97)*, pages 202–208, Washington, PA, USA, 1997. (Cité page 21.)

Xiaojun Wan, Liang Zong, Xiaojiang Huang, Tengfei Ma, Houping Jia, Yuqian Wu, et Jianguo Xiao. Named entity recognition in chinese news

comments on the web. Dans *Proceedings of the International Joint Conference on Natural Language Processing (IJCNLP'11)*, pages 856–864, Chiang Mai, Thailand, 2011. (Cité page 32.)

Mengqiu Wang, Wanxiang Che, et Christopher D. Manning. Joint word alignment and bilingual named entity recognition using dual decomposition. Dans *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL'13)*, pages 1073–1082, Sofia, Bulgaria, 2013. (Cité page 48.)

Ye-Yi Wang, Alex Acero, et Ciprian Chelba. Is word error rate a good indicator for spoken language understanding accuracy. Dans *Automatic Speech Recognition and Understanding (ASRU'03)*, pages 577–582, St. Thomas, VI, USA, 2003. (Cité page 34.)

Lufeng Zhai, Pascale Fung, Richard Schwartz, Marine Carpuat, et Dekai Wu. Using n-best lists for named entity recognition from chinese speech. Dans *Proceedings of Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics (HLT-NAACL'04)*, pages 37–40, Boston, MA, USA, 2004. (Cité pages 34 et 82.)

Azeddine Zidouni. *Modèles graphiques discriminants pour l'étiquetage de séquences : application à la reconnaissance d'entités nommées radiophiniques*. Thèse de doctorat, Université de la Méditerranée, France, 2010. (Cité page 62.)

Résumé La Reconnaissance des entités nommées est une sous-tâche de l'activité d'extraction d'information. Elle consiste à identifier certains objets textuels tels que les noms de personne, d'organisation et de lieu. Le travail de cette thèse se concentre sur la tâche de reconnaissance des entités nommées pour la modalité orale. Cette tâche pose un certain nombre de difficultés qui sont inhérentes aux caractéristiques intrinsèques du traitement de l'oral (absence de capitalisation, manque de ponctuation, présence de disfluences et d'erreurs de reconnaissance...). Dans un premier temps, nous étudions les spécificités de la reconnaissance des entités nommées en aval du système de reconnaissance automatique de la parole. Nous présentons une méthode pour la reconnaissance des entités nommées dans les transcription de la parole en adoptant une taxonomie hiérarchique et compositionnelle. Nous mesurons l'impact des différents phénomènes spécifiques à la parole sur la qualité de reconnaissance des entités nommées. Dans un second temps, nous proposons d'étudier le couplage étroit entre la tâche de transcription de la parole et la tâche de reconnaissance des entités nommées. Dans ce but, nous détournons les fonctionnalités de base d'un système de transcription de la parole pour le transformer en un système de reconnaissance des entités nommées. Ainsi, en mobilisant les connaissances propres au traitement de la parole dans le cadre de la tâche liée à la reconnaissance des entités nommées, nous assurons une plus grande synergie entre ces deux tâches. Nous menons différents types d'expérimentations afin d'optimiser et d'évaluer notre approche.
Mots-clés Reconnaissace des Entités Nommées, Reconnaissance Automatique de la Parole, Champs Conditionnels Aléatoires, Modèle de Langage.

Abstract Named entity recognition is a subtask of information extraction. It consists of identifying some textual objects such as person, location and organization names. The work of this thesis focuses on the named entity recognition task for the oral modality. Some difficulties may arise for this task due to the intrinsic characteristics of speech processing (lack of capitalisation marks, lack of punctuation marks, presence of disfluences and of recognition errors...). In the first part, we study the characteristics of the named entity recognition downstream of the automatic speech recognition system. We present a methodology which allows named entity recognition following a hierarchical and compositional taxonomy. We measure the impact of the different phenomena specific to speech on the quality of named entity recognition. In the second part, we propose to study the tight pairing between the speech recognition task and the named entity recognition task. For that purpose, we take away the basic functionnalities of a speech recognition system to turn it into a named entity recognition system. Therefore, by mobilising the inherent knowledge of the speech processing to the named entity recognition task, we ensure a better synergy between the two tasks. We carry out different types of experiments to optimize and evaluate our approach.
Keywords Named Entity Recognition, Automatic Speech Recognition, Conditional Random Fields, Language Modeling.